Fri 11/4/2011 4:03 PM

Response to request for comments

As a researcher who conducts research on human subjects, and who is active in training IRBs and others on the ethical issues surrounding human subject research, particularly online, I want to express my concern about the apparently conflicting objectives being pursued by the federal government when it comes to research data associated with humans (whether technically human subjects or not). As has become very clear through highly-publicized examples (the AOL search dataset, the Netflix Prize dataset), and as will become even more clear with any datasets involving personal or medical data, the ability to re-establish links between "deidentified data" and individuals is increasing, and data sets that were once believed to present no risk would today be considered too much at risk for re-identification. There is no reason to believe that this continuing evolution will change.

Part of this challenge is fundamental. It is always easier to seek a match for an individual in a large dataset than to reidentify the full set. Consider a dataset that contains only date of birth, three digits from a social security number, three digits from a phone number, and some sensitive information (e.g., HIV status). From that data set, it would be computationally hard, if not intractable, today to build a list of HIV-positive people. But a person with access to one individual's information could trivially check whether that individual is in the dataset, and find his status if so. This asymmetry, combined with the pace of technological change, is rarely considered adequately in making plans to archive data for permanent use.

Hence, I strongly urge that any policy that addresses mandates for data archival have a very clear section on the policies and practices for opting out in cases where the usefulness of the data cannot be separated from potentially reidentifiable personal information, and for subsequent removal of datasets from archives if unexpected developments lead the datasets to present unacceptable risks to individuals from or about whom the data was collected.

JK

--
--
Joseph A. Konstan
Distinguished McKnight Professor and Distinguished University Teaching Professor
Associate Department Head
Department of Computer Science and Engineering
University of Minnesota

Mon 11/7/2011 11:21 AM

I agree in principle with the notion that whatever data is generated under federally funded research should be available to other researchers. But there is one, perhaps insurmountable problem: under current IRB regulations, an investigator must provide all IRBs overseeing a research project with the names and contact information of everyone who might access those data, even de-identified data. That is not possible under any data sharing agreement. And I don't know that I would agree with overriding individual IRB concerns and controls. If I were a subject in a research project, I would want to know who would be seeing and using my data.

Perhaps the solution would be having a 2-level consent statement (or 2-level waiver from the IRB): one giving the IRB-approved researchers access to the data, and one allowing the data to be shared broadly. A subject providing informed consent could decide whether anyone other than the investigators could see and analyze his or her data and could therefore restrict broad access without restricting local access by the study's investigators.

- Bill Tierney

William M. Tierney, MD, MACP
President/CEO, Regenstrief Institute, Inc.
Sam Regenstrief Professor of Health Services Research Associate Dean for Healthcare Effectiveness Research Indiana University School of Medicine Chief, Medicine Service, Wishard Hospital and Wishard Health Services

---

Mon 11/7/2011 1:56 PM

Comment on OSTP changes: public access to research data from federally funded research

I have the following comments regarding public access to research data from federally funded research:

- Such data is important competitive intellectual capital of the US and its tax-payers and should be available for peer review, re-purposing, and re-use
- Not all data is curate-able, re-usable and re-purposable; agencies should customize requirements based on the type of research proposed. This could be in the form of:
  -- A digital record in an Institutional Repository about the project, the data collected, the basic methodology, BUT not the actual data, rather a pointer to the PI for more information
  -- A curated, representative sample of the data, especially if the data is collected for a specific research purpose from constantly streaming sensor data (e.g. the Huble Telescope)
  -- Fully curated data sets for medium to small sized research projects.
- There should be criteria for whether data from a project is required to be deposited in a national subject depository (e.g. genomic data) or in a local (e.g. university) data repository.
- Standardized tool sets for collecting and preserving data and scientific workflows should be encouraged (guided?) by the specific agencies, such as the DataOne network for environmental data at the University of New Mexico.

Thank you,

Johann van Reenen
Professor and Associate Vice President, Research Initiatives
Office of the Vice President for Research
1 University of New Mexico
327 Scholes Hall, MSC 05 3480
Albuquerque, NM 87131

Below are thoughts on the importance of "GIS-Ready" data products to increase the accessibility to and usability of, in this case, remote sensing data. Such derived products would greatly increase the value extracted from the vast amount of data stored and being acquired.

It specifically addresses item (1) Preservation, Discoverability, and Access in the RFI, noting that "Access" includes not just physical access—data that is physically accessible but in a format or form that a potential user cannot utilize or understand is not "accessible".

Gary N. Geller, Ph.D.
- Deputy Manager, NASA Ecological Forecasting Program
- Conservation Liaison, ASTER Science Project
Jet Propulsion Laboratory
MS171-264, 4800 Oak Grove Drive
Pasadena, CA 91109-8099 USA
+1 818-354-0133 FAX: +1 818-393-1370 GMT -8
http://terralook.cr.usgs.gov

The Need for GIS-Ready Products
Concept Note, 7 November 2011, Gary Geller (gary.n.geller@jpl.nasa.gov)

Increased accessibility, usability, and availability of remotely sensed image products is essential if the needs of less technical users such as educators, natural resource managers, policy makers, and the general public are to be met. It is with these goals and users in mind that GIS-ready products and the coordination ideas suggested here are developed.

This concept note provides background information on what GIS-ready products are, why they are important, and who the target users are, then describes the current situation with respect to their availability. It is hoped that additional discussion on this topic will encourage more data providers to make GIS-ready data products available. And, because availability is already starting to increase, it is suggested that coordination among the providing organizations, captured as product guidelines, will benefit users by simplifying access and usability.

GIS-Ready Products. This imprecise but convenient term refers to georeferenced image products that can be incorporated into a GIS with minimal effort. They utilize simple, common formats that do not require sophisticated, specialized software, but rather can be viewed with common software such as browsers, IrfanView, PowerPoint, or of course any GIS software. Although a range of products of varying complexity can be considered "GIS-ready", here the emphasis is on the simple end of the spectrum with a focus on meeting the needs of less-technical users. And while jpeg images have many advantages, such as small size, any common

format would probably be suitable. GIS-ready products are most commonly surface images from optical sensors like Landsat, but radar images, digitized maps, and atmospheric products such as profiles are also possible.

Cost. Because GIS-ready products are derived from existing scientific products, the algorithms for generating them are simple and easy to code, and generally not a cost consideration. Thus most of the cost of adding these is the marginal cost of extending the system to offer a new product, a cost that will vary depending on the system.

Target Users. GIS-ready products are for any user that can benefit from easy access to images that otherwise would utilize complex formats geared towards remote sensing specialists. These complex formats include, for example, HDF, or any multi-band format that requires knowledge of image processing to compose an image. Target users need not be GIS experts, use GIS software, or even know what GIS is. Because GIS-ready products can be viewed with simple, common software they greatly increase access to and usability of scientific data products—and thus the value that can be extracted from those products. The number of users interested in GIS-ready products far exceeds the number of traditional scientific users.

Current Situation. No-cost access to the vast data stores of remotely sensed data has been gradually increasing. All NASA EOS data, starting with that coming from the Terra spacecraft, have been freely available, at no cost, since the first data reached the ground in 2000. Creation and availability of the Landsat Global Land Survey (GLS) products in 2000 was another big step forward. Eliminating user charges for the entire Landsat archive, which USGS did in 2008, was another (very) big step forward for open access. Brazil and China now provide greater access to the images in the CBERS archive, and India recently updated their data policy so that all images coarser than 1m are now accessible. Although this is an excellent trend, in most or all of these examples the products offered are very large, multi-band datasets that require users to select and combine some of the bands if they need an image. That step blocks access by the huge number of users needing a smaller, simpler, derived product—a GIS-ready product such as an image layer. TerraLook, a NASA-USGS effort, addresses this problem by making images of the GLS and ASTER (Advanced Spaceborne Thermal Emission and Reflection) archives available as georeferenced jpegs--but this approach needs to become more widespread. To summarize, there is still much room for progress on two fronts: 1) increasing the amount of no (or very low) cost data available and 2) increasing the usability of data by offering derived, GIS-ready products. These two activities may be related, as explained in the next paragraph.

What Happened with ASTER. Although the Earth Remote Sensing Data Analysis Center (ERSDAC), which oversees the ASTER dataset, charges for the full multi-band product, TerraLook encouraged ERSDAC to allow their products to be made available at no cost as TerraLook's simple georeferenced jpeg format. This opened up the entire ASTER archive, currently about two million images, to everyone; ASTER is now available to a much broader range of users than the highly specialized scientific audience that is the primary target for most

sensors. Because GIS-ready products cater to a different audience than the specialized scientific one, the increased access provided by GIS-ready products does not decrease the sales of the full product; in fact, it may increase sales by increasing visibility and interest. Perhaps the approach taken by ASTER can act as a model for other data providers and encourage them to make simple GIS-ready products available at minimal or no cost.

Guidelines. As more GIS-ready data products become available there is the possibility that they will be offered in a multitude of formats or with incomplete or parochial metadata. Such a divergence would stymie accessibility and decrease the value to users as well as the value extracted from the data. This scenario can be avoided by simple coordination among data providers to make GIS-ready products more consistent. Once a data provider has decided to offer GIS-ready products, consistency does not impose any incremental costs.

Compared to the target audience for full scientific products, the users for GIS-ready products have a much narrower range of product requirements—this fact makes convergence towards a limited number of formats possible. But since such convergence is unlikely to emerge on its own, as a first step forward it may be wise to propose a set of guidelines. These guidelines could cover basic issues such as format (eg, jpeg with World file, tiff with World file, or geotiff), metadata format, and metadata content, and they could include information to limit the number of, for example, coordinate systems used so as to enhance compatibility with common software. Note that the goal here is not to create a new standard, but rather to coordinate formats and metadata, using existing standards.

Regardless of the details, however, the key goal is to increase access to more datasets for a broader range of users. Simple, GIS-ready products are an excellent way to do that, and perhaps TerraLook and ASTER provide a useful model as a starting point.

Tue 11/8/2011 12:51 PM

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Data and associated software should have an entry in bibliographic listings together with the citation of the corresponding article. This should be enforced through the journals that publish the article.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data and associated software should have an entry in bibliographic listings together with the citation of the corresponding article. This should be enforced through the journals that publish the article.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

For computational modeling software, there are already developed standards (SBML, NeuroML, and related).

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Data and associated software should have an entry in bibliographic listings together with the citation of the corresponding article. This should be enforced through the journals that publish the article. This could be enforced through the PubMed database for example, as a requisite for correct citation in the database.

Response to this RFI is voluntary. Responders are free to address any or all the above items, as well as provide additional information that they think is relevant to developing policies consistent with increased preservation and dissemination of broadly useful digital data resulting from federally funded research. Please note that the Government will not pay for response preparation or for the use of any information contained in the response.

German Cavelier PhD, Office of Technology Development and Coordination

National Institute of Mental Health

The Cost of Free

Holly Cowan Shulman

26 October 2011

Digital publication is neither free nor cheap. Despite enthusiasm for scholarly publication at no cost to the end user, free publication will not solve the dilemmas that have upended the ecosystem of scholarly communications since the invention of the World Wide Web. As a community with many different stakeholders, ranging from scholars to universities, libraries to publishers, and funders to programmers, we need to think more clearly about the future of scholarly electronic publication.

The World Wide Web burst into public awareness as an open space for freely available communications in 1993 with the introduction of Mosaic, the precursor of Netscape.[i] By the end of the 1990s that space had been further transformed by increasingly effective search engines, especially Google in 1998. As we all know, the impact was, and remains, a virtual tsunami on the world of communications. The WWW has already changed the environment for newspapers, magazines, music, political organization, and personal communications. It threatens cable and broadcasting. It is in the process of upending publishing and scholarly communications as one piece of this global shift.

The impact of this revolution has played out differently in various arenas. The cost of journals – especially those in Science, Technology, and Medicine (STM) – has skyrocketed while the budgets of libraries and institutions of higher education have decreased. In STM fields the importance of early publication is important, and publication online is a solution to the time factor.[ii] In some areas of the humanities

online publication has posited a solution for the difficulties of finding traditional scholarly outlets. As the wages of most Americans over the past generation, including most academics, has flattened, the prospect of reducing the cost barriers of purchasing books, subscribing to journals and buying newspapers has been seductive, especially as new reading devices such as Kindle and iPad appear and their prices drop. The role of agents is threatened as new organizations are formed for authors who seek to self-publish their books.[iii] We are now at the point of vertical integration of the publishing industry as Amazon expands beyond a virtual bookstore to become a publishing house. As a recent report published by the British Research Information Network (RIN) said: "At a time of financial stringency for universities, research funders and publishers, it is important that all the stakeholders in the scholarly communications system work together to find the most cost-effective ways of fulfilling their joint goal of increasing access to the outputs of research."[iv]

The United States government has weighed into this debate. Research funded by public monies, they argue – especially STM – should be freely available to the American public. STM journals have responded by charging their authors to publish, a practice now known as "author-side payments." Already 40% of biomedical journals work this way.[v] There is also the policy of online publication after a time barrier, which most of us know through JSTOR and MUSE as two examples of aggregators that sell their product rather than give it away. Some members of the digital humanities community passionately believe that online scholarly communications should be freely available to all as a matter of policy in order to open academic discourse to the widest audience possible.

Libraries are integral to this new mix. Where once they provided shelves, cataloging, reference, and an occasional rebinding services, they now must subscribe to new online publications and host their own catalogs and other works. In some cases libraries embrace their university publishers or have become close working colleagues of the university publishing house. There is a movement for open repositories where any scholar in that institution is expected (if not required) to post their work.[vi] Libraries have also hosted a world of experimentation such as the Institute for Advanced Technology in the Humanities and the Scholars' Lab at the University of Virginia. Adding to the seduction of library publication are copyright rules that differ for educational versus commercial use. If a humanities scholar wants to publish images, for example, she or he may find it far simpler to stay within presently defined academic boundaries and publish their work through their library.

Within this shifting ecosystem of scholarly communications, lies the small world of reference works, and within that world the smaller arena of documentary editions, and even within that the very small field of documentary editions in history. This is the world of the "the papers of….": Thomas Jefferson, Albert Einstein, Elizabeth Cady Stanton, Margaret Sanger, George Washington, the Freedmen and Southern Society Project, Thomas Edison, James Madison, Eleanor Roosevelt, Naval Documents of the America Revolution, Frederick Douglass, Abraham Lincoln, Cordell Hull, Andrew Jackson, Martin Luther King, and dozens more including my own *Dolley Madison Digital Edition*. While some have private funding, most operate through grants received from the federal

government, specifically the National Endowment for the Humanities and the National Historical Publications and Records Commission. These funders are now considering the pros and cons of demanding not only electronic publication, but also free access to those electronic publications. The implicit question is: what are the costs of free? Do we need to unbind ourselves from the world of publishing in order to nurture creativity and open dialogue – or are these illusory goals that will destroy the armature of editions by replacing cost of purchase with cost of production, hosting, and maintaining?

*

It is my belief, to begin with, that a documentary edition should be neither a website nor an electronic book. Information on a website moves through a series of links that allows the reader to go from page to page. Its basic language is HTML. It is better suited to an electronic exhibition than an electronic archive. An electronic book transforms the printed page into displayable and searchable pixels that visualize a book for the consumer. It adheres closely to the traditions of print, plus perhaps a slider on the bottom, a page marker, a percentage number to locate the reader in the book, and various ways of highlighting the text and taking notes.

As a reference work and research tool, the documentary edition is best suited to the format of a digital archive, or a storage repository. After all, the edition may include thousands, if not millions of documents. Unlike a novel or an essay or a scholarly monograph, there is no argument being made, no narrative, no real spine at all on which to hang the bones of the edition. It is a collection that is searchable through a process of identification and constraint. Each piece of information is like a pebble on a beach, and

when you do a search you pull up from the beach a certain number of those pebbles. When you're done, those pebbles drop back to the beach. The scholar or student can read it from end to end, or browse through it, or search for what they want through different search tools. This representation, however, requires that whoever is publishing the work tag it in XML, most likely using the encoding guidelines of the Text Encoding Initiative (TEI), in a manner conformant to a whole series of programming demands. It also means that its display be reasonably intuitive so that the reader can navigate the text. In sum, it means that a well-done digital documentary edition should not be composed of PDFs taken from a printed book, or a series of pages simply coded in HTML, and posted on line. The internal structure of a digital edition is important. It may be altered; it may be expanded; but it cannot simply be thrown up on the web as can a blog, for example.

The central argument, however, revolves around what are the advantages and disadvantages of digital versus print. After all, a print edition is a wonderful object, replete with nicely formed pages, all kinds of fore matter such as preface and introduction, bibliographical information and a table of contents. It has an index that allows the contents to be searched. Why bother to move over to an electronic format? Are the reasons cost or scholarly benefit? And do these align on one side or the other of this argument over free?

There has long been an assumption that digital is cheap, in fact so cheap it can be given away for free. The hypothesis is in part built upon the way the medium grew in the 1990s when newspapers, for example, started publishing for free electronically as well as

for a price in print. They rue the day. Today newspapers are establishing cost firewalls to protect their bottom line, while they fight off competition from bloggers and radio and television online products. Who wants to pay for *The New York Times* if you can get it for free? The print world of daily news and magazines is demanding subscription fees while also adding new content to give additional value to the online subscriber reader. The *Times* is a leading example of value added electronic materials rendering a superb electronic publication – at a cost. We all know that the print world is struggling to balance production costs with revenue. Amazon.com set a radically low fee to entice readers to purchase e-books so they could sell the Kindle, and as observed above are now trying to further cut their own costs through vertical integration of production by starting their own publishing organization. Meanwhile, the rest of the book-publishing world struggles to keep up. The American Association of University Presses recently issued a report tackling the problem of communications transformation entitled *Sustaining Scholarly Publishing* in which they lay out their own issues and explore viable solutions.[vii]

The cost of creating quality electronic scholarly publications in the field of documentary editions is, in fact, not cheap. Currently, funding grants carry a project through to the point where they are assumed to hand it over to someone who will then take care of it and somehow get it online. For starters, this model portends the elimination of a number of important editorial interventions that publishing houses undertake. They copyedit. They provide peer review. They help develop a product. They market. All of this has long

been true in print. In the new world of digital, presses must also check data for consistency, extend code for new features, create new data interfaces, and so on. They must have someone who can deal with the challenging world not only of markup languages but also of the specifications provided by the TEI. Without these standards there can be no consistency. Without this regularity there can be no interoperability or cost scaling or industrial output; electronic publication would instead remain a craft in which each producer created their own unique widget that would not play with their neighbor's or peer's widget.

Every edition needs to be "served," or hosted on a server. A good one is costly; even purchasing a license to use a piece of it is expensive. Moreover open source software needs customization. Most editors simply cannot open up a box entitled Drupal, an open source content management system, and make it work for their project without external help. There are start up costs and there are platform costs and there are new content costs. And there are the continuing costs of maintaining a product in a world of constantly shifting technology in which what works today will not necessarily function tomorrow, of sustaining the machine and the human resources to maintain that content.[viii]

Imagine that you are the editor of a small project, and you have decided – eagerly or reluctantly – to be born-digital. That said -- to whom would you go? Who would you find as a programmer, an information systems person, or a designer? You would need a new way to control and track your work and your deadlines. Where would you find an appropriate, and free, content management system? What would your overall costs be?

Whereas in print you would have simply handed over the manuscript to your publisher, who would have counted on recouping all costs through the sale of books, now you can no longer do so. You may fare well at a large university, but suppose you are an editor residing at a small institution. Or perhaps you are a retired professor and want to edit an edition as a retirement project. To whom do you turn for advice; what questions do you even ask? And if you have conquered the challenges of DTDs, XML, and TEI for encoding ordinary text, then what do you do when you later want to add different kinds of records such as inventories or ledgers? Two years on, can you find the same team who first set you up? Without a publisher, who guarantees that the work is not only quality scholarship, but even legitimate? Large-scale, well-funded projects can hire developers. Small projects usually run on a shoestring. And yet to date there has been NO talk about adding subvention costs to grant funding. The bottom line is that digital publication is neither free nor cheap. Three months of a programmer's time might cost $25,000. The cost of a production-level license for a first-rate XML publishing environment like MarkLogic Server runs upwards of $30,000 at current prices.. And every time you wanted to add a new installment you would need editorial and technical work that would easily cost $5,000 to $10,000. And that is assuming there is a structure somewhere out there that can do for you what you want.

Electronic publication has important advantages: but only if done well. To begin with there is the compelling allure of cross-searchability, interoperability, and aggregation. The execution of this vision is, to date, tied to server and tagging issues. Thus Rotunda, the electronic imprint of the University of Virginia Press, can publish compilations such

as Founders Online and make these editions cross searchable. There is a huge scholarly reference advantage here, but had Hamilton, Madison, Washington, and Jefferson all been published in discrete models this integration would not yet be possible. The *Dolley Madison Digital Edition* is now beginning to include lists of 300 names or so that will require a new kind of tagging; Rotunda is there, ready and willing and able to help accomplish this goal. Single volumes in a series such as the *Papers of George Washington* are now tied together and the researcher can do one simple search that carries her or him through all 60 volumes. Documents Compass, a unit in the Virginia Foundation for the Humanities, has been working on collecting all of the names and biographical references in all the founding editions and creating both a biographical dictionary and a prosopography. Funded by the Andrew W. Mellon Foundation, this will be a Rotunda publication.

The model of author-side payment is not a solution for documentary editions. Nor may individually built editions have the resources or technical infrastructure to create quality publications. I know of no digital humanities tools that confront these basic issues, so while it may be useful to have citizen transcribers using a tool known scripto (scripto.org), it does nothing to solve the basic problems of publication, while it is unclear that it resolves issues of costs in order to professionally establish the texts at hand. While technology is upending the world as we once knew it, we should approach the problems of electronic editions more carefully and beware of solutions that promise free – at great cost.

[i] There were other web solutions in 1993 besides Mosaic but it is this author's memory that Mosaic stole the show. By 1994 a number of search engines started up including Yahoo! and Lycos, which the following year were joined by AltaVista, Magellan, and others. Google was launched in 1998. The combination of browser and search engine transformed communications.

[ii] STM fields tend to be rapidly developing areas of research in ways not true in the Humanities and Social Sciences, and even before the WWW depended upon prepublication of journal articles.

[iii] Julie Bosman, "New Service of Authors Seeking to Self-Publish E-Books," *The New York Times,* 2 October 2011, http://www.nytimes.com/2011/10/03/business/media/perseus-creates-new-service-for-authors-seeking-to-self-publish.html

[iv] *Heading for the open road: costs and benefits of transitions in scholarly communications* http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/heading-open-road-costs-and-benefits-transitions-s

[v] *Heading for the open road.*

[vi] There is the Coalition of Open Access Policy Institutions (COAPI) in the United States. See Jennifer Howard, "Universities Join Together to Support Open-Access Policies," *The Chronicle of Higher Education,* 27 October 2011, http://chronicle.com/blogs/wiredcampus/universities-join-together-to-support-open-access-policies/32632

[vii] *Sustaining Scholarly Publishing: New Business Models for University Presses* (AAUP, March, 2011).

[viii] Matthew Gibson to Humanist Discussion Group (Willard.mccarty@mccarty.org.uk), 26 October 2011.

Fri 11/25/2011 3:45 AM

Public Access to Digital Data

I'm a factory worker for G.M. and a taxpayer.

In my opinion, since it's federally funded, I.E. my tax dollars, the information should be free to anyone with a social security number, or to all levels of education. As for companies or other businesses who want to use some of the data, break out the check book.

I know companies pay taxes too, however they can also afford to kick in toward your budget. I on the other hand don't have much money to spend on information that interest me.

Thank you for your time.

Larry

Sat 12/3/2011 8:07 PM

Public Access to Digital Data Resulting From Federally Funded Scientific Research

I wrote an essay about that about the general topic of funding digital public works about a decade ago for the Markle Foundation when they requested comments on "Policy for a Networked Society": http://www.pdfernhout.net/on-funding-digital-public-works.html

A shorter version of that is here:
http://www.pdfernhout.net/open-letter-to-grantmakers-and-donors-on-copyright-policy.html

Copies of both are attached.

From the executive summary of the shorter one:

"Foundations, other grantmaking agencies handling public tax-exempt dollars, and charitable donors need to consider the implications for their grantmaking or donation policies if they use a now obsolete charitable model of subsidizing proprietary publishing and proprietary research. In order to improve the effectiveness and collaborativeness of the non-profit sector overall, it is suggested these grantmaking organizations and donors move to requiring grantees to make any resulting copyrighted digital materials freely available on the internet, including free licenses granting the right for others to make and redistribute new derivative works without further permission. It is also suggested patents resulting from charitably subsidized research research also be made freely available for general use. The alternative of allowing charitable dollars to result in proprietary copyrights and proprietary patents is corrupting the non-profit sector as it results in a conflict of interest between a non-profit's primary mission of helping humanity through freely sharing knowledge (made possible at little cost by the internet) and a desire to maximize short term revenues through charging licensing fees for access to patents and copyrights. In essence, with the change of publishing and communication economics made possible by the wide spread use of the internet, tax-exempt non-profits have become, perhaps unwittingly, caught up in a new form of "self-dealing", and it is up to donors and grantmakers (and eventually lawmakers) to prevent this by requiring free licensing of results as a condition of their grants and donations."

--Paul Fernhout (Edinburg, NY; Software Developer and Homeschooling/Unschooling Parent)
http://www.pdfernhout.net/ ==== The biggest challenge of the 21st century is the irony of technologies of abundance in the hands of those thinking in terms of scarcity.

OFFICE OF SCIENCE AND TECHNOLOGY POLICY
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

From: C. Carl Jaffe, MD, Professor of Radiology Boston University, carl.jaffe@bmc.org

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?
Any such effort must be framed by a strategic plan that is step-wise since not-sharing is ingrained in current science behavior and encumbered by legal precedent (e.g. Supreme Court re: William Catalona vs WUSTL decision 2008) With regard to human biologic and tissue resources/data, federal policy should promulgate and encourage pre-procedural human consent documents that enables and encourages, on the model of a health-care directive, the healthcare entity to develop sharing processes (allowing for cost-recovery) open to applications from qualified research entities. Except for organ donation, current legal precedent and policies applicable to healthcare tissues removed from patients lack a cohesive modern basis and treat human tissue as potential marketable property.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?
For activities whose funding is more than 50% contributed by public funds, provide guidelines for a uniform period, say 6 months (as is the case with publications), during which the data can be privately withheld by the investigators for exploitation of analysis. With databases that are continuously being assembled without clear endpoint, the principal investigator at the time of initial receipt of public funds should define specific plans for an acceptable release of partial data-sets. Although NIH has stated 'data sharing' policies for extramural research projects that meet a certain cost threshold, little implementation occurs often because no meaningful pathway to ease the burden of data conditioning pre-sharing is funded. Moreover, the policy seems not to be applicable to intramural investigations at NIH. It is not clear whether DOD sponsored healthcare research has defined a research data-sharing policy.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?
Any new policy must recognize that data-conditioning (such as de-identifying human data and mapping data fields to common-data elements before sharing) require effort and add cost. Local process development needs to be encouraged like IRBs which should consider whether the data is amenable and worthy of sharing (not all data is worth preserving if it doesn't meet quality standards or have public funding)

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?
Alternative pathways with different cost requirements need development. Storage costs keep falling and the main cost burden may be pre-conditioning the data and auditing it for quality control. Researchers should be aware that some organizations, federal and private, offer common vocabularies and common data elements that when used at the data acquisition planning greatly ease query retrieval

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?
Federal and professional encouragement resting on ethical and open-science principals that encourage scientists to recognize their self-interest in participating in self-forming cross-disciplinary teams that speeds confirmation of science postulates when data is accessible. This becomes self-evident in certain branches of science such as genetic 'signatures' that are based on statistical clusters since those conclusions are best verified by challenging them with different tools and models

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?
Providing different adaptable/adoptable pathways with alternative effort and cost bases that are case appropriate that investigators can choose from. Then publicizing successful model case examples

that can be emulated.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?
Compliance that benefits the investigator such as a two-step bonus award such as 1) placement of data in a publically internet available shared container 2) evidence of use-case impact of the data on other researchers

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?
Recognize the power of Program Announced public-private partnerships and Cooperative Agreements (such as the NIH U01 mechanism). Set goals for numbers of such classes of funding with the funding agencies (including DOD)

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported? Standards for Interoperability, Re-Use and Re-Purposing
Self-monitored by the investigator and reported to their advantage on followup grant applications by the same investigator

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.
NEMA-radiology cooperation has produced DICOM image standards but pathology would benefit from encouragement. Standards can never be imposed but must be eased in incrementally by advantages accruing to the investigator community

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?
DICOM because it was needed for clinical care interoperability between manufacturers. Hence the key incentives were driven by clinical healthcare operational needs before they were recognized for their greater research value.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?
Moving as fast as possible and trumpeting success when it occurs

(13) What policies, practices, and standards are needed to support linking between publications and associated data?
“Shared data lists” and links to supplementary material is already occurring but hasn’t been fully absorbed or exploited by the science community as of yet. See the success stories by NIH TCGA, NCI-CIP, caBIG and other initiatives

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

This is a response from the UCLA Social Science Data Archive http://dataarchives.ss.ucla.edu/
Prepared by Libbie Stephenson, Director

The Data Archive has been in operation since 1977 and serves the entire UCLA campus of faculty and students engaged in quantitative research, including archiving original collections of data through surveys, and providing access for the re-use of de-identified (public use) data by faculty and students for research and instruction. Faculty members who are engaged in survey research use the Data Archive as a support to data collection processes and life cycle management, and in making the data publicly accessible. The Data Archive is a member of the Inter-university Consortium for Political and Social Research, an organization of over 700 members, worldwide, engaged in the use and preservation of data used in quantitative research. "ICPSR maintains a **data archive** of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields." UCLA is one of the heaviest users of ICPSR public use data, well above (approximately 3x heavier usage) the median usage by other institutions with the same Carnegie Classification. The re-use of public data at UCLA is vital to the generation of scholarship and social science pedagogy.

Thank you for the opportunity to respond to this RFI. We support efforts of the Working Group on Digital Data "to encourage and coordinate the development of [federal] agency policies and standards to promote long-term preservation of and access to digital data resulting from federally funded scientific research." We agree that all such policies should "follow best practices for protecting confidentiality, personal privacy, proprietary interests, intellectual property rights, [and] author attribution…"

Summary of comments:

In this response we suggest that Federal agencies should develop policies reflecting the nature of open access as it is understood today, but also make provision for an evolution of what open access will be given continued changes in technology and the need to be able to address open access from a global perspective. We provide some considerations about the nature of open access followed by comments on the questions posed. Some key points include:

- Open access has associated costs; data are not free; Federal agencies should allocate funds for long term data management.

- Open access data use requires skills and expertise; Federal policies need to acknowledge that open access does not preclude the need for the user to have the necessary technical, statistical and analytical skills necessary to work with research data effectively.

- Open access has boundaries when it comes to issues such as protection of privacy and confidentiality, national security, data embargos, copyright and intellectual property. Policies established by federal agencies should reflect these boundaries in ways that promote the widest possible access.

- Open access goals assume the existence of a robust infrastructure; however, there is no data stewardship infrastructure that currently exists anywhere in the world with the capacity to manage the amount of data generated. Furthermore, there are not enough trained members of the workforce capable of managing the masses and varieties of data being produced. There has been no scientific study of the strengths and weaknesses of current data stewardship organizations, operations, technological approaches, or repository software. The Trusted Repository Audit Framework, the Data Seal of Approval and other assessment tools should be widely applied.

- Policy development aimed at achieving open access to research data can best be accomplished with community involvement of all stakeholders: researchers, archival professional groups and scholarly societies. Evolution of policy should flow between stakeholders and agencies based on scientific study of research methods among disciplines; archival organizations operating at international, national, regional, state and local levels; and of the variety of practices and standards used to manage data for the long term.

- Research proposals containing data management plans should be evaluated by experts in long term data stewardship, and by those with knowledge of best practices for organizing and documenting data. Compliance with data management plans can be verified through the use of registries of unique identifiers.

Some considerations about "open access"

Open access has associated costs

For many the idea of openness suggests no monetary cost. Digital data are not free. Their collection is funded through contracts and grants paid by taxpayers. The computing facilities (software and hardware), technology experts, administration, and data collection itself (such as through survey research centers) is funded partially by grants, but largely by academic institutions. The long term maintenance of digital research data is funded by universities and organizations; for example, the Inter-university Consortium for Political and Social Research (ICPSR) http://www.icpsr.umich.edu is a member-based organization where member dues contribute to data management. These costs are rarely addressed in a research grant and where funds are budgeted, they do not contribute to the long term sustainability of digital data. Therefore, we argue that funds for research data collection do not necessarily provide for free access to digital data. In order for this to be true, federal agencies that fund collection of research data must also allocate monies for the long term management of the data. Experts suggest that in addition to providing support for data preparation, agencies should provide sufficient funding (1-5% of an award) to guarantee long term management of open access data.

Open access data use requires skills and expertise

The term "open access" suggests the idea that data can be used by anyone, even by those who have no knowledge about the original data collection. In order for this to be possible, the use of federal funds for research should support best practices for organizing and documenting data. It is important to note that even with good documentation and publications, data are not necessarily usable by just anyone. Federal policies need to acknowledge that open access does not preclude the need for the user to have the necessary technical, statistical and analytical skills necessary to work with research data effectively. And, while it may be outside the scope of this RFI, it is worth noting that the American public is woefully statistically illiterate. We

encourage federal agencies to provide support to training and education programs so that users beyond the research community can make effective, informed use of digital data.

Open access boundaries

It is also important to consider that not all research data can be categorized as open access, either because of the sheer volume or complexity of the data, or for reasons of confidentiality, national security, embargo period and so forth. We concur with the views expressed by the American Association for the Advancement of Science "that the discussion surrounding public access must clearly distinguish between providing access to research results in support of scientific progress and access to scientific information as a crucial element of public engagement." That is, for some research, open access to underlying data may not be practical or possible; (e.g., the general public does not have the facilities required to personally analyze petabytes of data collected by astronomers), therefore access to information about the research, the analysis of the data, the conclusions reached, and so forth, through publications, may be the best option for providing the general public with usable information about research conducted using tax dollars. We encourage development of policies where these distinctions can be recognized.

Copyright and intellectual property rights boundaries can be addressed within an open access environment. The Creative Commons organization exists to provide a simple way for research data and results to be shared openly with all. http://creativecommons.org/ Creative Commons provides a set of copyright tools and options so that researchers can receive credit for their work and still share it as widely as possible. The result is a public arena where content can be shared, copied, edited, or reformatted, while still acknowledging the work of the original investigator. Depending on the nature of the data and the kinds of rights the investigator desires, a variety of boundaries of access can be established. We would encourage federal agencies to coordinate with or adopt processes to support a Creative Commons approach to issues of intellectual property and copyright parameters.

Open access goals assume the existence of a robust infrastructure

Currently only a very small amount of data collected with funds from Federal agencies is ever deposited for long term stewardship. In a white paper prepared by the International Data Corporation, IDC the authors noted that "In 2006, the amount of digital information created, captured, and replicated was 1,288 x 1018 bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes… This is about 3 million times the information in all the books ever written." Further, the authors projected that "[b]etween 2006 and 2010, the information added annually to the digital universe will increase more than six fold from 161 exabytes to 988 exabytes." **There is no data stewardship infrastructure that currently exists anywhere in the world with the capacity to manage this much data.** Furthermore, **there are not enough trained members of the workforce capable of managing the masses and varieties of data being produced.** And, the repositories and archives that do exist operate in very different ways, employing widely varying definitions of what it means to provide data stewardship and what technologies are needed to achieve desired levels of curation. **There has been no scientific study of the strengths and weaknesses of current data stewardship organizations or operations.**

In developing policies and framework for stewardship of research data in an open access environment, agencies will need to invest in expanding and building upon the archival infrastructure that currently operates. Funding is needed for:

- Physical facilities, such as data centers for storage of large amounts of data
- Viable technological solutions for carrying out data stewardship processes for versions control, evolving file formats, discovery, and access

- Programs to assess the validity, competency, and track record of existing archives and repositories
- Programs to educate and train a skilled workforce

Physical data storage facilities should be funded to provide capabilities on a global level. Research is not only conducted from multi-disciplinary perspectives, it is also an international phenomenon; facilities for data generation, storage and access will need to take advantage of cooperative alliances with all stakeholders internationally. For example, investment in secure satellite storage and retrieval facilities is essential.

Given the volumes of data being produced, technological solutions to long term data stewardship are going to have a vital role in whether or not research data can be saved. Different kinds of data have different requirements for ensuring usability over the long term. In the social sciences, best practices for managing versions and file formats focus on techniques such as format migration and media refreshing. Ways of carrying out these processes mechanically are being explored by such groups as the ICPSR, the California Digital Library, and the Chronopolis/iRODS program, along with numerous academic institutions where smaller archives and repository software tools (such as Islandora or Stanford University's Hydrangea project) have been built. However, none of the new approaches thus far proposed or promulgated have been tested. There is no empirical evidence demonstrating that any one technical approach is any better than another. It is unclear whether micro-service or rule-based solutions are best and for which kinds of data are such systems most suited. There have been no assessments evaluating and comparing repository systems on a side-by-side basis to establish the strengths and weaknesses of each.

While there have been many claims made by many institutions that research data preservation and curation services are provided, none of them have been assessed by any standard criteria. Such standards have been developed, including the Trusted Repository Audit Checklist and the Data Seal of Approval. Open access cannot be accomplished until there is a careful investigation of the existing infrastructure, its capabilities, and whether or not organizations involved are achieving the desired goals for long term stewardship. Further, there is no agreed upon set of principles about which data needs to be maintained, in what way and for how long.

In a robust open access infrastructure, enabling workforce capacity is a serious concern. There are relatively few programs for training personnel to handle the variety of tasks involved in data management. Some look to the training provided through library and information sciences, but such programs are uneven and there is no established formal set of skills that all training programs should provide. None of these programs have ever been evaluated and there is no across the board accreditation process. Much of the time, 'training" involves reading numerous articles and reports and discussing issues from a philosophical or theoretical perspective. There is very little effort to ensure that new data archivists have the statistical, computer programming, or data science practical skill sets needed to perform in a variety of academic, private, and public agencies and firms. Current workforce capacity in commercial firms, banks, industrial occupations, medicine and physical and life sciences is lacking in the same skill sets. For open access to be a path toward innovation and economic strength a trained workforce is essential to such an infrastructure.

Responses to selected questions

Question 1
What specific Federal policies would encourage public access to and preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Stewardship of data includes the development of detailed metadata, the long term curation of data, and technical mechanisms to enable discovery of and access to data. Ensuring that researchers and the public alike are able to find and use existing data will contribute to increased productivity and will expand the opportunity to increase knowledge. We suggest, along with the National Digital Stewardship Alliance that policies of funding agencies need to "go beyond the data management plan, and should explicitly recognize "data under stewardship" as a core indicator of scientific effort and include this information in standard reporting mechanisms." Broadly stated, researchers who document and ensure the long term management of their data should be rewarded when future applications for funding are submitted.

Federal policy should also focus on the data stewardship infrastructure to ensure that data are maintained by trusted digital repositories (For example, see Beagerie, et al. Trusted Digital Repositories http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf). Not all archival operations function equally. Federal funding programs should support existing archives and repositories to become certified as trusted repositories or to meet the disciplinary data stewardship requirements of the data being managed. Further, the existing archival infrastructure should be studied at the national, state, local and institutional levels to ascertain strengths and weaknesses and to establish funding streams to upgrade and re-engineer older archival operations to make use of and/or develop new data management tools, software, equipment and data management facilities. Finally a study of the skills and training possessed and/or needed by current data management professionals should be made and a program to develop and train new professionals should be established.

Question 2
What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

This response pertains to data gathered in social sciences research. It is currently the custom that data collected by social scientists belong to the universities at which they are employed. However, sharing of data and making it publicly accessible is expected. Projects funded by National Science Foundation (NSF) and National Institutes of Health (NIH) now require data management plans with a view toward open access and sharing of data. In order to protect intellectual property interests, researchers have a right to expect that users of their data will be properly cited and that the original researcher will be given due credit.

Professional organizations such as the International Association for Social Science Information Services and Technology (IASSIST) promote the proper citation of data. Several organizations have developed methods for providing unique identifiers for data. The plethora of possible uniquely identifying systems

is confusing to researchers who do not know the difference between a DOI, URN, universal handle or some other kind of registry. Federal policy could support the establishment and use of a single unique identifier and system for storing and verifying such identifiers, or for building interoperability among domain or discipline specific systems. Research into what the best system and mechanisms to accomplish this should be should be carried out.

Question 3
How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on management of data?

Agencies can work with discipline-specific professional archival and data management organizations for ascertaining best practices in each discipline. Many practitioners have established best practices within their disciplines for managing data. There are also best practices for different file formats, including text, image, video, audio, simulation, game, and so forth. Data management plans call for researchers to specify the kind and format of data they will be collecting. At the same time, research at academic institutions is increasingly inter-disciplinary. Federal policies and data management plans must take into account the possible hybridization of data produced in such activities. There is no one solution which will address data in every discipline. Still, policies can focus on some common data management procedures for authentication, for version control, for establishing unique identifiers, and for following standards for establishing authorship, author rights, and verifying trusted repository or archival entities. Any policies established will need to be flexible enough to adapt to changes in research data gathering practices, technological advances, and changes in best data management practices.

Question 4
How could agency policies consider differences in the relative costs and benefits of long term stewardship and dissemination of different types of data resulting from federally funded research?

An assessment of the costs and benefits of long or short term stewardship of data produced with federal funds should be part of the process of review when a grant proposal is first submitted. Each agency that provides funds for research has a set of review criteria. For example, the National Science Board has recently release new criteria for proposals submitted to the National Science Foundation http://nsf.gov/nsb/publications/2011/06_mrtf.jsp. While these criteria do not specifically address short or long term stewardship as review criteria, (though perhaps they should) the perceived intellectual merit and broader impact of the intended research can provide guidance on when and how to allocate resources.

Further, the ability of the investigators (or the specified archive in which the data will be eventually placed), to disseminate data in user-friendly public access modes should be considered; many agencies conducting their own surveys now provide ways to produce simple tables and visualizations while at the same time providing the research community with access to the underlying, complete data used to produce the tables, charts and reports. The same consideration could be applied to federally funded data collections. This provides both a short term solution to data access, and still allows for in depth research to take place.

Cost and benefit of long term stewardship can be considered from other aspects. For example, not all data needs to be kept indefinitely. Not all data needs extensive maintenance. In some cases, merely providing what is termed 'persistent access' will suffice, but in other cases the data may be considered to be so valuable as to need full data curation. This can be file format dependent. Data produced in software dependent formats will need more attention over time to ensure that the data can still be used even if the original software used to produce the data no longer exists. The cost to do so needs to be factored into the grant proposal budget, and should be considered by those who review such proposals. In some cases, the data collection and its management will require that a data archivist is embedded into the project at the very beginning. Doing so can result in a final data product whose long term stewardship is less costly than it would be if data management is handled at the very end of the project or even some years after the completion of the project. All of these factors should be part of an evaluation of a grant proposal, and such an evaluation should be carried out by those informed about long term data stewardship practices and institutions.

Question 5
How can stakeholders (e.g. research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

The kind of detail requested in data management plans asks researchers to think about data outside of an analytical context. Unless the researcher has frequently re-used data from previous studies they are unaware of the kinds of data management activities needed throughout a research project as well as once the project is completed. The preparation of a data management plan generally requires a consultation between the researcher and the data archivist with knowledge about stewardship of the type of data being collected and the research methods being used.

The stakeholders need to be able to advise the researcher about best practices for data and file formats, software and computing technology, metadata development, and on documenting their datasets. Stakeholders can best contribute to implementation of data management plans by having fully trained staff, by having certified data management operations and by possessing knowledge about the research applications of the data they are hoping to manage. Grant proposal reviewers will need to look beyond the data management plan itself to verify that the repository or archive selected is actually able to carry out the plan specifications.

Question 6
How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

There is no easy answer to this question. Many funding agencies allow for proposals to include data preparation costs which are intended, in part, to ensure that the resulting data product will be usable by someone not familiar with the original project. Agencies should continue to provide funds for the preparation of data collected with federal monies, but such support should also reflect the cost of ensuring long term access. When archives receive data in a well-documented form, with datasets built according to disciplinary standards, it is possible to carry out appraisal and ingest with low cost and the long term curation can be straightforward. The higher costs come from trying to manage complex, large volume, multi-format and poorly documented collections. Data produced in proprietary formats,

or software dependent formats also raises costs. It is important that these issues be raised at the grant preparation stage, within the data management plan and that proposal reviews include an assessment of the data management plan and proposed costs.

Question 7
What approaches could agencies take to measure, verify, and improve compliance with federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The best compliance will come from developing the best plans. When researchers understand what is involved in data management during and after a project and are informed about the kinds of support and archival expertise available they will be better able to carry out their data management plans. When data archivists are involved in the development of plans, or are partners in the research there is greater likelihood of compliance.

One possible approach involves the assignment of a unique identifier to a fully processed set of research data. The unique identifier can be used as a way to *verify* that the data management plan has been carried out, and would be assigned by the archive or repository with responsibility for housing a researcher's data and reported to the funding agency. There are a number of such operations for obtaining a unique identifier, including the use of DOI's, URN', persistent identifiers and there are any number of locally designed registries. Federal agencies could require that researchers must use one certain system in much the way that publishers require bibliographic citations in specific order and format. The unique identifier accompanies the data and is referenced in publications and citations.

Question 8
What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Enabling innovative re-use of existing data should be one of the core criteria used to evaluate data management plans. In order for this to happen, the data have to be discoverable, the usability needs to be assessable, the data have to be accessible, and the potential of the data has to be made evident. Data discovery mechanisms are searchable catalogs or directories containing complete citations to existing data. Users can assess the utility of data by being able to evaluate data content and format using searchable detailed levels of metadata. Data needs to be stored in platforms that make it easy to obtain copies for analysis on the user's own computing facilities. These features exist for a small number of domain specific archives, such as the Inter-University Consortium for Political and Social Research (ICPSR), the Data Preservation Alliance for Social Sciences (DataPASS) and in some academically based small archives or repositories. The vast majority of data produced is not managed by such facilities and is therefore not easily publicly accessible.

Ability to assess the potential of data for innovation requires a skill set and area of expertise commonly referred to as data science. This is an emerging occupation and there are not enough trained persons who can carry out this kind of work. Numerous commercial enterprises are searching for this kind of expertise; some of these companies spoke about this at the 2011 Web 2.0 Summit. For example, Bluefin Labs http://www.web2summit.com/web2011/public/schedule/detail/21613 has managed to

build what they call a "data genome" using data science technology and expertise. Scholars could advance their research if they were able to partner with such firms to contribute to and take advantage of investments firms make in data science based initiatives.

The need for data science expertise is also required in disciplines that collect or produce large quantities of data. Technologies such as Hadoop, used to pull together and mine big data exist but are not always easily employed by researchers and their use is as yet underemployed in business applications. Companies such as Microsoft are trying to bring these tools to the average researcher (http://www.web2summit.com/web2011/public/schedule/detail/21995).

There needs to be an effort by the Federal government to promote and enable more industry-university partnerships so that the unique research data being collected can be made more accessible, used in more unique ways and that there is a newly trained set of professionals who can best stimulate the use of research data in innovative ways. Funding to establish data science training programs is needed. Funding to universities to implement and contribute to the development of data mining tools, in partnership with industry is needed as well.

Question 9
What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

There has already been considerable effort which is ongoing to provide standards for citing data and providing attribution. DataCite is an example of an internationally supported organization devoted to just this area of work. http://datacite.org/ Further, there are examples of how to create citations available from many professional organizations, archives and repositories. The issue is in getting researchers to use them. Many publishers now require a citation to data used in a publication as do scholarly societies with active publishing enterprises.

Question 10
What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Within the social sciences data community, a standard for describing the content of data files has been established through the Data Documentation Initiative (DDI) and this has evolved into an internationally accepted and implemented standard http://www.ddialliance.org/. DDI is not only a metadata describing standard, it also provides for machine actionable metadata enabling analysis of specific data elements across studies. Tools for employing DDI have been developed and continue to evolve; Colectica is an example http://www.colectica.com/ and there are numerous open source resources shared by the DDI community of implementers. DDI has been engineered to be interoperable with other metadata standards such as the Statistical Data and Metadata Exchange (SDMX) http://sdmx.org/.

Question 11
What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

DDI was a social science community driven effort and was adopted through the use of training in using the standard, in forming working groups of practitioners, and in having the funds to continue development. There was a recognized need for the standard to enable archivists and researchers alike to organize, manage, promote discovery of and reuse data collected in social sciences research. Further there is an ongoing desire to promote interoperability among metadata standards so that data from many disciplines, not just social sciences, can be mined and used. The standard will be able to evolve to accommodate changes in research strategies and methods; initially DDI focused on a metadata standard for surveys but work is ongoing to accommodate research projects collecting qualitative information, and to accommodate different types of data gathering instruments.

Another example of a standard is the Data Seal of Approval (http://www.datasealofapproval.org/). The Data Seal of Approval is a set of guidelines that archives and repositories can use to provide depositors with some guarantees that the archive follows the best practices for long term stewardship of data. It has been adopted by organizations internationally, including the ICPSR. The Data Seal of Approval was developed by a single organization, but its utility has been recognized by the data community. Professional organizations such as the International Association for Social Science Information Services and Technology (IASSIST) http://www.iassistdata.org/ recognize and support the use of the Data Seal of Approval by its members.

Question 12
How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies can look to professional organizations which are international in scope for coordinating standards. Groups such as IASSIST (http://www.iassistdata.org/) address issues on data management, national policies on data use, data sharing, and data access. Members provide context for emerging policies in different countries; issues on data preservation and stewardship are discussed and standards for best practices have evolved through the organization. IASSIST also provides training in data management, DDI, data curation and repository management.

Another internationally based archival effort was launched through the International Household Survey Network (IHSN) working in collaboration with the Accelerated Data Programme (ADP) (http://www.ihsn.org/adp/) This work is designed to help countries develop their data and statistical programs according to standards and best practices and to promote the use of their data. These newly formed archives use the DDI standard for metadata, and as described in their website "provide technical and financial support to survey data documentation and dissemination, and to the improvement of survey methods. Key outputs include the establishment of national survey databanks, and the establishment of national data collection standards to foster comparability of data across sources." Federal agencies could promote and finance similar programs within the U.S. and its territories at the state and local level.

Question 13
What policies, practices, and standards are needed to support linking between publications and associated data?

Work to enable data access from publications that report on data use and analysis is ongoing; organizations such as the ICPSR are engaged in working with publishers to provide links from publications to data managed at ICPSR. The National Digital Stewardship Alliance has taken a leadership role in promoting collaborative stewardship and standards with respect to linking of data to publications.

Currently publishers receive copies of data from researchers, but publishing companies will soon be unable to manage the storage of so much data. Further, publishing companies do not possess the skill set or facilities required for long term stewardship of such data, nor do the publishers offer any guarantees that stewardship will even be provided. Other issues to be addressed in making linkages between data and publications have to do with being able to maintain version control, to be able to update or correct datasets, and the archive or repository must be able to provide access indefinitely. The use of unique identifiers can aid in version control but publishers will need to be able to accommodate updated material. Requirements to properly cite the use of data provided with publications need to be made; investigators must be given credit not only for their publications but also for the data they produce and share.

The indefinite access to data in a form that is usable despite the passage of time requires that the archive operate according to the best practices for the data format and discipline of the data being managed. Repositories and archives need to be assessed according to criteria for evaluating the ability of the facility to maintain and provide access to the data. Examples could include the use of the Trusted Repository Audit Checklist (TRAC) based on the ISO 16363 Standard for Trusted Digital Repositories and the Data Seal of Approval. Data management plans could specify that publications and data will be linked and that data will reside in a verified archive or repository. Further, such repositories should be able to demonstrate that they have a valid plan for succession in the eventuality that the repository ceases operation or experiences a disaster. For example, the Data Preservation Alliance for the Social Sciences (DataPASS) provides a complete succession plan for the management of all data shared by alliance members.

Tue 12/20/2011 8:33 AM
NDSA Response to OSTP RFI - Public Access to Digital Data

Dear members of the OSTP,

On behalf of the National Digital Stewardship Alliance (NDSA) Coordinating Committee, I submit this response to the *Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research*. The 110 member Alliance represents a broad cross section of organizations committed to broadening access to our nation's expanding digital resource. It is developing and coordinating sustainable infrastructures for the preservation of digital content and advocating standards for the stewardship of digital objects. This community of practice around the management of distributed digital collections brings over 10 years of research and implementation experience to inform the topic of the call.

Thank you for the opportunity to contribute to this important work of the Interagency Working Group on Digital Data.

Sincerely,

Tyler Walters

Chair, Coordinating Committee
National Digital Stewardship Alliance

Coordinating Committee Members

Academic

Micah Altman, Senior Research Scientist
Institute for Quantitative Social Science
Harvard University

Helen Tibbo, Distinguished Professor
School of Information and Library Science
University of North Carolina at Chapel Hill

Tyler Walters, Dean, University Libraries
Virginia Polytechnic Institute and State University

Commercial Content

Gene Mopsik, Executive Director
American Society of Media Photographers

John Spencer, President and Co-founder
BMS/Chace

Federal Government

Blane Dessy, Director
Federal Library and Information Center Committee

Non-profit

Michele Kimpton, Chief Executive Officer and Co-founder DuraSpace

Kate Wittenberg, Managing Director
Portico

State Government

Amy Rudersdorf, Director of Digital Information Management Program
State Library of North Carolina

NDSA
NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

Submitted by the National Digital Stewardship Alliance (NDSA)

January 2, 2012

Introduction to the NDSA

The National Digital Stewardship Alliance (NDSA) was founded in July 2010 to extend work begun in 2001 by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress. The Alliance has over 100 members from educational institutions, non-profit organizations, businesses and local, state and federal government agencies, as well affiliations with international organizations. Its mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. [1] Members of the Alliance are taking action to preserve access to our national digital heritage by:

- broadening access to our nation's expanding digital resources
- developing and coordinating sustainable infrastructures for the preservation of digital content
- advocating standards for the stewardship of digital objects
- building a community of practice around the management of distributed digital collections
- promoting innovation
- facilitating cooperation between government agencies, educational institutions, non-profit organizations, and commercial entities
- fostering the participation of diverse communities and relationships across boundaries
- raising public awareness of the enduring value of digital resources and the need for active stewardship of these national resources.

Supporting communities of practice for preservation and access

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally-funded scientific research. When applied, these values support the practical development of communities of practice capable of gaining consensus to support preservation and access to digital data. The shared expertise and common experience of these communities result in stakeholder buy-in and adoption of policies and

standards. The National Digital Stewardship Alliance member organizations are bound as a community by the following values.

Stewardship. Members of the NDSA are committed to managing digital content for current and long-term use. The members of the NDSA are actively ensuring sustained access to the digital content that constitutes our national legacy and empowers us as leaders in the global knowledge economy. Individually, these organizations support the management of digital resources; the Alliance is committed to protecting our nation's cultural, scientific, scholarly, and business heritage.

Collaboration. Collaborative work is the centering value of the Alliance; it is a value shared by all members and a priority in work with all organizations and associations. Approaching digital stewardship collaboratively allows the NDSA to coordinate effort, avoid duplicate work, build a community of practice, develop new preservation strategies, flexibly respond to a changing economic landscape, and build relationships to increase capacity to manage content beyond institutional boundaries.

Inclusiveness. The NDSA is a collaborative effort to preserve a distributed national digital collection for the benefit of current and future generations. We value the range of experience, the potential for innovation, and the fault-tolerance that heterogeneity brings. We believe the preservation of digital information is a pervasive challenge and that engaging across different communities strengthens the nation’s digital preservation practices and increases the likelihood of preserving content now and into the future.

Exchange. Members of the Alliance encourage the open exchange of ideas, services, and software. This leverages the commitments of each member to increase the capacity of the entire stewardship network. Participation and engagement result in innovations and benefits that can be shared by all. The Alliance is committed to transparency and all products generated or produced by the Alliance will be circulated under open licenses.

Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

Community-based approaches to the challenges of rapid change and high volume within the data domain have proven to be the most successful in the long term. The Blue Ribbon Task Force on Preservation and Access recommended that for research data “Each domain, through professional societies or other consensus making bodies, should set priorities for data selection, level of curation, and length of retention.” [2]

The report validated experience over the last ten years of digital preservation work. A study of the networks developed through the NDIIPP program indicated that participating institutions bring to the network their own resources, interests, and organizational culture. Under the auspices of a neutral convener and honest broker, natural networks emerge over time through participation in shared activities and problem solving. As these networks form, the larger network becomes more complex, but also stronger and better able to withstand stresses and strains. [3]

The Opportunities for Data Exchange (ODE) project supported by the Alliance for Permanent Access and the European Union also takes a cross-cutting community approach to preservation and access to digital data. "The potential answers to grand challenges of our times require…the inclusion of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-infrastructures…All stakeholders in the scientific process must be involved in the design of this layer; policy makers, funders, infrastructure operators, data centers, data providers and users, libraries and publishers…" [4]

An exemplar of collaborative community efforts is the Dataverse Network project [5] recently described by the National Research Council of the National Academies as the "State of the Practice in Data Sharing." [6] The Dataverse Network is "unique in being designed to explicitly support long-term access and permanent preservation. To this end the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats." [7]

Foster public values and support for stewardship of digital data beyond mandating data management plans.

Policy should assert the value of research data and provide mechanisms to support the preservation, discoverability and access. To relieve frustration and confusion about actions the policy should provide a clear direction for funders, researchers and stewardship organizations. The Blue Ribbon Task Force recommended "Funders should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving. They should explicitly recognize "data under stewardship" as a core indicator of scientific effort and include this information in standard reporting mechanisms." [8]

Leverage substantial national and international efforts for common practices that support interoperability.

Substantial efforts have been made to pave the way for interoperability, re-use and re-purposing. Emerging practices for data citation, licensing and protocols for data sharing and sustainable re-use are becoming enough to adopt more broadly. Notable in these areas are work on the Data Seal of Approval by the Data Archiving and Networked Services that promotes sustainable access to digital research and provides training and advice about archiving and reuse.[9] LOCKSS is a community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content. [10] The

Data-PASS organization promotes collaborative, institutional stewardship of research data, permanent data archiving, and citation that permits results to be verified and re-purposed. [11] DataCite collaboratively addresses the challenges of making research data visible and accessible through data citation.[12] The Creative Commons project, Science Commons, has focused on protocols for sharing scientific data that includes licensing and mitigating legal barriers.[13]

Summary of Major Recommendations

- Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Establish policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Foster public values and support for stewardship of digital data beyond mandating data management plans.

- Leverage substantial national and international efforts for common practices that support interoperability.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI. The responses below exemplify how the principles can be applied to the individual questions, and highlight relevant NDSA activities in these areas.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The most effective policies in this regard would mandate data deposit into publicly accessible repositories. In the absence of such a policy, there are already cases of data which have been lost. The Federal policy framework should move public access to data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use.

Many members of NDSA provide repository services at low cost or through cooperative arrangements. Members of the NDSA also provide repository services that provide legal, technical, procedural and statistical controls necessary to protect data confidentiality while ensuring long. And the NDSA provides a model of institutional collaboration that supports stewardship, discovery and accessibility An example of a free access service is ViewShare.org, a platform for empowering curators, archivists, and librarians to provide access to the digital collections they are preserving through a shared interface. This

service provides the dual benefit of making data more broadly available and accessible while also making it easy for end users to copy and make use of the data in other environments. [14] The NDSA content working group is also working toward developing a clearinghouse for at-risk digital collections to help match data to potential preservation partners.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Each domain and discipline should be empowered to set priorities for data selection through, level of curation, and length of retention, through professional societies or other consensus making bodies.

Notwithstanding, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. For most data, "open access" is needed not only for the short term, but for the long term. And scientific disciplines have focused primarily on short-term access. There are critical standards for metadata exchange, fixity information and verification, and persistent citation that can support long-term access to data, preservation, and the long-term reproducibility of public results. Such baseline standards should be applied all scientific data. Among the range of important new standards for preservation and access there is still little knowledge about which standards are being implemented in which situations. The NDSA Standards working group is working on inventorying these standards and exploring how they are currently being used by NDSA member organizations. More than advocating the need for standards there is a clear need to understand which standards are being used in which situations and use that information to promote the usage of standards that are leading to results.

(5) *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., the Data-PASS archives). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation being used loosely without associated evidence of meeting specific requirements. Memory institutions such as archives, libraries and museums have an extensive track record with these functions and collaborative organizations such as NDSA could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions. In this respect, work from the NDSA innovation working group toward developing a "Neighborhood Watch" system for repository quality assurance could serve as the basis for establishing clear, externally verifiable reporting. [14]. The group has identified a pressing need for

an objective, repeatable, independently verifiable and simple way for an external agent to periodically retrieve content, verify its bit level integrity and publicly announce the results. This is a clear example of how assertions about data management could be tested and certified.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The most important step would be to communicate that the real costs of preserving and making digital data accessible are indeed legitimate and necessary costs of the overall research enterprise. Researchers routinely include publication costs within their research proposals -- the costs of ensuring long-term access reuse of data should be treated in the same way.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Many NDSA partners are leaders in this area. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

References

[1] The National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa

[2] Berman, Francine, and Brian Lavoie, et al. 2010. *Sustainable Economics for a Digital Plant: Ensuring Long-term Access to Digital Information.* Final Report of the Blue

Ribbon Task Force on Sustainable Digital Preservation and Access supported by the National Science Foundation, et al. Washington, DC: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Library of Congress. 2010. *Preserving our Digital Heritage: The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report*. Washington, DC: http://1.usa.gov/hmw2lj

[4] Alliance for Permanent Access. 2011. "Opportunities for Data Exchange (ODE) Project": http://www.alliancepermanentaccess.org/index.php/current-projects/ode/

[5] King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173-99.

[6] National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

[7] Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

[8] Berman et al. 2010.

[9] Data Seal of Approval: http://www.datasealofapproval.org/

[10] LOCKSS: http://lockss.org

[11] DataPass: http://data-pass.org/

[12] DataCite: http://datacite.org/

[13] Creative Commons project, Science Commons: http://creativecommons.org/science http://wiki.creativecommons.org/Science

[14] ViewShare: http://viewshare.org

[15] Abrams, S, Cruse, P, Kunze, J, Minor, D, Smorul, M. 2011."Neighborhood Watch" for Repository Quality Assurance. Presented at Designing Storage Architectures for Preservation, Washington, DC: http://1.usa.gov/uXj2Mf

Tue 12/20/2011 10:58 AM

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

To whom it may concern

My name is Andrew Vickers, PhD, and I am on the faculty of the Department of Epidemiology and Biostatistics at Memorial Sloan-Kettering Cancer Center. I have a long interest in data sharing in medical research. My scholarly papers include: the rationale for data sharing (Trials. 2006 May 16;7:15: http://www.trialsjournal.com/content/7/1/15); an empirical study of data sharing (PLoS One. 2009 Sep 18;4(9):e7078: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007078); guidelines on how to share medical data (BMJ. 2010 Jan 28;340:c181: http://www.bmj.com/content/340/bmj.c181?view=long&pmid=20110312); an extensive set of raw data and code from a series of studies on radical prostatectomy (BMC Res Notes. 2010 Sep 2;3:234; http://www.biomedcentral.com/1756-0500/3/234) and an editorial on the policies of funding agencies (BMJ. 2011 May 4;342:d2323: http://www.bmj.com/content/342/bmj.d2323?view=long&pmid=21543405). I was also the author of a widely cited *New York Times* op-ed on data sharing (http://www.nytimes.com/2008/01/22/health/views/22essa.html).

I would like to address the first question about the specific Federal policies that would encourage public access to data arising from Federally-funded research. My own view is that this has to be a policy with teeth. We don't want to simply repeat the history of PubMed Central: this was originally voluntary, but compliance was dismal and so was then made mandatory. All researchers now know that, if they want to get more grants from the National Institutes of Health, they have to deposit papers to PubMed Central and prove that they have done so when they submit their next grant. Why not have something similar for raw data? I would propose that, if you want another grant, you have to prove that the raw data produced from your last grant is publicly available on a registry. Naturally, researchers could ask for waivers, just in the way that you can exclude women or children from clinical trials if there is a good reason to do so. For example, if a researcher can make a clear case that sharing data would pose a genuine threat to privacy (for example, genetic studies on unusual populations) , then that research could be exempt from the data sharing requirement. Moreover, researchers could request a reasonable period of time to exploit their data. For example, researchers could state that their raw data were deposited at a particular registry, but that the data will not be accessible for two years after the first publication describing the study results. Note that the proposal is not for a vague "data sharing plan" (which, could after all, be "we will evaluate your request and then refuse it"), but for mandatory depositing of raw data into a publicly accessible archive.

It is clear that there would be some technical obstacles to such a proposal. For example, how would registries to accept raw data be organized and who would run them? Just how "raw" should raw data be. However, it is clear and obvious that such obstacles are far from insurmountable and that they could be solved by methodical planning.

As regards question 9, attribution and credit, please note that this an issue that I have dealt with in the peer-reviewed literature (Trials. 2006 May 16;7:15). In that paper, I proposed some guidelines for conduct of investigators using raw data collected by another team and for journals publishing such data. In the following "independent investigator" is the individual wishing to publish a reanalysis of published raw data; a "trialist" is the individual who helped gather the data in the first place.

Code of conduct for independent investigators and journals

1. Independent investigators planning to *publish* a new analysis should contact the trialists before undertaking any analyses
2. One or more trialists should be offered a co-authorship on any resulting papers

3. If trialists disagree with the methods or conclusions of a new analysis:

a. They should not have veto power, unless this was agreed beforehand by the independent investigators

b. They should, however, be guaranteed the opportunity to write a commentary to be published alongside the new analysis

4. Journals should not publish new analyses of previously published data unless either a trialist is an author or a separate commentary from a trialist is attached

5. Published new analyses should cite the original trial

I would be delighted to respond to any questions or comments on these thoughts

Andrew Vickers

Memorial Sloan-Kettering Cancer Center

Wed 12/21/2011 4:07 PM

Response to RFI

I would like to fully endorse the comments made regarding the open sharing of data across scientific disciplines as advocated in the by the National Digital Stewardship Alliance in this document :

http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf

As a member of the Department of Energy community I fully believe access to and sharing of scientific data is critical for leadership and health of science in the United States.

Thank you for your consideration,

Miriam

<><><><><><><><><><><><><>

Miriam Blake
Research Library Director
Los Alamos National Laboratory

Phycological Society Of America

c/o Susan H. Brawley, President
5735 Hitchner Hall
School of Marine Sciences, University of Maine
Orono, ME 04469, USA
Telephone: 207-581-2973 Fax: 207-581-2801 brawley@maine.edu

December 21, 2011

The Office of Science and Technology Policy

To Whom It May Concern:

The Phycological Society of America (PSA) appreciates the opportunity to respond to OSTP's Task Force request for information from scientific societies regarding access to digital data from federally-funded research, as described in the Federal Register solicitation of 4 November 2011 per Section 103 (b)(6) of the America COMPETES Reauthorization Act of 2010 (ACRA). "Phycology" refers to the study of algae. The PSA is a non-profit scientific society that is incorporated in the State of Maryland and was founded in 1946 to advance research and education in all aspects of algal science. Our membership of about 1000 scientists and graduate students do research in universities, industry, state and federal government, and NGOs; about two-thirds of our members work/study in the US. Our members study a remarkable range of important topics, from key aspects of global carbon cycles to important health issues.

The PSA's *Journal of Phycology* is entering its 48th year (2012) of publication of basic and applied research on algae. PSA owns copyright to current and full back issues of the print and electronic journal (6 issues/year), and PSA controls all editorial decisions and content of our journal, which has been published since 1999 in association with Wiley-Blackwell Publishers, the publisher of the journals of about 283 other US scientific societies. We have found our association with Wiley-Blackwell (W-B) to be particularly valuable in terms of early establishment of the electronic version of the *Journal of Phycology* (begun in 1999). As with many other journals, extensive sets of digital data and detailed methodological information are now made available with the published article as electronic supplements. This particular response concerns the request by the OSTP task force for comment on how to increase "preservation and dissemination of broadly useful digital data resulting from federally funded research".

Question 1: What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federal funded scientific research, to grow the US economy and improve the productivity of the American scientific enterprise?

The Digital Object Identifier (DOI) system that was developed by the publishing industry through the not-for-profit CrossRef system now holds over 50 million DOIs of articles that increase interoperability and accessibility. We are aware from our publishing partner Wiley-Blackwell that discussions have begun about the possibility of assigning separate DOIs to supplementary electronic material, including digital data. This would appear to us to be useful in encouraging public access and preservation of these digital data in published articles, especially because the DOIs could be linked to the sponsoring grants on agency websites, if desired.

PSA has experienced some loss of supplementary videos from the earliest years of electronic publication of the *Journal*, as software operating systems of publishers were changed and upgraded. The development of DOIs for digital data seems likely to be helpful in assuring successful transfer of data in future digital innovations in publication. The federal government might establish a working group with publishers' representatives to address the general, long-term preservation issue for electronic data (Paper is archival; is any electronic publication really permanent?), and ensure that the federal government is sponsoring basic research that will lead to permanent archive mechanisms.

We note that sets of data from federally funded research are likely to be most useful after they have been peer-reviewed along with the article submitted for publication that is based on their analysis; peer review results in a value-added product that is copyrighted. In our case, expert reviewers and editors in our field donate their time to peer review; this results in modest profit that is devoted entirely to support for graduate student research, tuition to summer field courses dealing with algae, etc. (i.e., building US human capital in science); support for annual meetings of the Society (travel expenses of symposium speakers); public outreach to K-16 educators; and periodic, special supplements or innovations at the *Journal*.

Agencies might develop common formats for certain sets of digital data (e.g., Long-term Ecological Research [LTER] sites) that could make it possible for raw data (lacking peer review) to be valuable when archived by a federal agency, and such data could be required of the investigator by project's end. The NSF, for example, now requires data management plans as part of grant applications. Even in cases, however, such as GenBank at the National Center for Biotechnology Information (NIH), where archive of similar digital data prior to---or without---publication is possible, lack of peer review ultimately makes large collections of such data less valuable.

Question 2: What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded research?

The most important step with respect to the intellectual property rights of society publishers of science is to recognize that an expert community of scientists in the area of the digital data will provide strong peer review that adds value to the digital data. The peer review has cost, which leads to the article and accepted digital data being protected by copyright, which must be protected by the government.

Question 9: What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

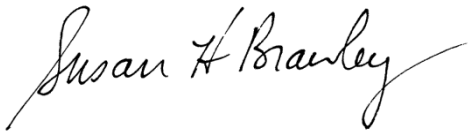
This requires standard professionalism and honesty of the secondary user. DOI tags will help the secondary user report the source(s) of their analyses, and the peer review system is the best safeguard against uncredited use of other investigators' data.

Question 13: What policies, practices, and standards are needed to support linking between publications and associated data?

Cooperation between scientific publishers to establish DOI tags for associated data would immediately make this linkage possible.

These responses were reviewed by the PSA's Executive Committee. We would be happy to respond to further questions.

Sincerely yours,



Susan H. Brawley
President

CAL POLY LIBRARY SERVICES
ROBERT E. KENNEDY LIBRARY

TO: Office of Science and Technology Policy

FROM: Anna Gold, University Librarian, California Polytechnic State University Library Services, San Luis Obispo, California

DATE: January 12, 2012 (revised response)

SUBJECT: Response to RFI: Public access to **digital data** resulting from federally funded scientific research.

About this response

The following information was prepared by the Kennedy Library at California Polytechnic State University in San Luis Obispo, California, in response to the request for information issued November 3, 2011, by the Office of Science and Technology Policy.

Information was primarily provided by Marisa Ramirez, Digital Repository Librarian at Kennedy Library, in consultation with Timothy Strawn, Director of Information Resources and Archives, and University Librarian Anna Gold, with additional input provided by David Beales, Associate Librarian for Engineering at Kennedy Library.

The Robert E. Kennedy Library and Cal Poly Library Services at California Polytechnic State University (San Luis Obispo) provide a comprehensive program of library services within a teaching-led comprehensive public polytechnic university, including access to scholarly and professional information, data services, and a digital institutional repository that recently passed the 1.3 million download mark. The Library's Data Services Team is actively working to define, develop, and sustain a library program of data and GIS services that contributes to Cal Poly's mission and programs of learning and research.

The mission of Cal Poly Library Services is to promote open and informed inquiry, foster collaboration and innovation, support the unique needs of every student and scholar at Cal Poly, and contribute to the cultural life of our community. In common with other libraries in higher education, the Library is in a unique position in our organization to do this through access to technologies that support the creation, reuse, sharing and preservation of new knowledge.

Background

The U.S. government funds tens of billions of dollars in basic and applied research each year, with the goals of speeding the pace of scientific discovery, fueling innovation, and improving the public good. At its core, the most significant research findings are supported with underlying data sets, often collected in digital form on a large or small scale.

Traditionally, libraries have served as steward of the record, and this continues to be true in the digital realm. In the last five years, there have been an increasing number of hires in the Library and Information Science (LIS) fields for data services librarians, data curators, digital curators, and digital repository services librarians. In this era of data-driven science, academic and research libraries are strategically repositioning

themselves to align more closely with their institutions' research dissemination strategies. Academic and research libraries are considering digital curation issues broadly, across all subject disciplines, in an effort to share information as a community and work together to determine best practices and standards.

A host of parties have interests in the curation of digital research data, and academic libraries are well-positioned to represent their interests in access, outreach and advocacy, information management, digitization and digital technologies, support for teaching and learning, and preservation. In collaboration with other campus partners, libraries also have a role to play in developing strategies to capture, collect, manage and preserve data streams created by faculty. Sharing and coordination of these efforts can contribute to the overall data management and preservation efforts.

The following further comments on the questions posed in the RFI are organized into two interrelated themes: preservation, discoverability and access; and standards for interoperability, reuse, and repurposing.

1. Preservation, Discoverability, and Access

Cal Poly Library Services recommends four major approaches to these interconnected challenges:

1.1 Develop a national infrastructure based on existing models.

Below are several models that we recommend be explored:

- **Australia National Data Service (ANDS)** has created the Australian Research Data Commons to support initiatives designed to enhance existing data creation and capture infrastructure commonly used by Australian researchers and research institutions. This will ensure that the data creation and data capture phases of research are fully integrated to enable effective ingestion into a local data and metadata store published through Research Data Australia. This integration enables researchers to contribute descriptions of data to the Australian Research Data Commons directly from the lab, instrument or fieldwork site. It also ensures that higher quality metadata, critical for reuse and discovery, is produced through automated and semi-automated systems.

- **Joint Information Systems Committee (JISC)** is a United Kingdom funding body that supports research by providing leadership in the innovative, shared use of information and communications technologies and infrastructure to support education, research and institutional effectiveness. JISC offers support at local, national and international level by creating and supporting shared resources, knowledge, expertise and services, particularly where it gives rise to immediate cost savings.

- The **Digital Curation Centre (DCC)** is the United Kingdom's leading hub of expertise in curating digital research data. Launched in 2004 by JISC, the DCC provides a national centre for solving challenges in digital curation that could not be tackled by any single institution or discipline. The DCC is responsible for developing resources, training opportunities, and funding projects that promote the development of innovative methods for the preservation, discoverability, and access of research data.

- **DRIVER i**s a pan-European effort whose primary objective is to create a cohesive, robust and flexible infrastructure for digital repositories, offering sophisticated services and functionalities for researchers, administrators and the general public. Aimed to be complimentary to GEANT2, the infrastructure for computing resources, data storage and data transport, DRIVER delivers resources that result from

scientific output, including scientific/technical reports, working papers, pre-prints, articles and original research data. The vision is to establish the successful interoperation of both data network and knowledge repositories as integral parts of the E-infrastructure for research and education in Europe.

1.2 Utilize and encourage integration of current sources of expertise in the library, archives and records management fields.

The library profession has many professional organizations devoted to exploring, adapting and implementing emerging digital curation services, technologies, and infrastructures, whether repository-based or platform-agnostic, born-digital or digitized, for the lifecycle management of research, scholarship and other academic activities.

We recommend that the federal government leverage these professional organizations to assist in developing methods to curate a variety of content in digital form; to use scalable, efficient, and sustainable methods to inform and educate librarians on digital curation trends and new technologies; and finally, to collaborate with other organizations within the library profession and academe on issues concerning digital curation. Such library organizations include, but are not limited to:

- Association of College and Research Libraries (ACRL), specifically the Digital Curation Interest Group and SPARC;
- American Society for Information, Science and Technology (ASIS&T), specifically the Digital Libraries Interest Group and the Research Data Access and Preservation Group;
- Association for Library Collections and Technical Services (ALCTS), specifically the Preservation & Reformatting Sections including the Intellectual Access to Metadata Interest Group, Digital Conversion Interest Group, Digital Preservation Interest Group;
- Association for Information and Image Management (AIIM), specifically the Electronic Records Management section; and
- Society for American Archivists (SAA), specifically the Electronic Records Section.

1.3 Cultivate a workforce capable of addressing the new challenges posed by data curation and cyberinfrastructure development.

Expanding current data curation and cyberinfrastructure activities and embarking on new ones will require investment in professional development for library staff, and in some cases the creation of entirely new positions. Funding should go towards identifying new facets of library graduate education and subsequent professional development to prepare librarians to support data curation and cyberinfrastructure activities.

A challenge in identifying suitable models is to provide sustained, practical professional development opportunities suitable for working professionals, and not limited to campus-based residential programs, though these also have an important role to play in fostering the knowledge, experience, and skills to contribute to data curation and cyberinfrastructure activities.

1.4 Integrate and universally adopt existing mechanisms to educate faculty regarding copyright and intellectual property, and improve compliance with federal data stewardship.

We suggest exploration and possible adoption of the following models and approaches:

- **United Kingdom Intellectual Property Office Audit Model, and the Hargreaves Review.** In the United Kingdom, a key process in managing the intellectual property stemming from research is to conduct an Intellectual Property audit. The purpose of conducting an audit is not simply a stocktaking exercise, but instead it is undertaken to further exploit the intellectual property assets in hand, to implement procedures to minimize the risk of litigation by infringing others' copyrighted material as well as an opportunity for researchers to ask questions they may have about Intellectual Property laws and for the University to identify areas of further education.

  In November 2010, the UK Prime Minister commissioned an independent review by Ian Hargreaves and a team of consultants of the UK's intellectual property framework. The review made ten recommendations designed to ensure that the UK IP system promotes innovation and growth in the 21st century, both nationally and internationally. The United States may consider adopting elements from the UK Audit Model as well as conduct a study to determine how to realize efficiencies within the existing IP system.

- **Develop tools to assert author rights to research data.** High-impact academic publishers such as Nature Publishing Group are now requiring authors to deposit their raw research datasets with them as part of the peer-review process. This raises concerns about intellectual access to the raw data: journals have traditionally required authors to sign over intellectual property rights in exchange for getting published, and may well extend these terms to the raw data. Criteria or tools must be developed to help authors assert their intellectual property rights to their research. Failure to do so could result in stifling academic creativity and intellectual progress.

- **Develop criteria to guide academic publishers' policies on embargo periods.** The purpose of an embargo is to protect the revenue interests of the publisher, but it is generally considered frustrating to academic researchers who rely on current publications to further their work. In essence, the publisher's embargo stifles academic creativity and intellectual progress. As publishers increasingly require raw datasets from authors in order to publish scholarly articles, a further concern is that these embargoes may be extended to limit access to research datasets. A more equitable balance must be reached, based on established and publicly available criteria, to better guide academic publishers' policies on embargo periods. Furthermore, timely deposit of research data in open disciplinary repositories as may be frustrated by a publisher embargo. It is more desirable that data be openly accessible without embargo, with deposit in a disciplinary or institutional repository to be preferred over deposit with publishers. This would not prevent publishers from requesting that authors provide data as part of the peer review process, nor prevent publishers from linking to that data for published articles, using community-based citation standards.

- **Investigate ways to integrate attribution initiatives into reporting systems to ensure compliance with Federal data stewardship policies. Ensure that appropriate attribution is provided to the creators of data by promoting methods such as:**

- **Open Researcher and Contributor ID (ORCID)**. ORCID aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community.

- **ResearcherID**, a multi-disciplinary scholarly research community developed by Thompson Reuters' Web of Knowledge, which assigns a unique identifier to each author to eliminate author misidentification while simultaneously adding dynamic citation metrics and collaboration networks to an author's profile.

- **Use existing academic channels to educate, verify and improve compliance with Federal data stewardship and access policies for scientific research.**

  Such channels include:

  - **Institutional Review Boards (IRB) and Institutional Animal Care and Use Committees (IACUC).** These are ethical committees formally designated to approve, monitor and review research involving humans and animals. Federal regulations have empowered these committees to approve, require modifications in planned research prior to approval, and to perform critical oversight functions for research conducted on human and animal subjects. Most, if not all, research institutions have one or both of these committees. These committees often require researchers to complete educational modules, such as the CITI Program, which is a service providing research ethics education to all members of the research community. A similar online program could be developed for data stewardship and could be required by local IRB and IACUC committees as a condition of local research approval.

  - **Campus departments such as Research and Grants Development Offices.** These campus offices can verify that grant applicants are including data management costs in grants. Grant applicants are typically required to report back to a central campus office on disbursement of funds and requirement compliance. Utilize this existing framework by leveraging their current activities including verifying grant compliance.

2. Standards for Interoperability, Reuse and Repurposing

It is suggested that metadata standards generally are most usefully considered within the limits of their user communities' standard practices. So long as they are XML-based, there is a useful degree of interoperability; the trend towards interoperable schema will continue while it is still useful. However, librarians are aware that any effort to maintain quality metadata standards is difficult. Metadata schema for data that are not directly related to the needs of the disciplinary community of interest are unlikely to be embraced wholeheartedly.

2.1 Recognizing the important role of disciplinary communities in developing their own descriptive and administrative metadata standards, and recognizing the powerful potential of XML-based and semantic web

approaches to assuring interoperability, it will also be useful to continue to exploit existing national and international structures to develop and promote common standards. These structures include:

- **ISO (International Organization for Standardization)** is a network of the national standards institutes of 162 countries and is the world's largest developer and publisher of International Standards. Because ISO wields influence in both the public and private sectors, this organization has the ability to form consensus on solutions that meet both the requirements of government, education and the broader needs of society. ISO, for example, may be a relevant standard for the registry, management, sharing, and delivery of research data across digital repositories and discovery services.

- **Cross-disciplinary metadata initiatives, such as the Dublin Core Metadata Initiative (DCMI), and web content interoperability standards, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE).** The DCMI provides core metadata vocabularies that are widely used in digital repositories to support management, discovery and interoperability of resources, such as Darwin Core. Darwin Core is a stable and versatile standard that facilitates the discovery, retrieval, and integration of information about modern biological specimens and their supporting evidence housed in digital or physical collections. The Open Archives Initiative has its roots in the open access and institutional repository movements, and has developed widely used interoperability standards (such as OAI-PMH and OAI-ORE) that aim to facilitate the efficient dissemination, description and exchange of content.

2.2 *Develop permanently funded tools that enable wide scale registering and verification of data repositories.*

Two examples of these include:

- **DataBib**, a grant-funded project by the Institute of Museum and Library Services, aims to create a community-driven, annotated bibliography of research data repositories. Once funding ends, however, it is unclear how this resource will be maintained. Ideally, this would be a project that could be developed externally, but then adopted and permanently manage by public sector stakeholders.

- **OpenDOAR**, a directory of open access academic repositories, is a current example of how best to develop a single comprehensive, authoritative list which requires registration, verification and harvesting of data repository metadata.

Thu 12/22/2011 2:17 PM
Comments on RFI: Public Access to Digital Data

To whom it may concern:

I'd like to submit the following comments regarding the Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research. I write representing only myself, as a professional librarian working to assist researchers with data management and data management planning. My views do not necessarily represent those of my employer.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

- Requiring data management plans is a good start. More specific guidance and stronger encouragement to openly share at least a portion of a project's data would help.
- Funders SHOULD support infrastructure for the data resulting from the research they fund, especially the NSF, which provides less of this support than other federal agencies (see the the NSB 2005 report on long-lived digital data: "Participants agreed to a considerable extent on the main policy issues, even though there is one stark difference between NSF and many other agencies: the vast majority of long-lived data collections supported by the NSF are managed by external research organizations, while other agencies, such as the National Aeronautics and Space Administration (NASA) and the National Oceanographic and Atmospheric Administration (NOAA) focus more heavily on archiving and curating many such data collections themselves.") Not adequately supporting domain-based services forces researchers in domains without appropriate data centers to adopt ad hoc approaches that may not serve well in the long run. Even those in a domain with a data center may find their data is not a 'fit' - for example, CUAHSI's HIS only accepts georeferenced data, yet researchers may do lab-based hydrologic research. Existing data centers, particularly those that receive federal funds, should be encouraged to accept and curate a very broad range of data in their disciplines.
- That said (funders should support infrastructure), there will likely always be cases where some data sets do not fit into existing infrastructure, the policies and management of existing infrastructure are not aligned with a data producer's needs, or a researcher or institution has valid reasons for wishing to manage the data "closer to home" (i.e. in their home institution). This should be permissible.
- Funders should make public the specific criteria by which data management plans in grant proposals are evaluated.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

See my comments on (1) - funded, domain-based data centers with an ongoing mandate would serve well to ensure long-term access to date. In addition, the long-term value of a particular digital data set may not be known until well after its creation. Establish processes to identify and "promote" data that are recognized as worthy of preservation, and do not burden the researcher with the cost of subsequent preservation when value of their data sets becomes evident long after their creation. The NSB's 2005 report (referenced in (1)) describes three categories of data collections, research, resources, and reference, based primarily on the breadth of the user community. The report that notes that collections may evolve and change categories; it's also possible individual data sets might move from one category to another. We don't have a good way to make this happen.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

There are several roles that research libraries can play in this arena:

- Connect researchers with campus and external services.
- Manage data resources locally, when appropriate.
- Collaborate on the development and operation of domain-based infrastructure.
- Promote and support the use of existing infrastructure. Assist researchers in identifying appropriate infrastructure and data repositories, advise on best practices for preparing data and metadata for deposit, etc.
- Participate in the development of standards.
- Maintain a current awareness of institutional and funders' policies and assist researchers in meeting them.
- If sufficient expertise is available on staff, train graduate students on new requirements and best practices for meeting them.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Specify periods of data retention in data sharing policies (where that isn't already the case) so researchers have clear guidelines to inform budget planning.

Best regards,

Gail Steinhart
Research Data & Environmental Sciences Librarian, Albert R. Mann Library (mailing address)
Fellow, Digital Scholarship & Preservation Services, Cornell University Library
Cornell University
Ithaca, NY 14853

AIP American Institute of Physics
One Physics Ellipse, College Park, MD 20740-3843 H. Frederick Dylla, Executive Director and CEO

22 December 2011

Submission for the Record: **Response to November 4, 2011 Federal Register Notice of Request for Information, OFFICE OF SCIENCE AND TECHNOLOGY POLICY, Public Access to Digital Data Resulting From Federally Funded Scientific Research; FR Doc. 2011–28621**

Submitted by: H. Frederick Dylla, Executive Direct and CEO, American Institute of Physics
Tel. +1 301-209-3131; Dylla@aip.org

Electronically submitted to: digitaldata@ostp.gov

The American Institute of Physics (AIP) appreciates this opportunity to submit comments and would be delighted to continue working with OSTP and other federal partners through a process of active engagement.

About AIP

The American Institute of Physics (AIP) is a 501(c)(3) not-for-profit membership corporation created in 1931 for the purpose of "the advancement and diffusion of knowledge of the science of physics and its applications to human welfare." AIP is an organization of 10 physical sciences societies representing more than 135,000 scientists, engineers, and educators. As one of the largest publishers of scientific information in physics, AIP employs innovative publishing technologies and offers publishing services for its Member Societies. AIP's suite of publications includes 15 journals, three of which are published in partnership with other organizations; magazines, including its flagship publication *Physics Today*; and the AIP Conference Proceedings series. AIP delivers valuable resources and expertise in education and student services, science communication, government relations, career services for science and engineering professionals, statistical research, industrial outreach, and the history of physics and other sciences.

Enabled by Internet technologies, AIP disseminates more information, more widely and more affordably, than ever before in its history, reaching more authors, subscribers, and users than ever before. This accomplishment requires heavy investments in technology and infrastructure (such as an online platform) and business-model innovation to deliver the option of free or low-cost access: open access, pay-per-view, or article rental, recognizing that the value of the final published article needs to be paid for to remain sustainable.

Introduction

AIP's highest goal is to achieve the widest possible dissemination of the research results it publishes, including any pertinent associated data and context information. As a scholarly publisher, AIP believes that better discoverability and reuse of original research data are to be encouraged at all levels and among all stakeholders. AIP also believes that data resulting directly from federally funded scientific

research should be made freely available in a sustainable manner and that this is best achieved through appropriate policies that leverage public-private collaboration.

AIP believes that it would be in the best interest of the United States and its government, as well as in the best interest of all other stakeholders, to strike a balance between public access and sustenance of the scholarly publishing industry because of the impact and value it brings to the progress of science and its contributions to American society and economy. Such a balance can be achieved based on shared principles such as the importance of peer review, the recognition of economic realities through adaptable and viable publishing business models, the need to ensure secure archiving and preservation of scholarly information, and the desirability of broad access. Policies should recognize that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost, which, in certain circumstances, the host should be entitled to recover. One way to achieve this balance is for government to adopt a sensible, flexible, and cautious approach to drafting public access policies—an approach that engages all concerned parties, including federal agencies, scientists, university administrators, librarians, publishers, and the public.

Consistent with the recognition of economic realities, it is AIP's position that government agencies should develop their public access policies through voluntary collaborations with nongovernmental stakeholders, including researchers and publishers. Any policies should be guided by the need to foster interoperability of information across multiple databases and platforms. Agencies' efforts then could be directed toward facilitating cyberinfrastructure and collaboration programs with and between agencies and the stakeholders to develop robust standards for the structure of full text and metadata, navigation tools, and other applications to achieve interoperability across the scholarly literature. More detail on this is provided later in the document. AIP believes that any scholarly publication access policy needs to be flexible to accommodate agency-specific needs and have the capacity to evolve in response to the rapidly changing nature of scholarly publishing.

AIP specifically recommends that federal grants set aside funds to support researcher data management and deposit efforts. Federal agencies could also play a role in supporting and encouraging the establishment of discipline-specific data archives where these are currently lacking. The amount and type of support should be determined in collaboration with key stakeholders involved in the deposit, storage, and preservation of data.

Federal policies should also focus on supporting and encouraging the development of community standards for the citation and reuse of data sets, thereby facilitating the creation of a system that gives researchers an incentive to share data resulting from federal grants.

AIP Responses to RFI Questions

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

We would make the distinction that it is not "public access" in the broadest sense that is important but rather access by other scientists who can use the digital data for the further advancement of science.

As data are not copyrightable, policies about access become policies about deposit by the data owner or proxy into an accessible system. It should be noted, though, that any policies should recognize and take into account differences between 'databases' (information products created for the specific display and retrieval of data) and 'data sets' (sets or collections of raw relevant data captured in the course of research or other efforts). Policies could require that data generated from federally-funded research be deposited in a certified and openly accessible repository; furthermore, researchers could be encouraged to make these deposits upon submission of their first manuscript showing results that were based on the data set. Although some agencies already have a preservation/access role (for example, DOE Order 241.1B), AIP agrees with the Interagency Working Group on Digital Data that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." Agency policies should support and encourage such a distributed system for both access and preservation; that is, policies should recognize and build upon the broad set of capabilities that exist for both access and preservation within the library and publishing communities for both documents and data – Portico, LOCKSS.

The integrity of preserved data would also need to be taken into account and supported by any policy.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

All policies should comply with current copyright and patent law. Data should be embargoed to the principle researcher until conclusions drawn from the data can be published in the research literature. An additional maximum embargo of one year would also provide for the filing of patents by the grantees (or their institution) as allowed by many, if not all, funding agencies (HR 1249 Sec 102(b)(1)(A)). See also the distinction between databases and datasets as addressed response to question 1.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Differences between scientific disciplines and different digital data must be taken into account by domain experts at the time of proposal review (note the language used in the Data Management Plan FAQ's of NSF in a variety of instances: "to be determined by the community of interest through the process of peer review and program management.") Only such experts will be able to determine if the data to be generated by the proposed research will be of longer term value to the scientific community of interest and if its type conforms to acceptable community standards.

Metadata—data about the data—which would include information both about what the data is and how it was collected, is addressed further in this response.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Policies must first recognize that not all data is worth preserving. Every type of data should be assessed regarding long-term stewardship. Policies would have to take into account not just the size of the datasets but also long-term usability, which depends on the rate of technology change, and level of documentation required. Along with the data, enough information needs to be preserved to reproduce the dataset. As noted in the answer to question 3, agencies will need to call upon data experts as well as scientific experts.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

There needs to be an interconnected system for access to and sharing and preservation of data based on community-developed standards and best practices. The system needs to encourage innovation and must support multiple solutions—data as an information resource is inherently more complicated than scholarly articles. Each stakeholder will then need to contribute based on their specific skills and expertise. Libraries, through Institutional Repositories, could take on a stronger preservation role. Publishers have been adding value to the research process and providing access to and preservation of the scholarly literature for hundreds of years and could extend this to data, well beyond current support for supplemental material. Universities and research institutions have both scientific domain knowledge and data and information experts. Any system will need to preserve incentives for innovation.

Consider, for example, work being done by the Data Preservation Alliance for Social Sciences through their partnership with the Library of Congress, LOCKSS, and Dataverse to prototype a policy-based replicated data archive.

Other examples include:

- linking between datasets and their resulting scholarly publications based on community-accepted standards, thus ensuring datasets become part of the scientific literature;
- Having clear standards and guidelines for the certification and auditing of data repositories; encouraging a system that incentivizes data repositories to maintain the accuracy or integrity of the data once it has been deposited;
- Incentivizing the deposit of datasets and ensuring that the administrative burden this imposes on researchers minimal.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Require data management plans and coordinate plan requirements across agencies and to community standards (see the Open Archive Information System Reference Model – ISO standard 14721:2003). What constitutes data that needs to be preserved should be clearly identified through the process of peer review and program management. Preserving and disseminating digital data should then be considered “part of the cost” of funding and doing research, not “an additional cost”. Funding agencies could emphasize that proposals must take into account data fit for reuse and preservation. Again, this

should be the approach across agencies. Research labs/institutions/university overhead rates would need to include cost of data preservation.

As pointed out in the final report from the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (*Sustainable Economics for a Digital Planet*): "Policy mechanisms can play an important role in strengthening weak motivations" as there is often "misalignment of incentives between communities that benefit from preservation (and therefore have an incentive to preserve), and those that are in a position to preserve (because they own or control it) but lack incentives to do so."

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

If data is created in the course of federally-funded research, then the funding agency could require that any such data deemed to be "preservation data" be deposited in a recognized archive. Through direct agency involvement in creating a "comprehensive framework for data access and preservation" based on community-accepted standards and best practices for data citation and reuse, agencies would maintain lists of certified repositories. Certified repositories could be similar to the data center members of the DataCite organization (of which DOE's Office of Scientific and Technical Information is a member) or participants in the SafeArchive program of Data-PASS. In addition, grantee data management plans could be required to identify all datasets expected to be produced from funded work.

Certification of compliance would then simply require grantee reporting to include in reports on their funded proposal the data citations and the repository where the data was deposited.

As work is already being carried out to develop standards in this area (i.e. *The ISO 16363 Standard for Trusted Digital Repositories),* it would be more expedient for federal agencies to work within and help support such standards.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

AIP agrees with the statement from the Interagency Working Group on Digital Data (IWGDD) in its report, *Harnessing the Power of Digital Data for Science and Society,* that "the current landscape lacks a comprehensive framework for reliable digital [data] preservation, access, and interoperability". We feel that there is a very important role for the federal government and its science funding agencies to play to help create and promulgate such a comprehensive framework.

Federal investment in creating stable, standardized, and accessible data will be an essential base from which innovation can occur. The ease of reuse could then lead to developments akin to IBM Research's "Many Eyes" product for data visualization (www-958.ibm.com), or spur the private sector to offer data services for researchers.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

This ecosystem of attribution and credit already exists with respect to scholarly articles. A researcher's standing in their field is largely a result of their list of authored scholarly articles and the number of citations to those published articles. The credit comes in the form of respect from peers, funding for further work, and career advancement, and rests in large part on the underlying quality control provided by peer review. Not providing appropriate attribution is considered unethical scientific behavior and can lead to the retraction of published work.

The mechanisms to be developed would support an extension of this system to cover data. The elements to support are:

- data must be recognized as a primary research output,
- data must have unique and persistent identifiers and be fully citable, thereby allowing its use and reuse to be tracked and recorded in the same way as scholarly publications, and
- data citation information must be used for research evaluation and reward.

Persistent identifiers for data could be handled through use of digital object identifiers already used for scholarly articles or similar (see Datacite.org). There are also examples of recommended practice for citing data. [For example: creator (publication year): Title, Publisher, identifier; see http://datacite.org/whycitedata and DOE's Data ID Service.]

Publishers could support the development of such a system by requiring that all data needed to reproduce the results and conclusions of a published scholarly article must be cited according to community standards.

Funding agencies could support the development of such a system by recognizing data that has been archived and made available to the research community as "first class research objects" at the same level as articles. Agencies should also recognize any reuse of these data which could then be counted via citations.

See the Australian National Data Center's "Building a Culture of Data Citation" poster available at http://ands.org.au/cite-data/index.html.

For a hybrid example spanning the world of digital data and scholarly publication, see the *Journal of Physical and Chemical Reference Data*, a long and successful collaboration between AIP and the National Institute of Standards and Technology.

Standards for Interoperability, Reuse and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

First, it is important to separate metadata standards from data format standards. Metadata standards could be developed that are lightweight enough to be widely interoperable and extensible so as to accommodate discipline-specific needs (within the XML publishing standard). These standards would need to cover both bibliographic information (data creator, date of creation, what the data describes, where it can be accessed, etc.), and how it was collected (experimental apparatus, experimental conditions, location, etc.).

Data format standards that would enable reuse and repurposing would need to be developed at the discipline-specific level. There need not be one solution per discipline: it may be that the communities in question need a handful of solutions that correspond to the various types of data and/or modes of scientific research that produces the data. So while it is true that actual data solutions need to be discipline appropriate, there may be logical clusters of solutions for the connections between publishing and data depending on the nature of the data.

There is a role for federal agencies in coordinating across discipline boundaries (covering all funded areas) and internationally. In its October 2011 report, *Federal Engagement in Standards Activities to Address National Priorities: Background and Proposed Policy Recommendations,* the Subcommittee on Standards of the National Science and Technology Council noted that "There was agreement among respondents that the US government should continue to play the role of participant in private sector standards setting processes. There was also general agreement that the effectiveness of government participation depends on the level and consistency of involvement and commitment of resources, both staff and budgetary, to the process. Lack of coordination among agencies…was cited by many respondents as having a negative impact on government effectiveness. "

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Digital Object Identifier, or DOI, is an example of a successful standard. Its development and adoption involved a multi-stakeholder, community-driven approach that solved a practical problem and provided benefit to the end-user.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

AIP supports the recommendation of the Interagency Working Group on Digital Data (IWGDD) that an NSTC Subcommittee for digital data preservation, access, and interoperability be created. This subcommittee would then be able to provide coordination among the US funding agencies and collaborate with its international counterparts. Coordination at the national level should extend beyond

science funding agencies as relevant work is being done elsewhere within the US government (for example, the work of the Library of Congress through its National Digital Information and Infrastructure Program [NDIIP], particularly its "partnership with the National Science Foundation in 2005 to undertake a program of pioneering research to support advanced research into the long-term management of digital information").

In addition, this subcommittee could ensure that each Federal agency is itself required to adopt and implement digital data standards developed within the global community.

Federal agencies can support conferences and other initiatives on a discipline level by funding standards and preservation work as well as pure research.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

See answer to question 9. The mechanism for linking between publications and associated data essentially exists with the digital object identifier, which is already used widely for linking between publications. The federal government could provide additional logistics and financial support for making this mechanism standard practice with respect to data and coordinating/aligning policies across federal agencies to encourage use of those standards by grantees.

Agency involvement and/or support of current initiatives such as the NISO/NFAIS Working Group on Supplementary Journal Information (www.niso.org), which is working on recommended practices for publishers who are increasingly attaching data sets as supplementary information appended to publications, would also help address some of the issues at a practical level.

RSC Publishing

Response to Office of Science and Technology Policy (OSTP) Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

On behalf of the Royal Society of Chemistry, UK

To: Office of Science and Technology Policy (OSTP)
Washington, DC 20502, USA

via e-mail to: digitaldata@ostp.gov

9th January 2012

From:

James Milne PhD
Editorial Director & Acting Managing Director
RSC Publishing
Royal Society of Chemistry
Thomas Graham House
Science Park, Milton Road
Cambridge, CB4 0WF, UK

About RSC Publishing

RSC Publishing is one of the largest and most dynamic publishers of chemical science information in the world. We publish 34 international peer reviewed scholarly journals, approximately 95 scientific books per annum, two highly acclaimed magazines, and a number of successful databases.

Not-for-profit

We are a not-for-profit publisher wholly owned by the Royal Society of Chemistry. Our authors, readers and customers are truly international and our publishing activity dates back to 1841.

Authoritative

RSC Publishing is a member of ALPSP, the Association of Learned and Professional Society Publishers, and we adhere to the ALPSP principles of scholarly-friendly journal publishing practice.

All research articles published by the RSC are peer reviewed. The journals are considered to be of the highest standards in their field, with an average impact factor of an impressive 5.4. Through the professional management of the publishing process, from submission through to publication, RSC content satisfies the pillars of scholarly publishing:

- Certification (validation of quality and integrity)
- Registration (recognition of achievement)
- Accessibility (unparalleled online access, worldwide)
- Archiving (reliable perpetual accessibility)
- Navigation (industry leading services to identify content)

Award-winning

RSC Publishing has been recognised by a number of prestigious awards, including the 2011 ALPSP Best New Journal Award for the high impact journal *Chemical Science*, and several innovation Awards for its free online chemical database ChemSpider.

Professional

The publishing operation is based in Cambridge, UK, and employs around 275 people on the Science Park. These professional publishing staff engage in the preparation, peer review, selection, editing, production, marketing and distribution of information in the chemical sciences. Additional international publishing staff are based in Philadelphia and Raleigh, USA; Beijing and Shanghai, China and Tokyo, Japan.

Investing for the Research Good

As a Not For Profit organization, the RSC sustains its proven and established publishing activities primarily through subscription revenue. This model also enables the RSC to invest in new highly valued services for the community, generally at no additional cost to the user.

By way of example, during 2009 RSC Publishing acquired ChemSpider, a structure centric database for chemists. ChemSpider provides searchable access to over 26 million chemical structures and is considered to be one of the richest single sources of structure-based chemistry information worldwide. RSC Publishing provides free access to this service, as part of its publishing operations. Ref: www.chemspider.com

We welcome the opportunity to respond to the Office of Science and Technology Policy (OSTP) Request for Information (RFI): Public Access to Digital Data Resulting From Federally Funded Scientific Research.

Our comments are presented below, in response to the questions posed in the RFI.

DATA - Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

- We broadly support Government and funding agencies encouraging and endorsing data preservation as part of grant applications and awards
- Consideration should be given to acknowledging the differences between data types or needs, by: subject, level of curation, collection types, etc.
- Additional benefit could be gained from accreditation of a variety of approved repositories
- It is likely to prove important for policies to encourage and incentivise data deposition in approved repositories
- Thoughts should also be given to who may wish to access this data (publicly) and how to make the data discoverable
- Measures should be put in place to ensure ongoing data integrity
- Such activities are likely to support the US and international economies, and encourage international collaboration.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

- Policy and statements should be clearly defined to differentiate what data may be openly (publicly) shared, and what may not.
- Through the creation and adoption of individual repositories, appropriate recognition should be given to the authors, and curators. Suitable referencing (citation) between the data and associated works (e.g. journal articles) will also be both beneficial and necessary.
- The long-term sustainability of the system must be preserved, changes which impact on intellectual property rights must consider the impact on existing policies, processes and systems
- Policies should acknowledge the costs involved in hosting, maintaining, preserving and making available data sets. The hosts should be entitled to recover these costs to ensure sustainability of the systems

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

- Consultation with the appropriate communities will be able to provide this. Considerable research and understanding is already available on differences between scientific disciplines when it comes to data expectations, needs or opportunities.
- Federal agencies may play a lead role in facilitating such discussions, involving scientists, publishers, and information professionals
- Consideration should be given to supporting the utilisation of existing subject specific data repositories, such as ChemSpider (www.chemspider.com) for chemical compounds, or CCDC (www.ccdc.cam.ac.uk) for crystallographic data.
- Where existing repositories are not available, financial support to create suitable data archives could be necessary

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

- If the principle is to provide public access to all publicly funded data, then all research should be treated the same.
- However certain areas have more obvious societal benefits in the short to mid term (such as health related areas). For this reason, the fundamental principle may not be truly appropriate, and emphasis should be given to areas where there are greater opportunity benefits to society.
- Attention could therefore be given to areas where there is genuine public interest in accessing research data.
- It may also be worthwhile consulting with other groups who have already undertaken research on data management, such as:
    - DataCite: http://datacite.org
    - CoData: www.codata.org
    - Opportunities for Data Exchange (ODE): www.ode-project.eu

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

- All stakeholders should collaborate to define a sustainable approach for this complex and broad ranging activity
- All stakeholders would then need to engage and adopt the agreed practices
- Areas such as investments, incentives, liabilities, risks, penalties and sustainability would need to be assessed and fully addressed.
- A detailed communication plan, possibly including training is likely to be necessary to the research community, to ensure their support. To this end, the processes required from researchers to deposit their data should be as simple and administratively simple as possible.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

- Working with established repositories and systems is likely to be the most cost effective solution, where such repositories exist (e.g. ChemSpider, CCDC, as referenced in Q3)
- Investing in new systems that replicate existing systems is likely to be less effective and more expensive
- Adequate funding to support data management systems is necessary to ensure a sustainable solution, especially if data preservation is encouraged or mandated.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

- Defining a minimum/appropriate number of repositories by need and accrediting these internationally would facilitate more effective monitoring of depositions.
- Regular re-accreditation and inspection will ensure the highest standards are maintained

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

- Market forces are likely to lead to innovative use of publicly accessible data
- Repositories openly identifying what data is publicly available, would help facilitate appropriate use or reuse of such data. This would avoid confusion over restricted vs open material which may reside on the same platform.
- We encourage the support of projects, involving researchers, publishers and information professionals, to explore what experimental or innovative uses could be derived from publicly accessible research data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

- Established citation and reference standards should be applied to data records.
- Involvement with DataCite, a not for profit organisation aiming to make data easier to access, would help in this regard.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

- Many subject specific formats exist for data management. Working with the individual communities will ensure appropriate (internationally recognised) standards and formats are adopted.

- Once a community-driven standard has been adopted then journal publishers can work with the community to aid compliance to the standard (i.e. that adopting the standard is an expectation or requirement of publication)

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

- The many chemistry based standards have come around from there being a very strong community desire for a standard to be agreed and adopted. By being community driven, and hence strongly supported by the community, consensus and support is much improved
- Other characteristics for success including fully involving the multiple stakeholders in formulating suitable standards, communicating the process and the outcome to the community (often via societies and/or scientific journals)
- The most important element is to ensure the scientific community, database providers/curators and journal publishers work in partnership to aid the definition of, and compliance to the standard

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

- There are a number of established organisations seeking to define digital data standards in an international context (including groups mentioned in response to Q4).

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

- Through working with publishers and publishing associations, standards and practices can most effectively be defined. Much progress has already been made in this regard, as all parties are supportive of achieving this common goal.

Stephanie Wright
swright@uw.edu
Data Services Coordinator / Atmospheric Sciences Librarian
University of Washington Libraries
Seattle, Washington

January 12, 2012

Office of Science and Technology Policy
The White House

RE: Comments in response to Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting From Federally Funded Research
Federal Register Doc No 2011-28621
http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28621.htm

Thank you for the opportunity to comment on this issue of such great importance to the future of scientific research in this country. I have provided my responses to those questions below to which I felt I had the most relevant expertise.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

At the most basic level there needs to be a federal policy that mandates federally funded research data is deposited in a manner that makes the data openly accessible and fully reusable, with exceptions to this rule being applicable only in instances where access to the data poses privacy or security concerns. Exceptions should be based on ethical, not proprietary, considerations. "An open data regime not only maximizes the benefit of the data, it also simplifies most of the other issues around effective research data stewardship and infrastructure development."[1] There is already research that shows open access to research data increases citations and reuse of data.[2]

While immediate access by the public to this data would be ideal, an embargo period would not be out of the question to allow the original researcher/s to capitalize on the publication opportunities. That being said the allowable embargo period should not be so long as to unnecessarily restrict access for an extended period of time and slow down advancements that can be made through the reuse of that data. This will maximize the return on federal investment in the original research.

The federal policy should also include a clearly defined provision for free and open re-use of the data either by mandating that the data be in the public domain or at most maintaining that

[1] Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. GRDI2020. http://www.grdi2020.eu/Repository/FileScaricati/e31a1aab-b01e-4e7e-9b10-0fd93d4b710f.pdf Accessed 12 January 2012.
[2] Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

subsequent users of the data must provide attribution along the lines of the requirements for the Creative Commons CC:BY license.[3]

Free and open access to research data provides new opportunities for commercial development of not only one's own intellectual property but also that of others. It opens up opportunities for everyone and accelerates scientific and commercial innovation.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Permitting free and open use of research data through a license similar to the CC:BY license, will allow credit to go to the researchers who invest significant effort into the research and collection of the data while also allowing for any subsequent researchers to be clear on the rights surrounding its reuse. This will also minimize the reluctance to reuse data by alleviating fears of lawsuits as current copyright law is poorly understood by most researchers. Even the current copyright law stating that facts are not copyrightable leads to misunderstanding surrounding its availability for reuse. Clear licensing of the data in this manner will minimize the complexities in instances where data is being used by researchers in international collaborations from different countries with significantly different or conflicting copyright law.

There would need to be development on a standard for data citation, preferably one that allows for tracking of citations between publications and their related datasets. This standard for citation would also need to take into account reusability, merging of datasets and versioning so credit can be given and shared appropriately.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report proposed that legislation be enacted that would permit the Library of Congress "to designate specially qualified institutions as agents for mandatory deposit for specific types of content."[4] By taking this further and making the qualified repositories domain specific, it can simplify identification of the appropriate repository for a particular dataset and improve the findability of relevant datasets, much as subject classification does for print materials. It would also minimize the number of different types of data any one repository would need to accommodate and maintain and can build on existing infrastructure. This could lead to enhanced collaboration between stakeholder communities and increased likelihood of standard creation relevant to the domain.

---

[3] http://creativecommons.org/licenses/by/3.0/

[4] National Digital Information Infrastructure and Preservation Program. (2010). Preserving Our Digital Heritage: The National Digital Information Infrastructure and Preservation Program 2010 Report. Retrieved from website: http://www.digitalpreservation.gov/multimedia/documents/NDIIPP2010Report_Post.pdf. Accessed 12 January 2012

When it comes to creation of a federal policy, allow funding agencies to develop relevant guidelines for researchers in different domains such as NSF's data management plan guidelines which vary by directorate.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Grant funding to researchers and their institutions will need to be increased to help allay the costs of maintaining the qualified repositories mentioned in Question 3 with different amounts of funding based on the data type requirements of those domain specific repositories. This will help maintain the sustainability of these repositories by sharing the burden between multiple institutions, the organizations that support the domain as well as the federal government, all of whom would benefit from access to the data in the repository.

For domains where there are not already established repositories, the funding agencies can provide startup money to libraries and their institutions to develop a relevant archive with the stakeholders in that domain (research institutions, publishers, other domain-related organizations, related industries) developing a plan to cover long-term costs of the repository.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Shared responsibility among all the stakeholders is the most effective way to achieve the goal of maximizing the investment into research and its outputs. All the stakeholders can work together to develop, promote and support a domain-based infrastructure. They can collaborate on developing standards and tools for metadata capture and shifting the current data-silo culture to a one of data sharing.

Publishers and research organizations can enable and encourage ethical data sharing by providing recognition to researchers through data citation and cross linking. Universities can increase the incentive to share and reuse data by tying both of those to the tenure and promotion process. They can also make sure that data management is applicable earlier in the career of the researcher by making data management plans required for dissertations and theses with related research data. Libraries and universities can educate researchers on the importance and value of proper data management. The government can encourage this by providing funding for creation of data management curriculum.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Along with those suggestions provided in Question 2, there could be development of author and institutional identifiers as is being attempted through the ORCID (Open Researcher & Contributor ID) organization.[5] Buy developing a method of citation that incorporates data

[5] http://orcid.org/

creator identifiers, recognition could more easily be given to those who not only did the original work but subsequent researchers who add value to the original data as well.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

In conjunction with suggestions made in Question 2 and Question 9, it would be in the best interest of the researchers, their institutions and publishers to work together to create a standard method of tracking and linking citations of datasets with the related works published in their publications. They could work with an existing organization such as DataCite to extend the functionality of their EZID permanent identifier service to meet that need.[6] By allowing cross linking of datasets and their related publications, not only is the dataset given increased visibility but so is the publication derived from the data.

[6] http://datacite.org/cdl_launch_ezid

Mon 1/2/2012 7:20 PM

Public Access to Federally Funded Research

I am an ordinary US citizen, 67 years old, who is most DEFINITELY in favor of public access to ALL federally funded research in ALL areas of endeavor. I also think it should cost very little to make such access available digitally at this point in time since most people now have computers and access to the internet. I think had we had access to federally funded medical research, in particular, in past years, there would not have been so many reports after the fact of drugs and medical procedures that have harmed people who used them.
Access to information, PUBLIC access made available to anyone desiring it with enough time and interest to read it, is one of the hallmarks of this republic, and most certainly should be allowed as much as possible when the means is available to make it accessible to as many as possible.

I doubt these comments by an ordinary citizen will make one fig's worth of difference to whoever is collecting these "comments," but I'm sending them anyway.

Gail Kearns

Mon 1/2/2012 8:07 PM
response to RFI on data sharing

Recommendations on ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research

George B. Moody
Harvard-MIT Division of Health Sciences and Technology Cambridge, MA

I am one of the founding members of PhysioNet (http://physionet.org), an NIH-funded resource that curates and provides free web access to many large collections of recorded physiologic signals and time series, and to related open-source software [1,2]. PhysioNet, established in 1999, is intended to stimulate current research and new investigations in the study of complex biomedical and physiologic signals. About 45,000 visitors use PhysioNet each month, accessing data and software contributed by federally-funded researchers and others worldwide. As a measure of effectiveness, a Google Scholar search returns (as of January 2, 2012) over 7000 articles and patents citing PhysioNet or making use of data or software provided by PhysioNet [3].

The observations and recommendations below are mine alone, although they are informed by personal experience with PhysioNet and with related data-sharing efforts beginning in about 1978. They may not reflect the opinions of my employer.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

When applying for funding, researchers understand that their opportunities to work on the problems they may have invested years of their lives to pursue depend critically not only on the quality and relevance of the ideas they propose to study, but also on their scientific productivity and the impact of their work, performance metrics based on their publications.

Many researchers are understandably reluctant to share data they may have invested considerable effort in collecting, when they have little or no reason to expect recognition (i.e., improved likelihood of obtaining research funding) for having done so. This reluctance may be compounded by apprehension that a competitor for funding may learn something from their shared data that will improve his or her own chances.

It may not be possible to provide a set of motivations for data sharing that can overcome these perceived incentives for data hoarding in all cases.
Funding agencies can take a major step in this direction, however, by directing peer review panels charged with ranking research proposals to consider the applicants' data-sharing record. Agencies wishing to encourage data sharing might require that assessments of applicants' scientific productivity

and impact of their work should reflect the amount, quality, timeliness, and relevance to current research of the data they have shared.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

This question assumes that researchers may have IP interests that need to be protected from damage caused by disclosure of data collected using public funds.
It is not obvious that this assumption is warranted.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

In my field, and in others in which data may include identifiers that permit them to be associated with individual human subjects, privacy concerns are often perceived to trump the public interest in accessing research data. At a minimum, researchers must anonymize (deidentify) data before making them publicly accessible. This process can become expensive, and it is often difficult to determine if all identifiers have been removed.

As a result, data sets that might help to address major public health challenges are often not shared, and agencies fund redundant data-collection efforts that fall short of what might be possible if larger numbers of subjects were studied (by combination of data sets) or if the population entered into a single study were examined in greater depth by other investigators.

Agencies might help by establishing repositories for controlled access to data that have been scrubbed to a high standard that nevertheless may fall short of complete deidentification, indemnifying researchers who contribute data to such repositories, and requiring those who access the contributed data to refrain from any use other than scientific research.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

[no recommendation]

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Institutions such as university libraries may be particularly appropriate curators of shared research data. They are vital to the educational missions of the institutions to which they belong and from which they receive stable long-term support. Academic researchers can collaborate with libraries to develop new and more effective tools for dissemination of their research data to students and educators as well as

researchers. Universities may perceive opportunities to strengthen their research activities by promoting data sharing to build collaborations among their faculty members.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

First, it is appropriate that funding applicants should propose a plan to preserve the data they will collect and to make them accessible, and that this plan must include a budget for doing so.

Second, funding agencies might wish to consider awarding small supplementary grants for data-sharing to grantees whose data are nominated by others as likely to be valuable if shared, on the basis of the progress reports. This might go a long way towards addressing the objection that the costs of data sharing cannot be carved out of a research budget painlessly. For a typical NIH R01, perhaps an award of $15-25K would be sufficient. Such awards would be especially useful in the case of existing grants for which data-sharing is mandated or desired but not explicitly budgeted for.

Third, it is appropriate that applicants' past performance with respect to data sharing be given significant weight during evaluation of new proposals, just as their publications are considered as evidence of the impact of their work and of their productivity. If the expectation is that research data derived using public funds belong to the public, then those who have shared their data have met expectations. Those whose data have been used by other researchers, as evidenced by secondary publications and letters of support, may deserve an extra measure of credit. Those who have not shared their data, especially after receiving a grant under terms that require data sharing, should not receive additional funding until they have met the terms of any previous grants.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Recipients of funding are already required to produce periodic progress reports detailing the results of their funded research. It should require very little if any additional effort for grantees to document progress on their data sharing plans.

In my field, it is unfortunately common for researchers to put off efforts at data sharing until the conclusion of the major study, since it is also common to withhold data from other researchers until that time. The all-too-frequent result is that funding runs out, the graduate student who collected the data (and who is the only one who knows how they are organized) has moved on, and the principal investigator is writing the next grant application. Data sharing becomes an afterthought, and if it happens at all it will be late, incomplete, unnecessarily expensive, and sub-optimally organized for re-use. These experiences may not be typical of projects in other fields, but I suspect it is the norm for many of them in which there is not already a well-established culture that expects data sharing.

Agencies can help to steer their grantees away from this counterproductive pattern by encouraging or requiring that data be deposited at the times that progress reports are due, in an archive accessible to the agency. Although I would not expect a funding agency to review all such deposits, they should be considered as appendices to the progress reports, and grantees should expect that they will be examined if questions arise with respect to research progress. Researchers should also be encouraged

to share these data appendices with external advisors who may be able provide early feedback on (re)usability.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

PhysioNet sponsors annual open challenges aimed at accelerating research on significant problems that can be addressed using freely available data [4].
These events offer an opportunity for anyone interested to work on a worthwhile question and (perhaps) to make progress towards its solution without requiring the lengthy, difficult, and expensive data-collection effort that would otherwise be needed in order to begin serious work.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

It is already in a researcher's best interest to cite sources of data, just as it is expected that researchers will cite publications that are relevant to new work. Doing so establishes a basis for understanding the innovative aspects of the new work; at the very least it allows one to demonstrate awareness of what has been done by others. Citing a well-known and generally available data source can often answer the otherwise difficult questions that peer reviewers may raise with respect to methods of data collection, selection of subjects, and experimental conditions. In scholarly writing, there is no incentive to misattribute or plagiarize data, because proper attribution only adds credibility to the work.

Nevertheless, in some circumstances, explicit disincentives may be needed to deter data plagiarism or misattribution.

Funding agencies wishing to provide such disincentives might begin by adopting a policy that applicants for funding must cite data sources in all of their publications (or presentations, or grant applications) making use of them, and that failure to do so constitutes scientific misconduct that will impact the offender's eligibility for funding to the same extent as any other fraud.

Researchers wishing to avoid having their shared data be plagiarized can make them freely available, and specifically accessible to search engines such as Google. It then becomes trivial to check for suspected plagiarism, and the high likelihood of exposure acts as a further disincentive to misbehavior, if indeed one is needed.

Finally, to the extent that data sharing becomes the norm in a field of research, papers that report on data that have been neither shared by the authors nor attributed to another source should not pass peer review, and will be judged unreliable and likely fraudulent. Journals may be able to accelerate this evolution of the norm by adopting standards that require (or encourage) data sharing.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature

Genetics 29, 371) is an example of a community-driven data standards effort.

MIAME and related efforts (see MIBBI [5]) address the issues of how to describe a data set (e.g., in English) with sufficient detail to permit reuse. In general, these "minimum information" projects are aimed at specifying the content of the description, rather than its format or the format of the data themselves.

There is much less to say about data formats, but there are nevertheless important points about formats to be considered in the context of data standards:

Most important is the use of open formats. Readability of data must not be dependent on availability of a specific computing platform (operating system and CPU) and/or a specific reader application. Ideally, creating a reader for a new computing platform should be a simple process.

When research data must be deidentified before public access can be allowed, it is especially important to avoid the use of proprietary formats that may conceal data identifiers thought to have been deleted.

In my field, data files often contain lengthy time series of observations. Frequently, short intervals ("events") within a long series are of particular interest, hence there are advantages to formats that permit efficient random access. do not require that files be read from its beginning in order to locate a desired time interval.

Finally, one of the lessons drawn from PhysioNet is that the use of common data formats is important. We encourage all contributors to use open formats for their data, in part to ease readability, but also so that researchers who may use contributed data will not need to learn to use a new set of tools for each new data set. This is an important component of the value of shared data.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

In my field, EDF (the European Data Format for polysomnograms and EEGs [6]) is an excellent example of a highly successful standard for a data storage format. It was designed by a handful of medical engineers who met at a conference in 1987 and published the specification in 1992. EDF is currently used by at least 90 companies, including all of the major manufacturers of polysomnographs and EEG recorders worldwide. The (free) specification fits on a single well-written page, and an EDF reader or writer can be implemented from scratch in a day or less by a competent programmer. The format is reasonably storage-efficient (important since EDF recordings can be quite lengthy).

Another example is SCP-ECG (Standard Communications Protocol for computer assisted Electrocardiography), which was designed between 1989 and 1991 and then redesigned in a formal standard development process between 1995 and 2001 in the US, continuing until 2005 in Europe. The SCP-ECG standard is described in a document of about 200 pages [7]. Embedded metadata selects combinations of many alternative formats included at the behest of representatives of manufacturers who participated in the standard development. Most of the variant formats are highly storage-efficient.

SCP-ECG has not been widely adopted, even by the manufacturers represented on the committee responsible for the decade-long standards process.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

It is vitally important that digital data standards be open. Standards that are encumbered by proprietary technology divide the worldwide community into those who have the right to use them, and those who don't. The only reason to have a data standard at all is to promote communication. Hence the most valuable advocacy role for Federal agencies in data standards development is to resist the intrusion of proprietary technology into data standards.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

It is not clear that any action is needed in this regard, but it may help to accelerate what is already a growing trend of citing data sources by mandating such citations when shared data have been used.

References

[1] A.L. Goldberger, L.A.N Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[2] G.B. Moody, R.G. Mark, A.L. Goldberger. "PhysioNet: Physiologic Signals, Time Series, and Related Open Source Software for Basic, Clinical, and Applied Research," Proc. 33rd IEEE EMBS 8327-8330 (2011). [PDF attached]

[3] http://physionet.org/publications/

[4] http://physionet.org/challenge/

[5] http://mibbi.org/index.php/MIBBI_portal

[6] http://www.edfplus.info/

[7] ANSI/AAMI EC71:2001

33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

PhysioNet: Physiologic Signals, Time Series and Related Open Source Software for Basic, Clinical, and Applied Research

George B. Moody, Roger G. Mark, and Ary L. Goldberger

Abstract— **PhysioNet provides free web access to over 50 collections of recorded physiologic signals and time series, and related open-source software, in support of basic, clinical, and applied research in medicine, physiology, public health, biomedical engineering and computing, and medical instrument design and evaluation. Its three components (PhysioBank, the archive of signals; PhysioToolkit, the software library; and PhysioNetWorks, the virtual laboratory for collaborative development of future PhysioBank data collections and PhysioToolkit software components) connect researchers and students who need physiologic signals and relevant software with researchers who have data and software to share. PhysioNet's annual open engineering challenges stimulate rapid progress on unsolved or poorly solved questions of basic or clinical interest, by focusing attention on achievable solutions that can be evaluated and compared objectively using freely available reference data.**

I. INTRODUCTION

PhysioNet (`http://physionet.org/`) is a research resource intended to stimulate current research and new investigations in the study of complex biomedical and physiologic signals. It has three major components:

PhysioBank is a large and growing archive of well-characterized digital recordings of physiologic signals, time series, and related data for use by the biomedical research community. PhysioBank currently includes more than 50 collections of cardiopulmonary, neural, and other biomedical signals from healthy subjects and patients with a variety of conditions with major public health implications, including sudden cardiac death, congestive heart failure, epilepsy, gait disorders, sleep apnea, and aging (see Tables I and II). These collections include data from a wide range of studies, as developed and contributed by members of the research community.

PhysioToolkit is a large and growing library of software for physiologic signal processing and analysis, detection of physiologically significant events using both classical techniques and novel methods based on statistical physics and nonlinear dynamics, interactive display and characterization of signals, creation of new databases, simulation of physiologic and other signals, quantitative evaluation and comparison of analysis methods, and analysis of nonequilibrium and nonstationary processes. A unifying theme of many of the research projects that contribute software to PhysioToolkit is the extraction of "hidden" information from biomedical signals, information that may have diagnostic or prognostic value in medicine, or explanatory or predictive power in basic research. All PhysioToolkit software is available in source form under the GNU General Public License (GPL).

PhysioNetWorks is a virtual laboratory for development of data and software resources that will eventually become components of PhysioBank and PhysioToolkit. By providing large, secure workspaces with redundant backup to active researchers who can easily share them with colleagues anywhere, PhysioNetWorks encourages investigators to create well-organized and documented, *usable* data and software repositories during the conduct of their research. When the research is complete and the major results have been published (or at any time the researcher wishes) the repository can be shared with a colleague, a group of colleagues, or the research community at large.

II. BACKGROUND

PhysioNet was established in 1999 as the outreach component of the Research Resource for Complex Physiologic Signals[1], [2], a cooperative project initiated by the authors at Boston's Beth Israel Deaconess Medical Center, Harvard Medical School, and MIT, together with colleagues at Boston University, McGill University, and later at numerous other institutions. Beginning in the mid-1970s, members of the PhysioNet team who were then working on some of the first microcomputer-based instruments for cardiac arrhythmia monitoring foresaw the usefulness of establishing shared databases of well-characterized ECG recordings, as a basis for evaluation, iterative improvement, and objective comparison of algorithms for automated arrhythmia analysis. A five-year effort culminated in the publication of the MIT-BIH Arrhythmia Database in 1980, which soon became the standard reference collection of its type, used by over 500 academic, hospital, and industry researchers and developers worldwide during the 1980s and 1990s. Other databases of ECGs and eventually other physiologic signals followed. By 1999, the MIT group distributed CD-ROMs containing 11 such collections, and had participated in the development of several others.

The MIT group contributed its 11 databases, and the software it had developed for exploring and analyzing them,

PhysioNet is supported by the US National Institutes of Health (NIH), its National Institute of Biomedical Imaging and Bioengineering (NIBIB), and its National Institute of General Medical Sciences (NIGMS), under NIH cooperative agreement U01-EB-008577.

GBM and RGM are co-PIs of PhysioNet, and ALG is its program director. GBM is the architect and technical director of PhysioNet, and is with the Laboratory for Computational Physiology (LCP) in the Harvard-MIT Division of Health Sciences and Technology (HST). RGM directs the LCP, and is a Distinguished Professor in HST and EECS at MIT. ALG directs the Rey Institute for Nonlinear Dynamics in Medicine at Boston's Beth Israel Deaconess Medical Center, and is a Professor of Medicine at Harvard Medical School.

Address for correspondence: George B. Moody, Massachusetts Institute of Technology, Room E25-505A, 77 Massachusetts Ave., Cambridge, MA 20139 USA. Email: `george@mit.edu`

to establish PhysioBank and PhysioToolkit. Free availability of these resources via the Internet catalyzed an even greater explosion of interest in them, as researchers and students worldwide who had no previous access to such data or software began new programs of research, and specialists began comparing their methods. These initial contributions were quickly supplemented by additional collections of data and software from their collaborators, and soon after, from many researchers worldwide. PhysioBank and PhysioToolkit have grown to many times their original sizes, and most of the growth has been thanks to the hard work and generosity of an international community of researchers.

III. ACTIVITIES

Shortly after PhysioNet was established, we initiated an annual series of open engineering challenges, in cooperation with the annual IEEE-EMBS-sponsored conference, *Computers in Cardiology* (now *Computing in Cardiology*, or CinC). We hoped to introduce PhysioNet to our international colleagues who would be attending CinC, by encouraging participation in an activity that made effective use of the facilities provided by PhysioNet to stimulate rapid progress on an unsolved problem of practical clinical significance. A timely contribution of data from Thomas Penzel made it possible to create the first PhysioNet/CinC Challenge, which attracted the attention of more than a dozen teams to the subject of detecting sleep apnea from the ECG[3]. Their efforts were broadly successful, they discussed their findings at CinC 2000, and an annual tradition was born.

In complementary ways, PhysioNet and CinC catalyze and support scientific communication and collaboration between basic and clinical scientists. The annual meetings of CinC are gatherings of researchers from many nations and disciplines, bridging the geographic and specialty chasms that separate understanding from practice, while PhysioNet provides on-line data and software resources that support collaborations of basic and clinical researchers throughout the year. The annual PhysioNet/CinC Challenges seek to provide stimulating yet friendly competitions, while at the same time offering both specialists and non-specialists alike opportunities to make progress on significant open problems whose solutions may be of profound clinical value.

The use of shared data provided via PhysioNet makes it possible for participants to work independently toward a common objective. At CinC, participants can make meaningful results-based comparisons of their methods; lively and well-informed discussions are the norm at scientific sessions dedicated to these challenges. Discovery of the complementary strengths of diverse approaches to a problem when coupled with deep understanding of that problem frequently sparks new collaborations and opportunities for further study, as occurred when participants in the first Challenge combined their efforts to obtain an even better solution to the Challenge problem[4].

Recent challenge topics have included predicting acute hypotensive episodes in intensive care unit patients[5]; developing robust methods for filling gaps in multiparameter physiologic data (including ECG signals, continuous blood pressure waveforms, and respiration), with applications in detection of clinically important events and in reduction of false alarms in the ICU[6]; and (in progress) improving the quality of ECGs collected using mobile phones.

In a prototype implementation, PhysioNetWorks supported the 2010 Challenge, collecting and scoring entries from the participants. During its first five months (to mid-June 2011) the reimplented PhysioNetWorks has attracted more than 400 members, who are using it to support 12 collaborative projects and 15 others in preparation, in addition to the challenge in progress. Current PhysioNetWorks projects include data collection and annotation development efforts, as well as efforts aimed at improving or evaluating physiologic models and other software projects focused on creation or improvement of tools for research. New contributions of data and software are channeled through PhysioNetWorks, allowing their creators to participate in all aspects of curation of their contributions, and allowing the community to provide early feedback to influence decisions that may affect usability and value to researchers.

Significantly, many PhysioNetWorks projects are being established early in the grant cycles of the associated research projects. Use of PhysioNetWorks throughout the active phase of research encourages investigators to organize their work in a way that makes it easy to use community-developed exploratory and analytic tools by the research team and collaborators in the short run, thus making their final contribution more valuable to the research community at large, and avoiding the pitfalls of attempting to fulfill a mandated data-sharing requirement after funding for the project has ended and those who understand the data have begun work on other projects.

IV. CONCLUSIONS

In its first 12 years, PhysioNet has made a wide variety and large quantity of well-characterized data and related open-source software collected and created for biomedical research, often at great expense, available for re-use and further study at no cost by a worldwide community of over 40,000 researchers, clinicians, educators and students, and medical instrument and software developers. Through its open engineering challenges, it has stimulated development of inexpensive and minimally disruptive technology for detection of sleep apnea and sleep quality, prediction of adverse events and reduction in false alarms in the intensive care setting, and telemedicine. A Google Scholar search for PhysioNet and related terms finds over 5000 publications and citations as of June 2011. Finally, PhysioNetWorks has allowed us to scale up our capacity to bring new data collections and software packages on-line without requiring a proportional increase in the size of our team.

TABLE I

PHYSIOBANK COLLECTIONS OF MULTIPARAMETER AND ECG SIGNALS AND TIME SERIES

(AS OF JUNE 2011)

Collection	*Subjects*	*Duration (typical)*	*Signals and time series*	*Other*
MGH/MF Waveform Database	250	90-120 min	ECG (3 leads), ABP, PAP, CVP, respiration, airway CO_2, ...	beat annotations
Stress Recognition in Automobile Drivers	17	60-90 min	ECG, EMG, GSR, respiration	
Apnea-ECG Database	70	8 hours	ECG (subset includes respiration)	apnea annotations
Fantasia Database	40	2 hours	ECG (subset includes uncalibrated NIBP)	beat annotations
MIMIC Database	121	20-40 hours	ECG, BP, respiration, SpO_2, ...	beat labels, ICU monitor alarms
MIMIC II Waveform Database	20935	3-10 days	ECG, BP, respiration, SpO_2, ...	
MIMIC II Clinical Database	32536	3-10 days	hourly vital signs, medications, ...	ICD9 codes, lab tests, discharge summaries, ...
MIT-BIH Polysomnographic Database	16	8 hours	ECG, ABP, EEG, respiration, ...	apnea and sleep stage annotations
Sleep-EDF Database	8	8 hours	EEG (2), EOG, ...	hypnograms
SVUH/UCD Sleep Apnea Database	25	8 hours	ECG (3 leads), EEG (2), EOG (2), EMG, oronasal airflow, ribcage and abdomen movements, SpO_2, snoring, body position	apnea and sleep stage annotations
ANSI/AAMI EC13 Test Waveforms	10	1 minute	ECG	
European ST-T Database	90	2 hours	ECG (2 leads)	beat, rhythm, ST and T change annotations
Long-Term ST Database	86	24 hours	ECG (2 or 3 leads)	beat, rhythm, ST and signal quality annotations
MIT-BIH Arrhythmia Database	48	30 min	ECG (2 leads)	beat, rhythm, and signal quality annotations
MIT-BIH Noise Stress Test Database	15	30 min	ECG (2 leads) with calibrated noise	beat annotations
BIDMC Congestive Heart Failure Database	15	20 hours	ECG (2 leads)	beat annotations
Post-Ictal Heart Rate Oscillations in Partial Epilepsy	7	90-220 min	ECG	beat and seizure annotations
QT Database	100	15 min	ECG (2 leads)	annotations of onsets, peaks, ends of P, QRS, and T waves
AF Termination Challenge Database	30	1 min	ECG (2 leads)	beat annotations
Creighton University Ventricular Tachyarrhythmia Database	35	8 min	ECG (2 leads)	beat and VF annotations
Intracardiac Atrial Fibrillation Database	8	3-5 min	ECG (3 surface and 5 intracardiac leads)	beat annotations
Long-Term AF Database	84	24 hours	ECG (2 leads)	beat annotations
MIT-BIH Atrial Fibrillation Database	25	10 hours	ECG (2 leads)	beat and rhythm annotations
MIT-BIH ECG Compression Test Database	168	20 sec	ECG (2 leads)	
MIT-BIH Long-Term Database	6	24 hours	ECG (2 leads)	beat annotations
MIT-BIH Malignant Ventricular Arrhythmia Database	22	30 min	ECG (2 leads)	rhythm and signal quality annotations
MIT-BIH Normal Sinus Rhythm Database	18	24 hours	ECG (2 leads)	beat annotations
MIT-BIH ST Change Database	28	20-40 min	ECG (2 leads)	beat annotations
MIT-BIH Supraventricular Arrhythmia Database	78	30 min	ECG (2 leads)	beat annotations
Non-Invasive Fetal Electrocardiogram Database	1	5-20 min	ECG (maternal and fetal; 55 recordings over a 20 week period)	maternal beat annotations
PAF Prediction Challenge Database	100	30 min x2	ECG (2 leads)	beat annotations
PTB Diagnostic ECG Database	549	2 min	ECG (15 leads)	clinical summaries
St Petersburg INCART 12-lead Arrhythmia Database	75	30 min	ECG (12 leads)	beat annotations
Sudden Cardiac Death Holter Database	23	8-24 hours	ECG (2 leads)	beat annotations
T-Wave Alternans Challenge Database	100	2 min	ECG (12 leads, some 2 or 3)	beat annotations

TABLE II

PHYSIOBANK COLLECTIONS OF RR INTERVALS, GAIT, BALANCE, NEURO- AND MYOELECTRIC SIGNALS AND TIME SERIES (AS OF JUNE 2011)

Collection	*Subjects*	*Duration (typical)*	*Signals and time series*	*Other*
CAST RR Interval Sub-Study Database	809	24 hours	-	RR intervals
Congestive Heart Failure RR Interval Database	29	24 hours	-	RR intervals
Exaggerated heart rate oscillations during two meditation techniques	46	10 min - 6 hours	-	RR intervals
Normal Sinus Rhythm RR Interval Database	54	24 hours	-	RR intervals
Spontaneous Ventricular Tachyarrhythmia Database (Version 1.0 from Medtronic, Inc.)	135	5-10 min	-	RR intervals
Gait Dynamics in Neuro-Degenerative Disease	64	2 min	foot pressure	stride intervals
Gait in Aging and Disease Database	15	6-15 min	-	stride intervals
Gait Maturation Database	50	10 min	-	stride intervals
Gait in Parkinson's Disease	93	5 min	multiple foot force signals	stride intervals
Noise Enhancement of Sensorimotor Function	27	5-10 minutes	postural sway	
Unconstrained and Metronomic Walking Database	10	1 hour	-	stride intervals
CHB-MIT Scalp EEG Database	23	1-4 days	23-26 EEG signals	seizure annotations
EEG Motor Movement/Imagery Dataset	109	1-2 min	64 EEG signals	task annotations
Effect of Deep Brain Stimulation on Parkinsonian Tremor	16	1 min	rest tremor velocity	
Evoked Auditory Responses in Normals across Stimulus Level	8	5 min	evoked auditory response, oto-acoustic emissions	stimulus annotations
Term-Preterm EHG Database	300	30 min	EHG	
Examples of Electromyograms	3	10-30 sec	EMG	

REFERENCES

[1] A.L. Goldberger, L.A.N Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[2] G.B. Moody, R.G. Mark, and A.L. Goldberger. "PhysioNet: a Web-based resource for the study of physiologic signals, *IEEE Eng in Med and Biol* 20(3):70-75 (May-June 2001).

[3] G.B. Moody, R.G. Mark, A.L. Goldberger, and T. Penzel. Stimulating rapid research advances via focused competition: the Computers in Cardiology Challenge 2000, *Computers in Cardiology* 27:207-210 (2000).

[4] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray and G. Moody. Systematic comparison of different algorithms for apnoea detection based on electrocardiographic recordings, *Medical & Biological Engineering & Computing* 40:402-407 (2002).

[5] G.B. Moody, L.H. Lehman. Predicting Acute Hypotensive Episodes: The 10th Annual PhysioNet/Computers in Cardiology Challenge, *Computers in Cardiology* 36:541-544 (2009). (PMID:20842209)

[6] G.B. Moody. The PhysioNet/Computing in Cardiology Challenge 2010: Mind the Gap, *Computing in Cardiology*, 37:305308 (2010).

Tue 1/3/2012 12:50 PM

The DuraSpace organization is a member of the NDSA(National Digital Stewardship Alliance) and is in full support of the joint statement crafted by the member organization and in collaboration with the Library of Congress which can be found in the attached document, fully supporting the stewardship and open access of data from federally funded projects.

best,
Michele Kimpton
Chief Executive Officer
DuraSpace organization

NDSA
NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

Submitted by the National Digital Stewardship Alliance (NDSA)

January 2, 2012

Introduction to the NDSA

The National Digital Stewardship Alliance (NDSA) was founded in July 2010 to extend work begun in 2001 by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress. The Alliance has over 100 members from educational institutions, non-profit organizations, businesses and local, state and federal government agencies, as well affiliations with international organizations. Its mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. [1] Members of the Alliance are taking action to preserve access to our national digital heritage by:

- broadening access to our nation's expanding digital resources
- developing and coordinating sustainable infrastructures for the preservation of digital content
- advocating standards for the stewardship of digital objects
- building a community of practice around the management of distributed digital collections
- promoting innovation
- facilitating cooperation between government agencies, educational institutions, non-profit organizations, and commercial entities
- fostering the participation of diverse communities and relationships across boundaries
- raising public awareness of the enduring value of digital resources and the need for active stewardship of these national resources.

Supporting communities of practice for preservation and access

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally-funded scientific research. When applied, these values support the practical development of communities of practice capable of gaining consensus to support preservation and access to digital data. The shared expertise and common experience of these communities result in stakeholder buy-in and adoption of policies and

standards. The National Digital Stewardship Alliance member organizations are bound as a community by the following values.

Stewardship. Members of the NDSA are committed to managing digital content for current and long-term use. The members of the NDSA are actively ensuring sustained access to the digital content that constitutes our national legacy and empowers us as leaders in the global knowledge economy. Individually, these organizations support the management of digital resources; the Alliance is committed to protecting our nation's cultural, scientific, scholarly, and business heritage.

Collaboration. Collaborative work is the centering value of the Alliance; it is a value shared by all members and a priority in work with all organizations and associations. Approaching digital stewardship collaboratively allows the NDSA to coordinate effort, avoid duplicate work, build a community of practice, develop new preservation strategies, flexibly respond to a changing economic landscape, and build relationships to increase capacity to manage content beyond institutional boundaries.

Inclusiveness. The NDSA is a collaborative effort to preserve a distributed national digital collection for the benefit of current and future generations. We value the range of experience, the potential for innovation, and the fault-tolerance that heterogeneity brings. We believe the preservation of digital information is a pervasive challenge and that engaging across different communities strengthens the nation’s digital preservation practices and increases the likelihood of preserving content now and into the future.

Exchange. Members of the Alliance encourage the open exchange of ideas, services, and software. This leverages the commitments of each member to increase the capacity of the entire stewardship network. Participation and engagement result in innovations and benefits that can be shared by all. The Alliance is committed to transparency and all products generated or produced by the Alliance will be circulated under open licenses.

Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

Community-based approaches to the challenges of rapid change and high volume within the data domain have proven to be the most successful in the long term. The Blue Ribbon Task Force on Preservation and Access recommended that for research data “Each domain, through professional societies or other consensus making bodies, should set priorities for data selection, level of curation, and length of retention.” [2]

The report validated experience over the last ten years of digital preservation work. A study of the networks developed through the NDIIPP program indicated that participating institutions bring to the network their own resources, interests, and organizational culture. Under the auspices of a neutral convener and honest broker, natural networks emerge over time through participation in shared activities and problem solving. As these networks form, the larger network becomes more complex, but also stronger and better able to withstand stresses and strains. [3]

The Opportunities for Data Exchange (ODE) project supported by the Alliance for Permanent Access and the European Union also takes a cross-cutting community approach to preservation and access to digital data. “The potential answers to grand challenges of our times require…the inclusion of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-infrastructures…All stakeholders in the scientific process must be involved in the design of this layer; policy makers, funders, infrastructure operators, data centers, data providers and users, libraries and publishers…” [4]

An exemplar of collaborative community efforts is the Dataverse Network project [5] recently described by the National Research Council of the National Academies as the “State of the Practice in Data Sharing.” [6] The Dataverse Network is “unique in being designed to explicitly support long-term access and permanent preservation. To this end the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats.” [7]

Foster public values and support for stewardship of digital data beyond mandating data management plans.

Policy should assert the value of research data and provide mechanisms to support the preservation, discoverability and access. To relieve frustration and confusion about actions the policy should provide a clear direction for funders, researchers and stewardship organizations. The Blue Ribbon Task Force recommended “Funders should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving. They should explicitly recognize “data under stewardship” as a core indicator of scientific effort and include this information in standard reporting mechanisms.” [8]

Leverage substantial national and international efforts for common practices that support interoperability.

Substantial efforts have been made to pave the way for interoperability, re-use and re-purposing. Emerging practices for data citation, licensing and protocols for data sharing and sustainable re-use are becoming enough to adopt more broadly. Notable in these areas are work on the Data Seal of Approval by the Data Archiving and Networked Services that promotes sustainable access to digital research and provides training and advice about archiving and reuse.[9] LOCKSS is a community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content. [10] The

Data-PASS organization promotes collaborative, institutional stewardship of research data, permanent data archiving, and citation that permits results to be verified and re-purposed. [11] DataCite collaboratively addresses the challenges of making research data visible and accessible through data citation.[12] The Creative Commons project, Science Commons, has focused on protocols for sharing scientific data that includes licensing and mitigating legal barriers.[13]

Summary of Major Recommendations

- Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Establish policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Foster public values and support for stewardship of digital data beyond mandating data management plans.

- Leverage substantial national and international efforts for common practices that support interoperability.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI. The responses below exemplify how the principles can be applied to the individual questions, and highlight relevant NDSA activities in these areas.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The most effective policies in this regard would mandate data deposit into publicly accessible repositories. In the absence of such a policy, there are already cases of data which have been lost. The Federal policy framework should move public access to data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use.

Many members of NDSA provide repository services at low cost or through cooperative arrangements. Members of the NDSA also provide repository services that provide legal, technical, procedural and statistical controls necessary to protect data confidentiality while ensuring long. And the NDSA provides a model of institutional collaboration that supports stewardship, discovery and accessibility An example of a free access service is ViewShare.org, a platform for empowering curators, archivists, and librarians to provide access to the digital collections they are preserving through a shared interface. This

service provides the dual benefit of making data more broadly available and accessible while also making it easy for end users to copy and make use of the data in other environments. [14] The NDSA content working group is also working toward developing a clearinghouse for at-risk digital collections to help match data to potential preservation partners.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Each domain and discipline should be empowered to set priorities for data selection through, level of curation, and length of retention, through professional societies or other consensus making bodies.

Notwithstanding, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. For most data, "open access" is needed not only for the short term, but for the long term. And scientific disciplines have focused primarily on short-term access. There are critical standards for metadata exchange, fixity information and verification, and persistent citation that can support long-term access to data, preservation, and the long-term reproducibility of public results. Such baseline standards should be applied all scientific data. Among the range of important new standards for preservation and access there is still little knowledge about which standards are being implemented in which situations. The NDSA Standards working group is working on inventorying these standards and exploring how they are currently being used by NDSA member organizations. More than advocating the need for standards there is a clear need to understand which standards are being used in which situations and use that information to promote the usage of standards that are leading to results.

(5) *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., the Data-PASS archives). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation being used loosely without associated evidence of meeting specific requirements. Memory institutions such as archives, libraries and museums have an extensive track record with these functions and collaborative organizations such as NDSA could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions. In this respect, work from the NDSA innovation working group toward developing a "Neighborhood Watch" system for repository quality assurance could serve as the basis for establishing clear, externally verifiable reporting. [14]. The group has identified a pressing need for

an objective, repeatable, independently verifiable and simple way for an external agent to periodically retrieve content, verify its bit level integrity and publicly announce the results. This is a clear example of how assertions about data management could be tested and certified.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The most important step would be to communicate that the real costs of preserving and making digital data accessible are indeed legitimate and necessary costs of the overall research enterprise. Researchers routinely include publication costs within their research proposals -- the costs of ensuring long-term access reuse of data should be treated in the same way.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Many NDSA partners are leaders in this area. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

References

[1] The National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa

[2] Berman, Francine, and Brian Lavoie, et al. 2010. *Sustainable Economics for a Digital Plant: Ensuring Long-term Access to Digital Information.* Final Report of the Blue

Ribbon Task Force on Sustainable Digital Preservation and Access supported by the National Science Foundation, et al. Washington, DC: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Library of Congress. 2010. *Preserving our Digital Heritage: The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report*. Washington, DC: http://1.usa.gov/hmw2lj

[4] Alliance for Permanent Access. 2011. "Opportunities for Data Exchange (ODE) Project": http://www.alliancepermanentaccess.org/index.php/current-projects/ode/

[5] King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173-99.

[6] National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

[7] Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

[8] Berman et al. 2010.

[9] Data Seal of Approval: http://www.datasealofapproval.org/

[10] LOCKSS: http://lockss.org

[11] DataPass: http://data-pass.org/

[12] DataCite: http://datacite.org/

[13] Creative Commons project, Science Commons: http://creativecommons.org/science http://wiki.creativecommons.org/Science

[14] ViewShare: http://viewshare.org

[15] Abrams, S, Cruse, P, Kunze, J, Minor, D, Smorul, M. 2011."Neighborhood Watch" for Repository Quality Assurance. Presented at Designing Storage Architectures for Preservation, Washington, DC: http://1.usa.gov/uXj2Mf

Tue 1/3/2012 8:28 PM
Response to OST RFI on public access to digital data

Name: Victoria Reich

Organization: Stanford University Libraries

City, State: Stanford, California

Comment 1:
I am writing to endorse the National Digital Stewardship Alliance (NSDA) response to the OST RFI on public access to digital data resulting from frederally funded scientific research, dated January 2, 2012. This response is attached.

NDSA
NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

Submitted by the National Digital Stewardship Alliance (NDSA)

January 2, 2012

Introduction to the NDSA

The National Digital Stewardship Alliance (NDSA) was founded in July 2010 to extend work begun in 2001 by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress. The Alliance has over 100 members from educational institutions, non-profit organizations, businesses and local, state and federal government agencies, as well affiliations with international organizations. Its mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. [1] Members of the Alliance are taking action to preserve access to our national digital heritage by:

- broadening access to our nation's expanding digital resources
- developing and coordinating sustainable infrastructures for the preservation of digital content
- advocating standards for the stewardship of digital objects
- building a community of practice around the management of distributed digital collections
- promoting innovation
- facilitating cooperation between government agencies, educational institutions, non-profit organizations, and commercial entities
- fostering the participation of diverse communities and relationships across boundaries
- raising public awareness of the enduring value of digital resources and the need for active stewardship of these national resources.

Supporting communities of practice for preservation and access

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally-funded scientific research. When applied, these values support the practical development of communities of practice capable of gaining consensus to support preservation and access to digital data. The shared expertise and common experience of these communities result in stakeholder buy-in and adoption of policies and

standards. The National Digital Stewardship Alliance member organizations are bound as a community by the following values.

Stewardship. Members of the NDSA are committed to managing digital content for current and long-term use. The members of the NDSA are actively ensuring sustained access to the digital content that constitutes our national legacy and empowers us as leaders in the global knowledge economy. Individually, these organizations support the management of digital resources; the Alliance is committed to protecting our nation's cultural, scientific, scholarly, and business heritage.

Collaboration. Collaborative work is the centering value of the Alliance; it is a value shared by all members and a priority in work with all organizations and associations. Approaching digital stewardship collaboratively allows the NDSA to coordinate effort, avoid duplicate work, build a community of practice, develop new preservation strategies, flexibly respond to a changing economic landscape, and build relationships to increase capacity to manage content beyond institutional boundaries.

Inclusiveness. The NDSA is a collaborative effort to preserve a distributed national digital collection for the benefit of current and future generations. We value the range of experience, the potential for innovation, and the fault-tolerance that heterogeneity brings. We believe the preservation of digital information is a pervasive challenge and that engaging across different communities strengthens the nation’s digital preservation practices and increases the likelihood of preserving content now and into the future.

Exchange. Members of the Alliance encourage the open exchange of ideas, services, and software. This leverages the commitments of each member to increase the capacity of the entire stewardship network. Participation and engagement result in innovations and benefits that can be shared by all. The Alliance is committed to transparency and all products generated or produced by the Alliance will be circulated under open licenses.

Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

Community-based approaches to the challenges of rapid change and high volume within the data domain have proven to be the most successful in the long term. The Blue Ribbon Task Force on Preservation and Access recommended that for research data “Each domain, through professional societies or other consensus making bodies, should set priorities for data selection, level of curation, and length of retention.” [2]

The report validated experience over the last ten years of digital preservation work. A study of the networks developed through the NDIIPP program indicated that participating institutions bring to the network their own resources, interests, and organizational culture. Under the auspices of a neutral convener and honest broker, natural networks emerge over time through participation in shared activities and problem solving. As these networks form, the larger network becomes more complex, but also stronger and better able to withstand stresses and strains. [3]

The Opportunities for Data Exchange (ODE) project supported by the Alliance for Permanent Access and the European Union also takes a cross-cutting community approach to preservation and access to digital data. "The potential answers to grand challenges of our times require…the inclusion of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-infrastructures…All stakeholders in the scientific process must be involved in the design of this layer; policy makers, funders, infrastructure operators, data centers, data providers and users, libraries and publishers…" [4]

An exemplar of collaborative community efforts is the Dataverse Network project [5] recently described by the National Research Council of the National Academies as the "State of the Practice in Data Sharing." [6] The Dataverse Network is "unique in being designed to explicitly support long-term access and permanent preservation. To this end the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats." [7]

Foster public values and support for stewardship of digital data beyond mandating data management plans.

Policy should assert the value of research data and provide mechanisms to support the preservation, discoverability and access. To relieve frustration and confusion about actions the policy should provide a clear direction for funders, researchers and stewardship organizations. The Blue Ribbon Task Force recommended "Funders should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving. They should explicitly recognize "data under stewardship" as a core indicator of scientific effort and include this information in standard reporting mechanisms." [8]

Leverage substantial national and international efforts for common practices that support interoperability.

Substantial efforts have been made to pave the way for interoperability, re-use and re-purposing. Emerging practices for data citation, licensing and protocols for data sharing and sustainable re-use are becoming enough to adopt more broadly. Notable in these areas are work on the Data Seal of Approval by the Data Archiving and Networked Services that promotes sustainable access to digital research and provides training and advice about archiving and reuse.[9] LOCKSS is a community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content. [10] The

Data-PASS organization promotes collaborative, institutional stewardship of research data, permanent data archiving, and citation that permits results to be verified and re-purposed. [11] DataCite collaboratively addresses the challenges of making research data visible and accessible through data citation.[12] The Creative Commons project, Science Commons, has focused on protocols for sharing scientific data that includes licensing and mitigating legal barriers.[13]

Summary of Major Recommendations

- Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Establish policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Foster public values and support for stewardship of digital data beyond mandating data management plans.

- Leverage substantial national and international efforts for common practices that support interoperability.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI. The responses below exemplify how the principles can be applied to the individual questions, and highlight relevant NDSA activities in these areas.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The most effective policies in this regard would mandate data deposit into publicly accessible repositories. In the absence of such a policy, there are already cases of data which have been lost. The Federal policy framework should move public access to data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use.

Many members of NDSA provide repository services at low cost or through cooperative arrangements. Members of the NDSA also provide repository services that provide legal, technical, procedural and statistical controls necessary to protect data confidentiality while ensuring long. And the NDSA provides a model of institutional collaboration that supports stewardship, discovery and accessibility An example of a free access service is ViewShare.org, a platform for empowering curators, archivists, and librarians to provide access to the digital collections they are preserving through a shared interface. This

service provides the dual benefit of making data more broadly available and accessible while also making it easy for end users to copy and make use of the data in other environments. [14] The NDSA content working group is also working toward developing a clearinghouse for at-risk digital collections to help match data to potential preservation partners.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Each domain and discipline should be empowered to set priorities for data selection through, level of curation, and length of retention, through professional societies or other consensus making bodies.

Notwithstanding, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. For most data, "open access" is needed not only for the short term, but for the long term. And scientific disciplines have focused primarily on short-term access. There are critical standards for metadata exchange, fixity information and verification, and persistent citation that can support long-term access to data, preservation, and the long-term reproducibility of public results. Such baseline standards should be applied all scientific data. Among the range of important new standards for preservation and access there is still little knowledge about which standards are being implemented in which situations. The NDSA Standards working group is working on inventorying these standards and exploring how they are currently being used by NDSA member organizations. More than advocating the need for standards there is a clear need to understand which standards are being used in which situations and use that information to promote the usage of standards that are leading to results.

(5) *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., the Data-PASS archives). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation being used loosely without associated evidence of meeting specific requirements. Memory institutions such as archives, libraries and museums have an extensive track record with these functions and collaborative organizations such as NDSA could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions. In this respect, work from the NDSA innovation working group toward developing a "Neighborhood Watch" system for repository quality assurance could serve as the basis for establishing clear, externally verifiable reporting. [14]. The group has identified a pressing need for

an objective, repeatable, independently verifiable and simple way for an external agent to periodically retrieve content, verify its bit level integrity and publicly announce the results. This is a clear example of how assertions about data management could be tested and certified.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The most important step would be to communicate that the real costs of preserving and making digital data accessible are indeed legitimate and necessary costs of the overall research enterprise. Researchers routinely include publication costs within their research proposals -- the costs of ensuring long-term access reuse of data should be treated in the same way.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Many NDSA partners are leaders in this area. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

References

[1] The National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa

[2] Berman, Francine, and Brian Lavoie, et al. 2010. *Sustainable Economics for a Digital Plant: Ensuring Long-term Access to Digital Information.* Final Report of the Blue

Ribbon Task Force on Sustainable Digital Preservation and Access supported by the National Science Foundation, et al. Washington, DC: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Library of Congress. 2010. *Preserving our Digital Heritage: The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report*. Washington, DC: http://1.usa.gov/hmw2lj

[4] Alliance for Permanent Access. 2011. "Opportunities for Data Exchange (ODE) Project": http://www.alliancepermanentaccess.org/index.php/current-projects/ode/

[5] King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173-99.

[6] National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

[7] Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

[8] Berman et al. 2010.

[9] Data Seal of Approval: http://www.datasealofapproval.org/

[10] LOCKSS: http://lockss.org

[11] DataPass: http://data-pass.org/

[12] DataCite: http://datacite.org/

[13] Creative Commons project, Science Commons: http://creativecommons.org/science http://wiki.creativecommons.org/Science

[14] ViewShare: http://viewshare.org

[15] Abrams, S, Cruse, P, Kunze, J, Minor, D, Smorul, M. 2011."Neighborhood Watch" for Repository Quality Assurance. Presented at Designing Storage Architectures for Preservation, Washington, DC: http://1.usa.gov/uXj2Mf

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research (PSTP request dated 4 Nov 2011 to implement Section 103 of America Competes Reauthorization Act of 2010)

submit electronically to: digitaldata@ostp.gov

2 January 2012

Comments of Paul Beier, paul.beier@nau.edu, School of Forestry, Northern Arizona University, Flagstaff AZ 86011-5018, 928 523 9341

My qualifications: I a wildlife ecologist who has engaged on research related to conservation of mammals, birds, reptiles, and amphibians for over 30 years. I am currently President of the Society for Conservation Biology, the largest professional society of conservation scientists. I am submitting these comments as an individual, but I believe they reflect the informed opinions of many other conservation biologists.

Scope of my comments: I confine my remarks entirely to one very simple type of biological data – namely the locations of species occurrences. These data consist solely of species name, date, x-coordinate, y-coordinate, and metadata that will allow users to infer the spatial resolution (precision) of the method used to determine the locations. Henceforth I refer to these as **species occurrence data**. These data (when used in conjunction with other freely available data on land cover, soils, topography, and the like) can be used to map species distributions and identify areas of suitable habitat that have not been surveyed for the species of interest. The data and subsequent analyses would inform management decisions by federal agencies, landowners, and others wishing to promote conservation of these species. Such data will be particularly useful for modeling shifts of species' ranges in response to climate change.

I address this one type of data because this data type is simultaneously important and simple. If the federal government cannot promote sharing of species location data, it will utterly fail to promote sharing of more complex types of data and metadata collected with federal funding.

My experience related to lack of availability of species occurrence data: Several experiences lead me to believe that sharing species occurrence data is crucial to advancing the science of conservation biology. I have served on 3 recovery teams for species listed under the Endangered Species Act, and on several panels attempting to make recommendations for other species at risk. In every case, our work was severely constrained by lack of access to existing species occurrence data. In every case most of these data had been acquired with federal funds, and in some cases the data had been acquired by federal agencies. With access to these data, we could have built reliable, empirical models of suitable habitat for the species of concern. Although many researchers voluntarily shared data in response to our requests, some of them no longer had access to the raw location data (e.g., because they had left their previous institution) and others never found the time to comply with our request. The process of request was cumbersome and slow because it required us to find out THAT the data existed before we could request it. In no case were we able to gather all of the relevant data we needed. This situation is inexcusable, and is depriving the public of the benefit of information that has been collected at taxpayer expense.

The following paragraphs are organized around specific questions in the request for information.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?"

I strongly believe that all federal agencies funding research related to biodiversity should require PIs on federally funded research projects to deposit species occurrence data with NBII as a condition of funding. PI's who have not complied with this requirement on a past grant or contract should not be eligible for future federal grants and contracts.

Arizona Game & Fish Department has funded several of my research projects, and has required me (and all funded PIs) to provide location data to their database management system before providing the final payment. Compliance has been almost painless for me as a researcher. The Department fuzzes the data to prevent abuse, and contacts me before sharing the data outside the regulatory agencies.

I have submitted one proposal to NSF under their new guidelines, which require a Data Management Plan as part of the proposal. In its current form, the NSF requirement for a data management plan is a feeble step forward that will do very little to improve data sharing. The data management plan, for instance, does nothing to alert the community of potential users that the data exist, nor does it provide a single access point for data from multiple PIs. It also cannot be monitored and enforced.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

The NBII should allow PI to specify how species occurrence data may be used during an initial period of, say, three to five years. For instance, the PIs could specify that no user can use the data to investigate the particular issues that the PIs designed their project to address. PIs could also specify that no peer-reviewed publications using the data can be submitted during the period of restricted use.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Some species occurrence data are sensitive, in that unscrupulous persons could use them to kill, capture, or harm individual animals or plants. Many agencies (Arizona's Heritage Database Management System, for instance, and other state databases) have already solved this problem using two simple measures: (1) The publicly available data consist only of low-resolution maps with locations "fuzzed" by up to a few km. This provides enough preliminary information for a potential user to determine if the data cover the area of interest to the user. (2) Precise location data are provided only to legitimate requestors who agree to specific terms on use of the data, including agreements not to depict or share precise locations in any way.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

For species occurrence data, the costs are miniscule and the benefits are large. I suggest that OSTP might for now require data sharing only for similar types of low-cost high-benefit data. OSTP could use the experience to start to produce reliable estimates of long term costs and benefits that could be used to guide future decisions.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

The Society for Conservation Biology (of which I am president) publishes several scientific publications. SCB could work with our publisher to require authors to archive their species location data with NBII. However, if only SCB took this step, some authors would submit elsewhere to avoid this extra responsibility. But a broad consortium of professional societies in ecology (SCB, Ecological Society of America, The Wildlife Society) and a handful of dominant publishers (e.g., Wiley-Blackwell, Elsevier, Springer-Verlag) could create a new culture in which data-sharing is viewed as a responsibility of publishing. I have appointed a Task Force in SCB to investigate how SCB could start a dialogue with our sister professional societies and the publishers of their journals to start to create this culture. It will take years, and there will be strong resistance from some academic PIs, but I believe this is an achievable long-term goal. Again, I think it makes sense to start with low-hanging fruit (e.g., species occurrence data); once the new culture of sharing has been in place for a few years, I think it will become obvious which other types of data to share, and how to share them.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The federal government should fund the NBII at a level that would bear virtually all the costs of preserving and sharing digital data on species occurrences.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

For some types of data, ensuring compliance will be difficult, but it should be relatively easy for species occurrence data. Federal funders of biodiversity-related research (NSF, USDA, DOD SERDP, EPA) could require the Data Management Plan in each proposal to list the species for which occurrence data will be collected. Funders should convey this information to NBII, who would need staff persons to track compliance and report non-compliance to all federal funders.

One more drastic measure is worthy of consideration: NBII could identify institutions with a pattern of non-compliant PIs and bar such institutions from future federal grants and contracts. This would motivate universities and other research institutions to monitor compliance.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Most scientists are reasonable people and realize that acknowledgments will facilitate future collaboration with other investigators. If asked to acknowledge data providers, most scientists will gladly do so.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

I believe NBII already collaborates with GBIF. Such collaboration for species occurrence data should be simple.

Thank you for the opportunity to comment on this important initiative. The Policy Committee and Science & Publications Committee of SCB are also considering your request for information, and will probably provide comments, which may or may not agree with mine. In any event, I would be happy to discuss with OSTP how to move this important initiative forward. I truly believe it is very important to conservation biology..

/s/ Paul Beier

January 4, 2012

Mr. Ted Wackler
Deputy Chief of Staff
Office of Science and Technology Policy
725 17th Street
Washington, DC 20502

Re: Public Access to Digital Data Resulting From Federally Funded Scientific Research (76 FR 68517)

Dear Mr. Wackler:

I am writing on behalf of Oxford University Press (OUP) in response to OSTP's Request for Information regarding "Public Access to Digital Data Resulting from Federally Funded Scientific Research".

The world's largest university press, Oxford University Press is an international publisher of scholarly and educational material with offices across the globe including major centers in New York City and in Cary, NC. OUP furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide and publishes over 270 peer reviewed scholarly journals (most of which are published in partnership with learned societies) and more than 2000 research monographs per year .

Preservation, Discoverability, and Access
As a scholarly publisher working closely with the academic community, OUP supports efforts to make data more available and more discoverable, and to ensure that appropriate data is preserved and curated for future use.

The first step is the creation of an environment and infrastructure where investigators are able and encouraged to deposit their data. While data management is built-in to the experimental design of so called 'big science', data management for the remaining 90% of scholarly endeavor could be improved. To this end, federal agencies should work with stakeholders including researchers, learned societies, library groups, and publishers and:

- set aside funds to support the establishment of subject-specific archives for data where they do not already exist
- develop policies and guidance on what data should be deposited
- develop policies which encourage researchers to deposit data and which balance the needs of investigators to gain credit by utilizing the data that they produce with the opportunities to advance knowledge by sharing data with other researchers and with the public
- encourage the development of metadata standards that can be used to describe data in order to improve discoverability
- develop preservation criteria regarding which data should continue to be preserved and curated

Standards for Interoperability, Re-Use and Repurposing
We see the key issues with respect to interoperability and use of primary research data as the development of:

- standards for the bi-directional linking between primary data and the peer reviewed research literature, and data citation
- standards for the acknowledgement of the use and re-use of data
- clear rules on modification of source data and how this modification is described
- security protection protocols to guard against unauthorized modification, damage, or deletion

OUP notes that standards for data stewardship are currently at an embryonic stage but that the following are examples of good projects / initiatives that we respectfully suggest Federal agencies should engage with:

APARSEN (www.alliancepermanentaccess.org/index.php/currentprojects/aparsen)
CASPAR (www.casparpreserves.eu)
CoData (www.codata.org)
DataCite (datacite.org)
DCC (www.dcc.ac.uk/)
DRYAD (www.datadryad.org)
nestor (www.langzeitarchivierung.de)
NISO/NFAIS Supplemental Journal Materials Working Group (www.niso.org/workrooms/supplemental)
OAIS (public.ccsds.org/publications/archive/650x0b1.pdf)
Opportunities for Data Exchange (www.ode-project.eu)
PARSE.insight (www.parse-insight.eu)
Planets (www.planetsproject.eu)
SHAMAN (www.shaman-ip.eu)

Yours sincerely,

Niko Pfund
President and Academic Publisher
Oxford University Press

Wed 1/4/2012 8:33 PM
National Digital Stewardship Alliance (NSDA)

I am writing to endorse the National Digital Stewardship Alliance (NSDA) response to the OST RFI on public access to digital data resulting from rederally funded scientific research, dated January 2, 2012.

Mary Ellen Petrich
Stanford University
Palo Alto CA

--
Mary Ellen Petrich
Content Specialist, LOCKSS
http://www.lockss.org/

January 5, 2012

To: Office for Science and Technology Policy (OSTP)
digitaldata@ostp.gov

From: Gary R. VandenBos, PhD, Publisher
gary@apa.org

Steven J. Breckler, PhD, Executive Director, Science Directorate
sbreckler@apa.org

American Psychological Association
750 First St., NE
Washington, DC 20002

Re: Request For Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research (FR Doc. 2011-28628)

Dear Sir or Madam,

On behalf of the American Psychological Association, we are writing to respond to the Request for Information requested by the Office of Science and Technology Policy (OSTP) in the Federal Register (Volume 76, Issue 214) of November 4, 2011, seeking public input on "approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research." We welcome this opportunity to provide recommendations regarding this matter.

As a professional association, we view public access to digital data as a primary objective. Thus, our publishing policies are moving in a direction that promotes the culture of data sharing in the field of psychology.

As a scholarly publisher, we believe that better discoverability and re-use of original research data are to be encouraged at all levels and among all stakeholders involved. As most publishers do, APA supports the view that Federal agencies should work with researchers and other stakeholders to create appropriate policies to make digital data resulting from federally funded scientific research freely available to the public. Every stakeholder has an important role to play. Governmental and other funding agencies have a special contribution to make in identifying international standards and best practices for the management of primary scientific data generated by taxpayer or other research grant funding. This role could also include standards for the interoperability of data repositories with the published research literature. To ensure that deposited datasets become an integral part of the record of science over the long term, publishers would encourage the establishment of common practices around the bi-directional linking of data and publications and around standards for the citation of data. We encourage agencies to investigate and establish contacts with a number of initiatives already underway or recently concluded that are examining data stewardship and public access issues in this context.

Along with other scholarly and professional publishers, APA recommends that Federal grants allocate specific funds to support researcher data management and deposit efforts, and to support the

establishment of discipline-specific data archives, particularly in areas such as psychology for which such mechanisms are not well developed.

Federal policies should establish clear rules for citation of data sets and acknowledgement of modifications to source data and should recognize the costs associated with hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term. This is important for creating incentives and rewards for data sharing, and to support recognition of data creators and collectors.

To foster greater legal certainty for data users and producers, APA recommends that Federal policy give clear direction as to what data may be shared publicly and establish penalties, e.g. grant bans for those who willfully misrepresent or distort data created by others, for the misuse or abuse of data.

Most publishers suggest that the Federal government investigate policies that create an incentives hierarchy for scientists to share their data – with the greatest reward for those who publish data with articles, but recognition also for those who publish data only or those who publish so-called descriptive data-publications

Along with other scholarly publishers, APA strongly supports the view that the Federal Government should be guided by "principles of transparency, participation and collaboration" as noted in the Transparency and Open Government Memorandum and Open Government Directive. We stand ready to work in collaboration with all partners to ensure the continued success, vibrancy, and innovation of the U.S. scientific community.

In addition to our general remarks above, we would like to comment specifically on some of the questions outlined in OSTP's Request For Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research, as follows:

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

APA recommends that Federal policies establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. Penalties should be established for the misuse or abuse of data and technical measures that ensure ongoing data integrity should be put in place. Key policy terms should be clearly defined to differentiate between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets'). To increase legal certainty for data users and producers, clear direction should be given as to what data may be shared publicly and what may not.

Federal policies should recognize that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost which, in certain circumstances, the host should be entitled to recover. Databases themselves – i.e. collections of data specifically organized and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – merit intellectual property protection, under copyright or database protection principles. These databases are often characterized by the sophistication of their data field structuring, searchability tools, and contain valuable and useful information for scholarly research. The value of individual researcher-validated data sets is different from larger-scale databases that have been organized and compiled to serve particular research needs.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

APA recommends that Federal agencies should work with researchers and other stakeholders to develop appropriate policies to make digital data resulting from federally funded scientific research freely available to the public. Where no standards or commonly accepted practices for making digital data publicly available are established, APA believes that the government has an important role to play in working with key stakeholders such as researchers and publishers to develop best practices that will advance scholarly communication and the public good.

APA also recommends that those disciplines—such as psychology—that currently do not have a well-developed infrastructure for subject-specific data repositories receive financial support to establish them. To ensure that data repositories are reliable, safe and secure long term preservation of the data, such data repositories should be subject to certification and audit procedures. The facility to link datasets and publications at the level of the data set should be a condition in such certification procedures, next to all necessary preservation requirements for the long term.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Key elements in successfully implementing data management plans include the following:

- Stakeholders can engage with their communities to encourage a culture of data sharing.
- Data management policies established in collaboration with researchers and other stakeholders such as publishers.
- Requirements for data management plans should be clear, complete and unambiguous, and they should specifically address liability issues.
- Data management policies should consider the practices of different research communities. They should be developed in collaboration with representative bodies of all stakeholders who will likely be affected – e.g. researchers, funders, publishers, data repositories, etc. In relevant fields of science, collaboration with publishers and editorial boards can help establish clear policies for the availability of research data in the context of publications that analyze and interpret federally funded research.
- Training courses, e-learning modules, and FAQs should be created for researchers to gain a more complete understanding of data management plan requirements as well as the data deposit process.
- Specific grant funds should be available to support data management and deposit activities.
- Incentives for researchers to deposit data after a clearly-defined and collaboratively set time frame should be provided as well as penalties for noncompliance.
- Data deposit, integrity, provenance, and access at repositories should be user-friendly, efficient and clear.
- Bi-directional linking between datasets in data repositories and publications is to be encouraged by clear citation guidelines to ensure that datasets become part of the record of science.
- Data repositories should be certified and audited. Linking possibilities at the level of specific datasets should be among the certification conditions. Researchers should not be required to maintain the accuracy or integrity of the data once it has been deposited but depositing researchers should have the right to modify or correct data they have deposited. Liability policies should protect researchers if data are corrupted or lost.
- The administrative burden on researchers should be kept to the lowest minimum possible.

APA believes that there is no one stakeholder, (e.g. publisher, government, research institution, library, university) or data repository that has, or should have, a monopoly on any of these activities. Stakeholders should work collaboratively to address these issues.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Federal agencies must recognize the time costs associated with data attribution and credit, and provide ways to fund these efforts at both the institutional level and at the level for those receiving grants.

APA recommends the endorsement of common practice and rules for the citation of datasets and linking between datasets and publications. Federal policies should establish clear rules for citation of data sets and acknowledgement of modifications to source data. They should also provide for the establishment of security protocols that protect stored data from unauthorized modification, damage or deletion and liability arrangements if data are lost or corrupted.

In this context, APA recommends that Federal agencies seek collaborations with DataCite (see http://datacite.org/). DataCite is a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, DataCite is currently working primarily with organizations that host data, such as data centers and libraries.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

As mentioned in the answers to other questions, APA supports the endorsement of standards for the citation of data and for the establishment of common practices around the linking between deposited datasets and related publications. In this context, APA recommends that Federal Agencies become involved with three initiatives already well underway in this area:

- Opportunities for Data Exchange (ODE, www.ode-project.eu) – whose aim is to gather and promote best practices around the way scientific data are treated. Its Report on Integration of Data and Publications is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/.

- The NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental). This group is preparing an initial draft of its recommendations, which are expected to include minimum metadata elements recommended to describe supplemental materials and establish their relationship to the main article, as well as optional elements that will more comprehensively characterize materials for future applications. A non-normative Document Type Definition (DTD) is also expected in draft form. This DTD is not intended to be, or become, an official standard. Instead, it is intended for use as a model to more precisely define a hierarchy for the recommended metadata, and as a starting point for organizations seeking to adhere to the NISO/NFAIS recommendations.

- DataCite (http://datacite.org/), a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be

verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, DataCite is currently working primarily with organizations that host data, such as data centers and libraries.

Thank you for this opportunity to offer APA's recommendations regarding public access to digital data resulting from federally funded scientific research.

Fri 1/6/2012 7:08 AM
Response to questions on Public Access to Digital Data

Submitted by: Walter S. Snyder

Department of Geosciences

Boise State University

Based on 10 years experience in geoinformatics, including being involved is starting the geoinformatics program at NSF, development and management of several community data systems (GeoStrat (NSF supported); Geothermal Data Exchange (GDEx) (DOE and NSF supported), the National Geothermal Data System (DOE)), and continued national and international collaborations on data issues.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The biggest mistake that the federal government can make is to assume that the agencies themselves will always be the best stewards of the data generated by federal funds. There is at least a two-tier issue here. First, data generated by personnel of the agency and through procurement contracts are perhaps best preserved by the agency itself (and therefore the agency must provide access to these data). Second, for data generated by extramural funds, through grants and cooperative agreements, the agency itself may not be, and perhaps in most cases won't be, the best stewards for these data. This issue is one of who can best determine the user's needs re the data - starting from data generation to management to working with data to publication to long-term, open access? Typically agency personnel are not the best people to make those decisions, and whereas their input is valuable, the decisions of how to construct and operation external data systems must reside with the user communities. It is important to note that these users are precisely the ones targeted by the COMPETES Act, and it would be presumptuous for agencies to assume they know better than the users what needs to be done. The approach to this two-tiered problem varies by agency and within each agency - and this is not a surprise. For example, in general NSF understands the importance of the users, and almost errs on the side of being too flexible and allowing unsustainable approaches to data on a project-by-project basis. For many other federal agencies, is completely understandable that they want to have and serve all data their funding has paid for and put these data on agency-controlled servers. servers and/or data sites - including that generated by extramural funding.

Two things here - this is fine for internal agency data, but is not optimal for data derived from extramural funding. This will lead to lower quality and incomplete data simply because it is a forced approach to data acquisition and management versus one that comes from and is for the user and data producing communities.

In short, policies must recognize that the agencies themselves have different missions and mandates than the user communities, and no matter how much one writes or talks about it, agencies are not representatives of the user communities. The agencies have to first and foremost worry about their own existence as a business entity, placed second is their role (depending on the agency or group within the agency) as a servant of the public, the user communities. That is not necessarily a bad thing, it is what it is and agencies should be the stewards of the data they produce for themselves, but they should not control (but can and should participate in) facilities that manage data generated by extramural funds. These need to be agency-supported community systems.

Policy: Data generated internally by an agency can be hosted by the agency; data generated by extramural funding of grants and cooperative agreements should be funded by the agency, but hosted by community-based data sites if at all possible.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Within a document, there is a need to distinguish between the written word and data contained in the document. The data, if it is the result of federal funding, should be immediately available to the public, and at a minimum be able to be used by anyone free of charge or other restrictions. The interpretations of those data however should be given the same copyright protection as any written document. These interpretations may have been possible because of federal funding, but they are the intellectual creation of individuals beyond the boundaries of a particular batch of funding and should follow standard approaches to intellectual property, including the individual passing the copyright to the publisher.

Policy: the data generated by federal funding should be freely and openly available within a reasonable time frame, but the interpretations of these data as presented in published and unpublished documents remain the intellectual property of their author or the author's assignee.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

This is a very insightful question and focuses on a critical issue - is there one best way to handle all digital data? Is there a singular approach that will work for all needs and communities? The answer is "no" - although there are many who think a single way can be created. The problem with that singularity view is that it would have to be forced, and in the process loose critical knowledge value of the data - this is a topic for a longer discussion. At a high level, data sharing among agencies will be possible by developing high-level standards for data discovery and sharing, but that should not dictate how data are captured, stored or even served to particular user communities.

Policy: each agency must assess the context and use goals of the data generated by their funds for two distinct groups of data: 1) internal data, and 2) for data generated by extramural funding. For this second group, the agency must consult with or assign this assessment process to users outside of their or any agency (i.e., to the user community). The agency should then compare their needs to what other agencies, groups and institutions are doing, and establish their own best practices.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

This is a very difficult question, because assessing the value of data is such a vague endeavor. Is it all about the cost of re-generating the data? Is it about its intrinsic value to understanding something else? Data coming from machines and sensors is far easier to capture and store - the problem being the quantity of the data - but that too is not longer a major problem. Some data have very long "shelf lives", e.g., geologic data, others very short life, e.g., medical research data. So one could ask agencies to perform a qualitative assessment of the data groups they handle - the emphasis here is "qualitative" - if you try to force a quantitative assessment, then you will, from the start, under-value some groups of data.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

These "stakeholders" or "users" need to self-organize, and that may require some funding from the relevant agency. They need to address all the questions poised here and many more. Then their recommendations need to be implemented by the relevant agencies - at least addressed in

an open way that also allows stakeholder rebuttal to a higher authority. OSTP could establish a clearing house for such input. Why is this important? Because the agency that provides the funding to the stakeholders can have undue influence and control. An a priory notion that the agency always knows best, while true the vast majority of the time, is not always the case, and the stakeholder has no recourse; OSTP should provide that recourse.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

We can speak from personal experience here. First, the agency needs to realize that preserving and making digital data accessible is not inexpensive, and a rule of thumb may be a minimum of about 3% of their operational budget for funding outside community data centers and perhaps 9% (total cost assessment) internally. Second, the agencies need to fund several key "community data centers" that not only handle data from extramural funding, but could also handle some of the agency's internal data at a cost lower than the agency can do internally (on a "total cost" basis). For extramural funds, if the agency funds one or more community data centers, then subsequent cooperative agreements and grants can utilize these centers for their data management, and include modest amounts in their budgets to pay to the center for this service. Thus, the agency funds the core operations of a data center, and each subsequent award pays for its data management needs in an affordable way.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Each project, whether internal to the agency (including procurement contracts) or extramural funded projects could produce a data management plan that details the data they intend to generate. This is not a difficult or burdensome exercise, and could take as little as two hours per project. Then there is a metric for comparison. These plans need to be open to amendment as the project proceeds. Then, if systems exist that can work with the data producers to capture the data as close in time to when they are generated as possible, you decrease the burden on the data producer, improve the amount and quality of data captured, and allow for an ongoing assessment of progress on data capture by each project. This need not be burdensome if it is integrated into the daily workflow.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

This can primarily be done through the establishment of "community data centers" - these can produce jobs for each center, but also potentially spin off collaborations with industry and business that may want to use the data in innovative ways. Simply put, if all the expenditure for preserving and making digital data accessible is contained within the agency there will be little stimulation of outside jobs and innovative outside use of these data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Again, this is from personal experience. In our data systems we now have established a first tie to a professional science society where the society, through its publications, will provide Digital Object Identifiers (DOI) for datasets. In addition our data system will host all data referred to in their publications (journals and books), including the said datasets and provide open, free access to these data. Thus the creator of a dataset gets publication credit for the data as well as the published paper, and each time another person references that dataset in subsequent publications receives another citation. This becomes a win-win situation that reduces the costs to the science society, and provides an easy and consistent way for the data producers to manage and get credit for their data.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The underlying problem is that there are numerous "community-driven data standards efforts" - so which ones do you use? Also, there is lots of hype around particular approaches, and although OSTP seems to have bought into the full story of the "Semantic Web" - other experts (see recent PEW report) have noted that it has not been as successful as billed. Nevertheless there are many "standards" that are now commonly used - for example the semantic web's web services, OGC/FGDC geospatial standards, etc. Some data groups are easier to set standards for than others - for example sensor/sensor array data is relatively easy, but science data based on field and laboratory analyses is far more difficult. Again, the key is what is implied in the question; to allow user communities to come together and work, over time on developing and refining the way they approach data. Agencies should follow these community-based groups, not tell them what to do. Convergence will happen as several studies point out, if you let it happen naturally and do not force it. A note of caution: defining a user community is difficult as well. What we have discovered is that the academic user community is quite distinct from that of state and federal agencies and that, whereas you want dialogue and to promote a path

towards convergence among academia and the agencies, initially these communities should acknowledge that they have different mission and mandates and that they need to get their own houses in order first - or you run the risk of the more powerful partner forcing "standards" which are not the best for the particular community and don't last.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Environmental Information Exchange Network (www.exchangenetwork.net) which has been operating for 7 years, was developed such that each node on the network can continue to do what its individual mission and mandate requires, yet a series of data exchange templets have been developed based on network-wide (i.e., community) input and agreement that allows the seamless exchange of data among the nodes on this network. An example of one that has not worked well is the U.S. Geoscience Information Network (USGIN); the story is in the details, but it has been imposed from effectively a single source and does not represent development in an open, true community environment, and will survive only as long as it has political support. The official written descriptions capture correct sentiments, but the key is how it has developed and its implementation. It seems to fit the needs of select state agencies and the US. Geological Survey, which is fine, but when it is forced onto academic and industry communities it fails because of the lack of openness and inclusiveness. The lesson here is that when you develop community-based groups where the smallest has as much say as the strongest member, where there is a spirit of collaboration and true exchange of ideas, you can build community standards, protocols and best practices that not only work theoretically, but practically and that can be sustained. A further lesson is that you can learn as much from failures as successes.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

U.S. Federal (and perhaps some state) agencies need to be at this table, but no more so than as equals to the nongovernment user communities. If the agencies are not producing what users need - why are they doing it? There must be a balance in power, votes and representation. There have been and will continue to be opportunities for such international collaboration. When they arise the relevant federal agencies should see these as opportunities to participate, and even support the efforts with funding, but be directed to not attempt to control the process or outcomes. In short, typical government agencies have difficulty cooperating within the U.S., much less internationally; user communities much less so. When opportunities arise they should join the efforts and participate and follow the lead of the user communities rather than being the leaders A notable example of a successful international effort is OneGeology (www.onegeology.org), which is a collaboration of mostly the national

geological surveys from over 117 countries. Its focus is geologic maps, and for the scale of the maps they are working with, it has been highly successful. (With one caveat - it effectively does not include, therefore represent the non-government geoscience communities).

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

See the discussion of question #9. What you don't want to do is have agencies take over the role of publication - that simply is too expensive, competes with private industry, and culturally won't work. What you have to do is provide mechanisms that allow for the natural convergence on this issue. In particular accept the fact that the words of explanation and interpretation about data are separate from the data themselves. Require the publication of data, then the linking of data to the publication will happen naturally, championed by the user communities.

Name/Email: OSU Libraries Scholarly Communication Working Group
Authors: Bonnie Avery, Michael Boock, Sue Kunda, Terry Reese, Janet Webster, Andrea Wirth

Affiliation/Organization: Oregon State University Libraries

City, State: Corvallis, OR

Comment 1 We propose that the federal government require data management plans to accompany funding proposals. Data management plans should include a standard way to estimate the cost for specific stages of data management including description, preservation and archiving, and dissemination. The federal government should provide researchers with a standard way of estimating these costs.

We propose a policy that requires grantees to explain how their data will be described and disseminated. Guidelines should be in place for what types of data can be embargoed or excluded from public dissemination.

Comment 2 It is our understanding that data are facts and not subject to copyright. A dataset is automatically in the public domain unless a contract is put on it; however, an original selection, arrangement or manipulation of data is subject to copyright law. In principle, we believe that federally funded data, whether it is subject to copyright and patent law or not, should be in the public domain, or, at the least made available under a Creation Commons Attribution license (CC-BY). Making digital data available in the public domain or under this type of license allows private and public interests to access and reuse data for commercial and non-commercial purposes as efficiently as possible. We recognize that there need to be exceptions to this rule, in cases where the research data is defined as “sensitive” or where research is on-going for example. The federal government should establish guidelines for these exceptions and require researchers and funding agencies to explain why their data should be an exception in their data management plans. Many restrictions on access to research data can be avoided if the data life cycle is well defined.

We also support efforts to make the code required to recreate the data available in the public domain or under an unrestricted license.

Comment 3 First, federal agencies could make a statement about one area where there is no inherent difference in the type of “data”: publications. A definitive statement that all publications resulting from publicly funded research data must be freely available to the public via a persistent link on the Internet is needed. A federal standard could be set for a reasonable lag between publication and availability in an open access repository which researcher could then quote to publishers when assigning their copyrights, rather than allowing the publisher to make this choice against the public interest -- which is now the case.

Second, all agencies should adopt the NSF precedent of requiring “data management

plans" (DMPs) as part of the application process. Differences between disciplines and types of data should be captured in these DMPs. Further if statements about adherence to or evaluation of the DMP become part of the research reporting process, agencies might improve DMP guidelines to enhance data sharing among disciplines.

Finally, federal funders might provide researchers with a standard way of estimating the cost for specific stages of data/preservation/archiving/curation by dataset type/format in their DMPs. (see comment #6)

Comment 5 Libraries at research institutions such as Oregon State University have a long-term interest in supporting their faculty. If not serving as the repository for institutional research results in their published form, academic libraries serve as the pointers to where these results can be found long after the authors cease to be part of the institution. As concern shifts from disseminating information in publications to dissemination of information about the data behind those publications, the library is uniquely placed to provide leadership in longer term research data management where the end goals are greater sharing and better management of research data as an institutional resource. Library professionals can both provide expertise concerning specific aspects of DMPs (how, when and which metadata standards to apply to research data throughout its life cycle; data repository options, linkage between datasets and published research, etc.).

Comment 6 When applying for federal grants, researchers should be required to submit plans similar to the NSF Data Management Plan. Plans should include projected costs for archiving and making digital data accessible, something that the NSF Data Management does not currently require. Awards should then include funding for the curation and distribution of data.

The federal government may want to adopt a central or regional approach for data preservation similar to what Australia is adopting with their National Data Service or the U.K. with their Digital Curation Center. Both recognize that a centralized approach protects against spending money on redundant infrastructures and inconsistent standards. A national model such as DataOne or a regional model of regional data warehouses makes more sense.

Comment 7 Grant recipients should be required to provide a link to their open data at the completion of their research. The government should track conformance with this requirement. Past compliance would be a tool for agencies to use in determining future awardees. Institutions would encourage their faculty to conform in order to ensure future funding.

Comment 8 Putting a national or regional data repository system in place, adopting metadata standards for organization and description of the data, and requiring dissemination are first steps toward broader and more innovative commercial uses of data. Make it easy to find, understand, and use data that is funded by taxpayers.

Comment 9 The federal government may wish to recommend a standard citation format for data produced with public funds. Data that is made more widely available in long-term digital repositories and that is easily cited according to standards supported by the federal government may encourage publishers to allow citation of data.

Comment 10 Digital data standards for specific scientific data will be largely domain specific. Within domains, there are many examples of community-driven data standards. More generally, the library and archival community has developed and participated in a number of standards based efforts to make data sharing easier. Standards like OAI-PMH (Open Archive Initiative) and RDF profiles for the library represent like efforts geared towards the dissemination and findability of data.

Comment 11 Within the library community, a number of successful standards efforts have effectively produced differing data standards. The primarily characteristics of these groups that have made them successful has been their inclusiveness. Almost universally, the standards processes that have been the most successful have utilized a transparent process that has allowed lots of feedback from the communities. But more importantly, those that work on the standards bodies tend to make up many different stakeholder groups allowing for a larger set of concerns to be addressed.

Comment 12 For starters, Federal agencies could seek to participate in the international community, rather than simply designing their own digital data standards. Within the international community, many countries have been wrestling with digital scientific data for a number of years. What's more, many governments are moving to open their own data to the general public (for example, the UK's recent decision to open public health information). Federal agencies need to resist the need to develop a uniquely U.S. solution and look at the work currently be done within the international community and become an active participant.

Sue Parchick
Fri 1/6/2012 9:05 PM
It's wrong for a Publically Funded study to be controlled for any reason. It's dishonest!!!!

It's wrong for a Publically Funded study to be controlled by any

publishing or private company for any reason, for any fee or profit. It's dishonest!!!

Anyone who voted for this shameful bill should be exposed, fined, and booted OUT of congress.

People whose purpose is self profit with public funds is an "embarrassment " to honest hardworking people.

Don Kirksey
Sat 1/7/2012 10:13 AM
Public Access to Digital Data

When making any decisions regarding the use and access to the scientific information generated through government-funded projects (most university research), please focus on protecting creation and preservation of intellectual property (IP). The public funding of university-based scientific inquiry is predicated on the expectation that "new discoveries will be translated into new products". This model (discovery/commercialization) only works when the IP is solid. The citizens are interested in "access to knowledge" rules and regulations only if they will improve the path to new products (diagnostics, medicines, and treatments). Please focus on IP!
Sent from my iPad

Elsevier submission to Office of Science and Technology Policy public consultation on Public Access to Digital Data Resulting from Federally Funded Scientific Research

January, 2012

Introduction

Elsevier's primary mission is to advance science by providing high-quality academic publications and services. We envision a future world in which data are much more broadly managed, preserved, and reused for the advancement of science. We want to work in partnership with other stakeholders to achieve this vision.

Professional curation and preservation of data is, like professional publishing, neither easy nor inexpensive. The grand challenge is to develop approaches that maximize access to data in ways that are sustained over time, ensure the quality of the scientific record, and stimulate innovation.

- We believe rich interconnections between publications and scientific data are important to support our customers to advance science and health.
- We recognize that scientists invest substantially in creating and interpreting data, and their intellectual and financial contributions need to be recognized and valued.
- Funders too invest substantially in these data and their contributions need to be recognized and valued.
- Where publishers add value and/or incur significant cost then our contributions also need to be recognized and valued.
- There are potential new roles, and we want to embrace an active test and learn approach.
- We will be sensitive to different practices and preferences between subject areas as we test and learn.
- Any role for Elsevier would not be exclusive, and we want to work in collaboration with other stakeholders to establish a sustainable framework for the discovery and use of scientific data.

Access to data

Preservation, discoverability, and access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

One of the biggest challenges for scientists today is the invisibility of data. While 97% of researchers in the US report good access to research articles, only 42% of researchers in the US express satisfaction with their access levels to data[1] and datasets, and this access gap clearly hinders scientific progress. It also leads to duplication of funding as different research funders pay repeatedly to have the same experiments run and the same data captured. Leadership and coordinated action by all stakeholders is needed to improve access to data by scientists.

Elsevier recommends the following:

- Federal agencies should work with stakeholders, including research institutions, funding bodies and publishers to develop and deploy standard approaches for linking publications and data.

[1] *Access vs. Importance, A global study assessing the importance of and ease of access to professional and academic information. Phase I Results. 2010. Publishing Research Consortium* http://www.publishingresearch.net/documents/PRCAccessvsImportanceGlobalNov2010_000.pdf

- Federal agencies should encourage authors to document their data and to deposit their data with an appropriate data center or service and to make their data available for reuse by others.
- All data should be assigned persistent, unique Digital Object Identifiers (DOIs) to aid their discovery, use, and citation. DOIs permanently identify and track scholarly items on the web, and are already used to link millions of items from hundreds of publishers and societies. DOIs integrate with the OpenURL and are completely access-model neutral.
- Appropriate metadata should be generated with the data to enable understanding and reuse.
- Stakeholders, including publishers, should encourage academics to cite datasets that have been used in their research and that are available for reuse via a data curation center or service and enable linking of data to the published journal article.
- Federal agencies should work with other stakeholders on policies for long term preservation of data, and accreditation systems/standards for digital curation services.

Federal agencies should also adopt policies that encourage publishers to continue to invest in their journals and in the development of discovery tools for data. For example our article linking tools facilitate entity text-mining (e.g. Arabidopsis Viewer), pull data associated with a published research article from a data store (e.g. Genome Viewer), support visualizations of data from a data store (e.g. Protein Viewer), link from published research articles to further detail in a data store (e.g. all 3 of the previous examples), and link articles to associated data in data stores (e.g. Pangaea or DRYAD). We are able to make these sorts of investments to make data more easily discoverable and reusable because we have sustainable business models for our journals. Unsustainable public access policies for journals could undermine these efforts.

Members of the public may also wish to access scientific datasets collected/created during federally-funded research projects. We recommend that careful work is done to understand actual needs, so that effective and sustainable approaches to filling these needs are developed.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Copyright does not protect ideas or facts, and so does not apply to raw datasets. Respect for copyright should be reflected in any policies, but should not become a smokescreen to prevent the sharing of raw datasets.

Creative investment in the presentation of data can attract copyright, for example in published journal articles. The entirety of a publication is covered by copyright, including any data presented within it, and so permission is required to use their contents unless the intended use is covered by a copyright exception.

Elsevier requests a non-exclusive license from its authors to use supplemental data if they are to be published. This policy means that supplemental data will continue to be owned / controlled by the original researcher. The researcher may, of course, have contractual obligations to his/her employer or funder that will guide whether and how these data can be reused.

A variety of data licenses have emerged which academics may wish to consider for their supplemental data, processed data, or raw data. Elsevier believes these re-use terms should be the choice of the researcher.

Incentives, rather than mandates, are needed to overcome data access challenges. There is currently a disparity between a researcher's willingness to use shared data and to supply it. Many researchers agree it is necessary, but decline to share their own data. An example is documented in a survey prepared for the launch of the EconomistsOnline repository in 2010 that indicated a majority of economists were in favor of accessing datasets, but when asked if they would post their own data only 15% indicated that they would be prepared to do so[2].

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals. We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD).

The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation. DOIs permanently identify and track scholarly items on the web, and are already used to link millions of items from hundreds of publishers and societies. DOIs integrate with the OpenURL and are completely access-model neutral.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

There is great variation across disciplines with respect to how data are collected or created, analyzed, documented, stored, and shared. For example, we see variance in the supplementary datasets submitted to underpin published journals articles (e.g. methods/protocols, raw datasets, executable code, videos of experiments, and more) across disciplines.

Terminology, taxonomy, methodology and communication vary significantly across disciplines (and often sub-disciplines). Controlled, shared vocabulary is essential for facilitating more automated approaches to processing information to underpin scientific discoveries. It is therefore very important for federal agencies to work with researchers, institutions, publishers and other stakeholders to develop (if necessary) and deploy appropriate subject area classifications and controlled vocabularies.

New professional data curation professionals are emerging with skills sets that combine academic expertise in a subject (or sub-discipline) with information science skills. Support for enabling more professional data curators and data curation facilities is essential. Initiatives such as the Archaeology Data Service, CASPAR, the Digital Curation Centre, Planets, OAIS, SHAMAN, and nestor are likely to prove useful examples of effective digital stewardship[3] including the development/deployment of shared vocabularies.

(4) How could agency policies consider differences in the relative costs and benefits of long term stewardship and dissemination of different types of data resulting from federally funded research?

[2] EconomistsOnline – http://www.economistsonline.org/home and http://itswww.uvt.nl/its/voorlichting/PDF/NEEO/D1.7-NEEO-Final-Report-2010.pdf section 7.3.5

[3] PARSE.insight Science Data Infrastructure Roadmap (see http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf)

US agencies can play a very important leadership role by convening stakeholders to work constructively together to identify and overcome barriers to the consistent curation and reuse of important scientific data.

OSTP itself could helpful ask the General Accounting Office to undertake a study of existing federal data archives and data curation centers to determine the full costs required for start-up, management, and ongoing access, preservation, and migration activity across different subject areas.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practices for sustainable and high-quality data curation services.
- Incentivize stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Incentives, rather than mandates, are needed to overcome data access challenges.

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals. We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD). The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation.

We encourage our authors to:

- deposit their data with appropriate data centers or services at the earliest possible opportunity, and certainly by the time of publication, recognizing that academics may need an exclusive period of time to analyze their data and publish results based on these analyses.
- seek support from an experienced data curator with expertise in their subject area (e.g. for privacy issues associated with patient images used in medical research);
- register Digital Object Identifiers (DOIs) for their datasets, implement data management plans, and use open standards to facilitate interoperability and successful data curation and
- cite datasets via their DOIs to encourage the fullest possible understanding of the research objectives, design, and methods prior to access/reuse of the data that underpin the publication

In cases where authors submit data alongside their articles, publishers have developed mechanisms to facilitate their upload and to make these data available for peer review. We have taken steps to ensure that reviewers can see all the material they need to complete their review, including data and supplementary materials.

Elsevier and other publishers can also:

- Champion the importance of long term preservation of data, and accreditation systems/standards for digital curation services.
- Communicate the benefits of data curation and reuse for different stakeholders in the scholarly communication landscape including authors, funders, publishers, researchers, and university administrators.
- Deploy our expertise in certification, indexing, and linking to add value to data (e.g. for search and mining).
- Use standard vocabularies, taxonomies, ontologies, and entity resources where possible rather than inventing our own.
- Support the creation and capture of linked data during the authoring and editorial process and maintain linked data through production processes.
- Facilitate the rich linking to and from publications.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

We agree this is an important issue and evidence-based policy making is crucial. It is essential for there to be clear business cases for, and sustainable business models, to underpin data curation. There are a number of active digital data centers and data repositories that commit to the documentation and digital preservation of their contents. More systematic study of the costs associated with these, and more rigorous attention to their long term sustainability would be extremely helpful to all stakeholders.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practice for sustainable and high-quality data curation services.
- Incentivize stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Some work is already underway in this area through groups such as the ISO Repository Audit and

Certification Working Group. We believe federal agencies should work with other stakeholders to develop (if necessary) and deploy standards and policies in this area.

Publishers can play a role to incentivize the deposit and reuse of data - just as we help to incentivize academics to publish by enabling them to register their scientific discoveries in widely accessed, cited, and respected journals. We have made data available alongside publications and support initiatives to help researchers to share data (e.g. Pangaea, CCDC and DRYAD). The publishing industry has also developed standards for inter-linking datasets and publications through the International DOI Foundation.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

OSTP itself could helpful ask the General Accounting Office to undertake a study of existing federal data archives and data curation centers to determine the full costs required for start-up, management, and ongoing access, preservation, and migration activity.

Federal agencies should provide funding to:

- Support researchers to document and deposit their datasets in data curation centers.
- Support desk-based research that involves reuse of datasets deposited by other researchers and accessible via data curation centers.
- Establish discipline-specific data curation facilities where these do not yet exist. Both the Open Archive Information System (OAIS) Reference Model and the report of a Blue Ribbon Task Force on Sustainable Digital Preservation and Access (available at http://brtf.sdsc.edu/) are both helpful in identifying best practice for sustainable and high-quality data curation services.

Federal agencies should also:

- Encourage the re-use of publicly funded and accessible research datasets by making them available under unambiguous non-exclusive licenses
- Commit to develop a culture of ethical re-use of data, for example by banning those who willfully misrepresent or distort data created by others from receiving grant funds
- Incentivize (e.g. by respecting copyright and other intellectual property rights) stakeholders who develop/deploy technical standards to facilitate the transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text, and tools. Create a level-playing field for different sustainable business models that emerge for these products and services

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Authors should be encouraged to registered DOIs for their data. With a DOI in place datasets become citable and can be linked to other data and to publications. The DOI contains the standard metadata which will supply the required metadata, including author, affiliation, related articles, etc. In addition, as the DOI has successfully worked with millions of published journal articles, the use of the DOI facilitates easy linking across objects, including data set and published articles.

Standards for interoperability, re-use and re-purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Digital curation standards are mostly in formative stages, and the following are good examples of active projects in this area:

- Opportunities for Data Exchange (www.ode-project.eu)
- DataCite (http://datacite.org/)
- APARSEN (http://www.alliancepermanentaccess.org/index.php/currentprojects/aparsen/)

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Recently NISO-NFAIS released Recommended Practices for Supplemental Journal Article Materials. Elsevier has been an integral partner in the development of these guidelines.

More broadly CrossRef and the International Digital Object Identifier Foundation have been transformational in developing standards to link publications and data in a persistent way. The history of these successful standards organizations is perhaps helpful to relate.

The International DOI Foundation was created in 1998 to support electronic publishing through the development and promotion of the DOI (Digital Object Identifier) System as a common infrastructure for content management. The Foundation is a registered not-for-profit organization, controlled by an Executive Board elected by the members of the Foundation.

In 1999 a technical demonstration was made of the DOI at the Frankfurt Book Fair. Representatives of the leading scientific, technical, and medical publishers recognized in this prototype that a lookup system based on the Digital Object Identifier (DOI) held the key to a broad-based and efficient journal reference linking system. They took the unusual step of joining together as the non-profit, independent Publishers International Linking Association Inc. (PILA), which was incorporated in January 2000 and CrossRef went live as the first collaborative reference linking service in June 2000.

CrossRef's mission is "to be a trusted collaborative organization with broad community connections; authoritative and innovative in support of a persistent, sustainable infrastructure for scholarly communication." CrossRef's general purpose is to promote the development and cooperative use of new and innovative technologies to speed and facilitate scholarly research. CrossRef's specific mandate is to be the citation linking backbone for all scholarly information in electronic form. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages through CrossRef Digital Object Identifiers (CrossRef DOI), which are tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article.

In parallel the International DOI Foundation has continued to evolve. Millions of DOIs have been assigned, and millions are accessed each month. DOI registration agencies have been appointed across

the globe. CrossRef was the first registration agency and has been followed by Office des publications EU (OPOCE) for government documents, *m*EDRA for multilingual resources, EIDR for movie and television assets, and DataCite for scientific data.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies could helpfully endorse existing initiatives and standards such as those mentioned in our response. Inter-working with the international academic publishing community on data issues is best done through the International Association of Scientific, Technical, and Medical Publishers (STM) and through the involvement of Elsevier and other publishing companies in working groups.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

As stated the publishing industry has developed standards for inter-linking datasets and publications through two not-for-profits: the International DOI Foundation and CrossRef. Elsevier is working with a number of organizations to link to appropriate data. For example, in earth sciences, we provide interlinking to earth sciences data through Pangaea and our published journal articles. Support by federal agencies for the Digital Object Identifier as the persistent identifier of choice for both data and publications would significantly improve the linking of publications and associated data.

Collaboration with DataCite could also be extremely helpful. This is an organization that aims to increase acceptance of research data as legitimate citable contributions to the scholarly record, and support data archiving to permit results to be verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, they are currently working primarily with organizations that host data, such as data centers and libraries.

Youngsuk Chi
Chairman
Elsevier
360 Park Avenue South
New York, NY 10010

To: **Office of Science and Technology Policy (digitaldata@ostp.gov)**
From: **University of California Libraries**
Subject: **Response to the OSTP RFI on Public Access to Digital Data**

Dec. 23, 2011

Introduction

The University of California applauds the recent Request for Information issued by the Office of Science and Technology Policy (OSTP) [1] to solicit comment and recommendations on approaches for ensuring long-term stewardship of, and broad public access to, digital data resulting from federally funded research. These research data play a fundamentally important role in the ongoing practice of scientific inquiry, discourse, and advancement, with many broader societal dependencies, relationships, and consequences, both direct and indirect, in commerce, education, and culture. However, data represented in digital form is inherently fragile with respect to the ever increasing pace of disruptive technological and institutional change. Thus, research data must be placed under careful, comprehensive, and proactive management and stewardship – in short, it must be properly *curated* – in order to ensure that it remains available for use, sharing, and re-purposing by current and future generations of scientists and scholars. Twenty-first century research is driven by the continual, exponential advances of computation, storage and bandwidth. It is imperative that data produced in this environment neither be lost nor remain unused. Unused data has no value.

The University of California believes that truly effective and sustainable stewardship solutions depend upon adherence to five broad strategic imperatives:

- *Know what you have* ("You can't manage what you don't measure")

  The objects of stewardship must be clearly documented in terms of significant form or structure, scientific meaning, and desired behavior in order to facilitate successful preservation, discovery, and use.

- *Express and share that knowledge widely* ("It takes a village")

  Comprehensive description of research data must itself be the object of affirmative custodial care. This information should be formally documented as an aid to widespread discoverability and potential transfer of stewardship responsibility. Note that the proper emphasis of all stewardship activity is on the persistence of the research data (and its concomitant description) itself, not the systems in which that data is managed, which are inherently ephemeral.

- *Make lots of copies* ("Be redundant, be redundant")

  Both coarse and fine grained replication and redundancy in all aspects of the stewardship infrastructure – technical, curatorial, and procedural – are one of the most powerful means to minimize the potential for debilitating single points of, systemic, or correlated failures.

- *Protect the copies* ("Trust, but verify")

An effective stewardship infrastructure incorporates affirmative safeguards into all technical systems and workflows, such as storage-level use of error correcting codes, media refresh, and fixity audit; data center high availability hosting and operational practices; and established procedures for obsolescence detection and mitigation.

- *Plan and watch* ("Proactive when you can, reactive when you must")

  Successful stewardship outcomes require a comprehensive program of services, policies, and practices. It is important to be prepared for anticipated eventualities through action plans, ongoing technology watch, and stakeholder engagement activities, while also being alert for, and responsive to, unexpected conditions requiring attention.

Preservation, Discoverability, and Access

The OSTP RFI properly casts preservation and access as complementary, rather than disparate, stewardship activities: access being dependent upon preservation up to a *point* in time, while preservation ensures access *over* time. Furthermore, data access enables new forms of scholarly collaboration and communication that will lead to revolutionary means of exploring knowledge. This can only be achieved with effective preservation and discovery. The full range of necessary stewardship policies and practices encompasses both technical and organizational facets. While effective technical systems are a necessary foundation, truly effective and sustainable stewardship solutions for long-term preservation and access must rely significantly on human competencies, analysis, and decision making.

(1) Encouragement of Access and Preservation

One of the most direct positive impacts that federal granting agencies can have in encouraging desirable behaviors for the long-term management, preservation, and sharing of research data is through instituting appropriate requirements as a pre-condition, and an auditable post-condition, for funding. (Current NIH and NSF data management requirements serve as exemplars.) Although compliance to new behaviors for preservation and sharing will initially be driven by external requirements, over time, as they are increasingly integrated and internalized into researchers' work practices, they will eventually come to be accepted merely as normative patterns of scientific activity. It would be useful for granting agencies to couch the intent underlying their requirements in terms of the many tangible benefits that may personally or organizationally accrue from following these practices, such as increased opportunities for collaboration, directed or serendipitous discovery, publication, and citation [2].

(2) Protection of Intellectual Property Interests

US law does not provide copyright or intellectual property protection to facts or factual data. US Copyright Office Circular Number 1 states, "[c]opyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these things are expressed." [1]

[1] See [3]. US copyright law contains additional exemptions and restrictions, and "several categories of material are generally *not* eligible for federal copyright protection": for example, "works that have not been fixed in a tangible form," "titles, names, short phrases and slogans," work created by the federal government, "ideas, procedures, methods, systems processes, concepts, principles," "works consisting

There are three questions that must be asked to determine whether an intended use of data triggers copyright or any intellectual property protection. The first question in respect to data and copyright and intellectual property protection is whether the intended use triggers copyright. Does the intended use of data fall within one of the copyright owner's exclusive rights? Title 17 USC Section 106 defines the exclusive rights of copyright owners [4], and if the intended use does not fall within those exclusive rights, then the use is permitted unless there is a prior contract that prohibits it. The mere extraction or copying of "ideas, facts, processes, or methods" or federal government information that is excluded from copyright protection does not trigger copyright (17 USC 102(b), 17 USC 105, and 17 USC 106). The reproduction and transmission rights of copyright owners cover only the making and or distribution of "copies," and not every "copy" qualifies. Specifically, "copies" that are not capable of being "perceived, reproduced, or communicated" and that are not sufficiently "fixed" do not qualify. The US Copyright Office defines the word "copy" for purposes of US copyright law as follows: a copy is "[t]he material object, other than a phonorecord, in which the copyrighted work is first fixed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device" [5].

The second question regarding data and copyright and intellectual property protection is, even if the intended use does trigger copyright, does it fall within copyright's limitations or exceptions such as Fair Use (17 USC 107), the Exemption for Libraries and Archives (17 USC 108), or other exemptions? If the intended use falls within an established exemption, then it does not require any further copyright or intellectual property protection.

The third question is, even if copyright does cover the intended use and the intended use is not covered by an exemption, is the work available under license such as Creative Commons or other public licenses? If so, then the use may fall under the scope of the license. Again, this use of data is pursuant to the scope of a license or contract, so no further copyright protection is necessary or warranted.

One growing use of digitized data is that of non-consumptive research. Non-consumptive research relies on computation and, typically, computational queries to a database or collection of digitized data in which the content and corpus of the database is not accessed for display or reading. This new form of research and discovery is a user's right, not a copyright owner's right, unless the user agrees to a contract or license that restricts computational and non-consumptive research or unless the database owner employs technology to prevent, prohibit, or restrict a user's right to engage in non-copyrighted computational research and discovery.

Clifford Lynch, director of the Coalition for Networked Information, notes that opportunities for new knowledge creation, production, and innovation require "new ways to think about the scholarly literature (and the underlying evidence that supports scholarship) as an active, computationally enabled representation of knowledge that lives, grows" and further "suggests ways in which information technology can accelerate the rate of scientific discovery and growth of scholarship" [6]. Given the enormous potential for public access and use of preserved and accessible publicly funded data, it would be a serious educational, scientific, and economic

---

entirely of information that is common property and containing no original authorship (for example: standard calendars, weight charts, tape measures, and rulers, and lists or tables taken from public documents or other common sources)" [3].

setback if opportunities for the creation of new knowledge and discoveries were limited due to incorrect interpretations and applications of existing copyright law and overly broad expansion of intellectual property protection. Current copyright, intellectual property, and licensing regimes provide more than adequate protection for creators, scientists, federal agencies, and publishers with respect to digital data.

(3) Disciplinary Differences

The broad strategic imperatives underlying long-term stewardship certainly apply, at least in the abstract, regardless of scientific discipline. Even at the more tactical level, many specific activities can be performed without particular regard to scientific discipline, for example, persistent identification, citation, storage, replication, fixity, and technical characterization. Most meaningful disciplinary differences come into play in terms of formats and tools, descriptive practices, methods and methodology, and modes of discovery and use, which can be highly specific to both broad and niche communities of practice. Funding agencies should recommend the widest possible use of the most common data formats and analytical tools. Comprehensive information documenting availability, deployment, and use of all such formats and tools, whether common or not, should be publicly available in well-known technical registries, such as PRONOM [7] or the Unified Digital Format Registry (UDFR) [8], being developed by University of California Curation Center (UC3) at the California Digital Library (CDL) as part of the Library of Congress's National Digital Information Infrastructure Preservation Program (NDIIPP).

Many of the discipline-specific descriptive practices have come into being as the result of long collaborative experience and represent an optimization of effort and productivity. As such, the development and codification of such practices should be accepted by federal agency policies, albeit with encouragement of their establishment with the widest possible scope. The primary challenge raised by narrow descriptive standardization comes from the rise of cross-disciplinary research, particularly when diverse communities rely on substantially inconsistent practices. Linked data and other semantic web technologies hold out great promise for facilitating automated cross-disciplinary discovery, and federal granting agencies should encourage the development and use of appropriate ontologies for this purpose.

Another potential disciplinary distinction is the "big-data/small-data" divide. Extremely large datasets, especially those arising in the hard sciences through large-scale simulation or fine-grained observational instruments, pose significant administrative and technical challenges. Funding agencies should continue their efforts to support, publicize, and move onto a sustainable footing high performance computing facilities. However, a recent survey of over 1700 *Science* peer reviewers reported that the largest dataset generated or used locally by over 48% of the respondents was less than 1 GB in size [9]. Thus, while the problems of big data receive much public scrutiny, it is important that small data, particularly in the life and social sciences, whose usage is undoubtedly much more diffuse, continues to receive adequate attention and support. It should also be noted that the Humanities is rapidly changing as well. Modern scholars working on interpreting the cultural fabric of our world are doing so in a largely data driven environment, so much so that recent trends in humanities' scholarship are taking the shape of data curation as a publication [10]. It is imperative for all forms of scholarship that data be prepared and shared with a networked mindset.

(4) Costs and Benefits

The allocation of scarce curatorial resources, whether financial or otherwise, is always based on evaluations of the current and future value proposition for the curated data. Evaluation criteria should include the scientific value, scope of applicability, and degree of uniqueness and reproducibility. Since any assessment of future value can be problematic with respect to fundamental underlying assumptions and ever-changing conditions, it is important that all plausibly useful research outputs are subject to minimally sufficient baseline practices, and that there is some level of ongoing curatorial assessment to select data deserving of added value attention in light of evolving circumstances.

(5) Stakeholder Contribution

Many well-established memory institutions – libraries, archives, and museums – have developed deep expertise, experience, and resources for dealing with the long-term preservation of and access to cultural heritage material, which in many cases can be directly applied to the stewardship of research data. These institutions should be encouraged to make available their preservation and curation systems and services to the research community. Many research universities, such as the University of California, already have mature local, centralized, and consortial solutions in place to address the long term stewardship needs of the University's digital assets.

As an example, an international group of academic libraries, research projects, and government and non-profit organizations have collaborated under the leadership of UC3, UCLA, and UCSD to create the DMPTool [11], a publicly available online system that aids researchers in creating data management plans meeting funder requirements. This system is configured to provide campus-specific guidance and advice regarding the availability of services appropriate for long-term stewardship.

(6) Funding Mechanisms

Based on UC3's experience working with data owners internal and external to the University of California, a "pay as you go" model for stewardship services may not be appropriate in all contexts. The majority of research data derives from grant funded project activities, which leaves little provision for sustainable funding beyond project completion. Most researchers are therefore eager to embrace an alternative "pay once" model whose charges can be built into grant proposals. In order to be sustainably viable, however, it is important that stewardship service providers understand the full gamut of lifecycle costs [12][13]. Note that this may require service providers to take on an unfamiliar fiduciary role in the long-term management of endowment funds. These funds should be dedicated for the purpose of sustaining the research data in perpetuity, and not be available for reallocation towards other service provider priorities.

(7) Policy Compliance

Voluntary compliance to Federal stewardship policies can be enhanced by proactively bringing all of the affected stakeholders – researchers, service providers, funders – into the process of developing those policies, so that all parties can feel that their particular needs and concerns have been considered and incorporated. Reporting requirements should be kept to a minimum and based on commonly accepted objective measures. If possible, these measures should be

independently verifiable, as suggested by the "neighborhood watch" concept proposed by the UC Curation Center [14], so that compliance can be determined with minimal intrusion and cost. Since funder requirements have only recently started to be promulgated, many in the research community are not fully aware of their intention or significance. Affirmative efforts are needed, both by funding agencies and local institutions, to raise the awareness of all of the implications of the new policy requirements.

(8) Innovative Use

The use and re-use of research data is largely dependent upon four factors: knowing that the data exist; knowing where to get it; having it delivered in a form that is easily integrated, either directly or through minimal transformation, into local work practices; and ensuring the long-term public access and use of publicly funded data. The first and second are questions of widespread dissemination of descriptive metadata in appropriate intra- or cross-domain discovery services integrated with datacenters and access repositories. The third factor is more difficult as it is facilitated by growing conformance to common data practices regarding the acquisition and representation of data. While some scientific communities have broad internal agreement regarding these practices, in many cases idiosyncratic usage is the norm. Funding bodies should consider supporting efforts at standardizing and codifying disciplinary practice on the broadest terms as part of a more general encouragement of translational research and centers of excellence. Finally, funding bodies and government agencies should require long-term public access to and use of publicly funded data.

(9) Attribution

Providing scientists with assurance of appropriate attribution and credit for making available their research output can be facilitated through support for formal data publication. While the historical practice has been to provide public visibility to only one of the many outputs of a research program – the summarizing paper or conference presentation – there is no reason why the other data products could not be similarly treated, wrapping those products in the familiar façade of academic publication [15]. Providing datasets with persistent identifiers and descriptive citations enables the entire scholarly publication infrastructure to come into play to provide sophisticated aggregation, indexing and abstracting, enhanced discovery, and attribution, all of which should combine to encourage more widespread use and repurposing.

Standards for Interoperability, Re-Use, and Re-Purposing

Effective solutions for the long-term preservation of, and access to, digital research data will almost certainly involve a community of committed stakeholders. Increased access actively promotes preservation outcomes: data that are used widely or frequently are much more likely to receive the appropriate stewardship attention. The global distribution of stewardship expertise and experience will always be uneven and it is natural to assume the orderly or ad hoc development of specialized centers of excellence offering tools, best practice recommendations, and services to the broader community. Furthermore, preservation is a serial, rather than a one time, activity. Given the inevitable evolution of organizational mission, resources, and priorities, over any sufficiently extended period of time it is likely that the responsibility for the physical and curatorial custody of research data will be transferred from institution to institution. Thus, broad

community conformance to accepted standards – both de facto and de jure – is a necessary concomitant to sustainable long-term success in preservation and sharing. We are in a research environment where technology is of secondary importance; information – and the widest possible distribution and sharing of that information – is what matters.

(10) Interoperability Standards

As discussed in the context of question (8), the use of research data is predicated on three factors: knowing that the data exist, knowing from where the data are available, and having it made available in a form that is easily integrated into local workflows. These suggest the need for common standards for data description, publication, discovery, and representation formats. Data description must be supported at sufficiently fine grain to enable direct and, ideally, automated determinations of the suitability of a given dataset for a particular local purpose. In other words, descriptive practice should extend down to the level of individual variable fields, units of measure, spatial and temporal coverage, normalization procedures, etc. It is also imperative that research data move from inside the academy to the outside. This suggests that descriptive standards should be developed in a manner that is usable to those with deep disciplinary expertise as well as broad synoptic understanding.

(11) Standards Process

The ecological science community has been successful in fostering a number of open source informatics standards and projects, including the EML metadata standard and its attendant tools. The ecoinformatics.org organization [17] provides a central platform and lightweight process for harnessing the voluntary collaborations of domain scientists in areas of broad concern and applicability. In accordance with open source principles, this work is self-directed and self-governing, leading towards community empowerment and commitment.

(12) Standards Coordination

Coordination and standardization of scientific data practices can be best performed on a disciplinary, rather than governmental, basis, taking advantage of long-standing disciplinary channels for intra- and cross-domain discourse and collaboration. Conformance to standards and best practices will be greatest when those practices are perceived as arising from within the community of concern and practice, rather than being imposed externally. That being said, governmental agencies and funding bodies can play an important role in encouraging and funding disciplinary working groups constituted on the broadest possible basis. Today, and certainly in the future, many innovative avenues of scientific advance result from research that transcends traditional disciplinary boundaries. It is therefore important that cross-disciplinary efforts at common standardization or standardized cross-walks be established and encouraged.

(13) Data and Publication

The DataCite consortium develops and promotes DOI-based standards and services for data publication, and is working with the scholarly publishing community to provide greater visibility to research data in the familiar context of discovery portals and indexing and abstracting services [18]. Federal funding decisions should be planned to encourage the support by publishers and data repositories of bi-direction linking between traditional academic publications and the data that underlies their analysis and conclusions, with all of the attendant mechanisms and incentives

for citation, attribution, and impact analysis.

References

[1] Office of Science and Technology Policy (2011), "Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research," 71 *Federal Register* 214 (4 November 2011), pp. 68517-68518.

[2] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma (2007), "Sharing detailed research data is associated with increased citation rate," *Public Library of Science* 2:3 <http://dx.doi.org/10.1371/journal.pone.0000308 >.

[3] US Copyright Office, Circular 1, *Copyright Basics* <http://www.copyright.gov/circs/circ01.pdf>.

[4] 17 USC Section 106 <http://www.copyright.gov/title17/>.

[5] US Copyright Office, *Definitions* <http://www.copyright.gov/help/faq/definitions.html>.

[6] Clifford A. Lynch, "Open computation: Beyond human-reader-centric views of scholarly literatures," *Open Access: Key Strategic, Technical and Economic Aspects*, ed. Neil Jacobs (Oxford: Chandos Publishing, 2006), pp. 185-193.

[7] National Archives [UK] (2001), *PRONOM* <http://www.nationalarchives.gov.uk/PRONOM>.

[8] UC Curation Center (2011), *Unified Digital Format Registry (UDFR)* <https://bitbucket.org/udfr/main/wiki/Home>.

[9] "Challenges and opportunities" (2011), *Science* 331:6018 (11 February 2011): 692-693 <http://www.sciencemag.org/content/331/6018/692.full.pdf>.

[10] UCLA Institute for Pure & Applied Mathematics, *Networks and Network Analysis for the Humanities: An NEH Institute for Advanced Topics in Digital Humanities*, August 15-27, 2010 <https://www.ipam.ucla.edu/programs/hum2010/>.

[11] Andrew Sallans (2011), "DMPTool: Supporting the data lifecycle," *NSF Workshop on Research Data Lifecycle Management*, Princeton University, July 18-20, 2011 <http://www.columbia.edu/~rb2568/rdlm/Sallans_UV_RDLM2011.pdf>.

[12] Serge J. Goldstein and Mark Ratliff (2010), *DataSpace:: A Funding and Operational Model for Long-Term Preservation and Sharing of Resource Data* <http://dspace.princeton.edu/jspui/handle/88435/dsp01w6634361k>.

[13] University College London/British Library (2011), LIFE: Life Cycle Information for E-Literature <http://www.life.ac.uk/>.

[14] Stephen Abrams, Patricia Cruse, John Kunze, David Minor, and Mike Smorul (2011), "'Neighborhood watch' for repository quality assurance," *Designing Storage Architectures for Preservation Collections*, Library of Congress, September 26-27, 2011 <http://www.digitalpreservation.gov/news/events/other_meetings/storage11/docs/cdl_neighborhood_watch_paper.pdf>.

[15] John Kunze, Rachel Hu, Patricia Cruse, Catherine Mitchell, Stephen Abrams, Kirk Hastings, and Lisa Schiff (2011), "Baby steps to data publication," *Beyond the PDF*, University fo California, San Diego, January 19-21, 2011

<http://sites.google.com/site/beyondthepdf/workshop-papers/baby-steps-to-data-publication>.

[16] John Kunze, Rachel Hu, Patricia Cruse, Catherine Mitchell, Stephen Abrams, Kirk Hastings, and Lisa Schiff (2010), *Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines*, Report for the Gordon and Betty Moore Foundation, <http://escholarship.org/uc/item/9jw4964t>.

[17] Ecoinformatics (2011), *Ecoinformatics Online Resource for Managing Ecological Data and Information* <http://www.ecoinformatics.org/index.html>.

[18] DataCite (2011), *DataCite: Helping You Find, Access and Re-use Research Data* <http://datacite.org/>.

January 9, 2012

U.S. Office of Science and Technology Policy
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research
Docket number OSTP–2011–0022
digitaldata@ostp.gov

Office of Science and Technology Policy,

The following comments are in response to the December 23, 2011 Federal Register notification (Vol. 76, No. 247, p. 80417-80418) inviting public comment on the "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research".

Preservation, Discoverability, and Access

Question 1 comment
An important aspect is whether the digital data is from a specific research project has value or if there is only value from the massive aggregation of such digital data. In the case of the digital data from a specific research project then policies can be established to encourage authors to make use of a journal's capability to make available supplementary data available for public access. In the case where only massive aggregation of such digital data has value then there need to be federal policies that encourage such massive aggregation as well as providing the resources to host such massive data sets.

Question 3 comment
Work with representative professional scientific organizations to find the most appropriate venue and manner to manage generated data. For those professional scientific organizations that also manage publications (e.g. American Chemical Society) they may be able to provide the capability to manage supporting information associated with publications.

Question 4 comment
If provided an efficient manner of making data available to their peers I believe most in the scientific community would see the value of doing so where it makes sense. I believe in the long term the scientific community will appropriately recognize those who make the effort to appropriately share their scientific data.

Question 5 comment
I believe that professional scientific organizations can best contribute through the encouragement of their members to advance the scientific enterprise by wider sharing of useful data with their fellow colleagues. Where those organizations also manage publications they can provide the capability to host associated supplementary information.

In relation to question 7, those professional scientific organizations that manage publications could provide reporting on the extent to which authors associated with specific institutions make use of the ability to provide associated supplementary information.

Question 6 comment

I believe that federal agencies, individually or collectively, should consider providing needed resources to voluntary consensus standards organizations to develop the needed digital data format standards that would greatly assist the free exchange of digital data. Separately, federal agencies would need to provide resources to provide hosting of some larger collections of digital data. While there are good examples of collections of digital data arranged on informal bases such collections are vulnerable to frequent moves and varying levels of support and maintenance.

For example the Mössbauer Effect Data Center is an example of such an informal collection of digital data. That data center is completing a move to the Dalian Institute of Chemical Physics, Chinese Academy of Science. Formerly it was located at the University of North Carolina at Asheville.

Through the provision of cloud computing resource the hosting and management of such digital data collections should be achievable on a more cost effective basis than is currently possible.

Question 7 comment
Work with professional scientific organizations that manage publications to provide reporting on the extent to which authors associated with specific institutions make use of the ability to provide associated supplementary information.

Question 8 comment
There may need to be specific funding mechanisms for federal agencies such as NSF to catalyze and initiate such endeavors.

Question 9 comment
Work with professional scientific organizations to establish policies regarding such attribution and credit policies.

Standards for Interoperability, Re-Use and Re-Purposing

Question 10 comment
The establishment of data format standards, where currently not existing or insufficient, would be needed to take maximum advantage of re-use and re-purposing of data. I believe that voluntary consensus standards organizations with appropriate assistance of federal agencies and professional scientific organizations are the best organizations to lead and maintain such developments.

Question 11 comment
I believe the development of ANSI/IEEE N42.42, "American National Standard Data Format Standard for Radiation Detectors Used for Homeland Security", is a good example of where there was strong federal support of the standards development process and strong interactions

with commercial vendors. On the down side the focus on U.S. Homeland Security usage and limited participation by the research community has not brought that format into more general usage in the scientific community. However, the ground work has been laid and with the right additional support by federal agency, voluntary consensus standards organization, and professional scientific organizations that format could become more generally used.

For other types of radiation detection instrumentation there is a need for standard data formats to be established before there can be the free exchange of digital data that could catalyze additional scientific advances.

Question 12 comment

Federal agencies should work collectively with voluntary consensus standards organizations (e.g. ASTM International) to determine what digital data format standards are needed and support their development. Such a step would be consistent with the National Technology Transfer and Advancement Act (NTTAA), Public Law 104-113. There could also be the involvement of major professional scientific organizations (e.g. American Chemical Society). Voluntary consensus standard organizations such as ASTM International have good experience in involving participation not only from U.S. interests but also international interests.

Question 13 comment

Many scholarly publications are using DOI (digital object identifier) to uniquely allow access to specific publications. The DOI may also find utility in linking a specific publication to the associated digital data. Other linking information could be achieved through contract / grant identifiers.

For digital data that is hosted on federal agency systems there should be consideration to using a form of DOI to allow for collections of data to be easily referenced in publications. Within individual collections of digital data individual digital data elements would also need a form of DOI to uniquely identify each data element. Data collections and individual data elements should also have the ability to be versioned in those isolated instances where corrections need to be made by the generator.

Sincerely yours

Mr. Donivan Porterfield
Los Alamos, NM 87544

DUKE UNIVERSITY
LIBRARIES

January 10, 2012

White House Office of Science and Technology Policy
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

RESPONSE from the Duke University Libraries.

Preservation, discoverability, and access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Federal policies regarding research data should start from the premise that more open is better, and then allow restrictions only as necessary to protect specific interests or provide specific incentives (and only if justified in the funding proposal). Open data can act a lever to maximize the investment made to create it, by allowing others to analyze it using methods other than those applied by the creator, and by allowing third parties (including entrepreneurial and commercial services) to combine it with data from other sources or layer innovate services on top. The classic example of this is data from the National Weather Service, which, because it is openly available, has provided the basis for an untold number of scientific and commercial projects, and created a whole new market for weather-related services that could not have been supported by the agency collecting the data on its own.

The best implementation approaches for such a policy would be those that take into account incentives that would encourage researchers and their institutions to preserve and share data, peer expectations being among the strongest of these for researchers. The data management plan requirement recently adopted by NSF is helpful in that it sets an expectation but doesn't require a specific implementation method, accounting for the variation in data types and practices across disciplines. Similar policies should be adopted by other federal funding agencies, encouraging broader access and preservation of research data to become an expectation in all disciplines.

However, setting policies will not be enough – it would be helpful for the federal government to stimulate the development of services that would make data sharing and

Box 90193 Duke University Durham, NC 27708-0193 Tel: 919-660-5800 Fax: 919-660-5923 library.duke.edu

preservation easier, and to foster standardization on a small set of generalized platforms and best practices to reduce the costs of managing research data.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Intellectual property issues around data are not well understood within the research community, and often barriers to access are put in place because of fear of potential misuse that could be allayed by the application of clear licenses that address specific concerns, like attribution. As part of their data policies, Federal agencies should recommend or require selecting from a specific set of data licenses that allow the openness that will promote innovation and scientific discovery while addressing legitimate concerns of the creators of the data and the agencies and home institutions that supported them.

Agencies should provide guidance to researchers at the proposal stage of a project on how to understand intellectual property issues related to their data, so that appropriate license selection and data management practices that take these into account can be implemented early on.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The policies and principles should be general, while the practices will need to be specific. Specific practices should emerge from specific communities, but providing common implementation platforms that can be the underpinning for variable practices would help.

Funding agencies should also be willing to provide funding to support data management expertise to be available locally at researchers' institutions (for example at their libraries) or through disciplinary repository services (such as NESCent Dryad) to assist researchers in applying data management approaches appropriate to their discipline. Examples of institutional services are the Distributed Data Curation Center at Purdue or the Scientific Data Consulting Group at the University of Virginia.

(4) How could agency policies consider differences in the relative costs and benefits of long- term stewardship and dissemination of different types of data resulting from federally funded research?

This is difficult to answer without some baseline data on what the relative costs and benefits are, and this baseline data will be difficult to come by unless comparable practices are put in place across different disciplines and outcomes are tracked over time. A good starting point might be to set a baseline allowable cost for data management and preservation (as a percentage of total project cost) for funding requests, and analyze after several rounds what approaches have been applied in different disciplines and how effective they are based on metrics like transaction costs for a third party to discover, retrieve, and make use of the data; verifiable integrity of the data at different year intervals, retrieval and use statistics, and so on.

If funding is provided to disciplinary repository services (such as Dryad, as mentioned above) they could be required to report on their methods and effectiveness, and comparing outcomes across different disciplinary repository services could be used as cost/benefit heuristics.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Researchers themselves should be able to keep their focus on what they know best – the subject and processes of their research. The organizations named in this question should provide support services (local and hands-on, where possible) to the researchers to facilitate best practices, find appropriate disciplinary standards and infrastructure, and implement approaches with a big-picture and long-term view in mind. These services should be made available at the beginning of projects (ideally, at the proposal stage) to facilitate best practices being put in place early and avoiding inefficiencies of retrofitting new practices mid way through or at the end of a project.

Currently, few organizations have the staff or infrastructure needed to provide this support. Federal agencies could provide funding and incentives to help build these support systems and encourage researchers to make use of them. The cost of developing these concentrated institutional support infrastructures will almost certainly be less than the distributed costs and inefficiencies of each researcher trying to figure out how to implement appropriate data management practices on their own, and likely doing it inconsistently or unsuccessfully.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

It will be important to recognize that not all costs for data management, sharing, and preservation will be directly attributable to particular projects, and that as these expectations become more routine a larger proportion of the costs will need to be considered indirect costs. Data management and publishing and preservation services will become the equivalent of library stacks and services today, and will need to have a

basis for persistence beyond the life of any given project. While disciplines or projects with exceptional needs will be able to articulate clearly their specific data management needs and costs, the majority of research projects are not likely to be able to do so, and will need to rely on baseline services provided by their institutions or disciplinary organizations, with more general formulas for funding allocation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

As noted in the response to #4 and other questions above, develop some reporting metrics that can be used for early efforts to improve effectiveness of approaches to data stewardship, provide support for organizations to help researchers meet a baseline set of expectations, and only when these are in broadly in place and researchers have no excuse not to do the right thing, then become stricter regarding compliance and verification.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Agencies should encourage and support use of open data licenses and platforms that make it easier for researchers to share data in standard ways. If organized and described well and provided through documented APIs and open licenses, data may be combined and analyzed using new analytical tools, or used for purposes not envisioned by their original creator.

Agencies could support data hubs, providing a discovery and access service to data even if it is hosted in distributed disciplinary repositories. Such hubs could act not only as registries of available data and how to get it, but could also feature examples of innovative uses of the data, to stimulate others to envision similar or unique uses of available data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data citation standards (such as those being developed by the DataCite project) and researcher identifier standards (such as those being developed by ORCID) are important, to encourage consistency of practice and enable machine-actionable analysis. But widespread use of such standards will depend on community expectations. To encourage data citation norms to be adopted by disciplinary research communities, agencies should require disclosure of data sources (using common data citation and researcher identification standards) in grant proposals and reports, and should encourage authors to prominently display their data sources and data citations in their

publications. We need to reach the point where data citation has become an expectation similar to publication citation. Requiring particular data citation practices in places where requirements are possible will make the practices more visible and more likely to be adopted in places where they are not necessarily required.

Standards for interoperability, re-use and re-purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community- driven data standards effort.

It's difficult to address individual standards because they will differ widely in different fields and require deep knowledge of practices in that field to be able to make reasonable recommendations. However, agencies should provide incentives and assistance to researchers to be aware of, choose, and use existing standards with broad community adoption rather than creating new ones. In proposals and reports, researchers could be required to justify what standard they have chosen, and agencies could encourage peer reviewers to look at these critically. Agencies should make sure to have experts who understand the commonly used standards for particular disciplines on review panels.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Some of the standards that form the basis of the Internet were created and are governed by collaborative processes through NGOs, with active participation from government agencies and the private sector. For example, groups like W3C, Apache, and Mozilla, have strong support from both public and private sector organizations and have developed open standards and systems that have formed the basis for the Internet economy. One of the key characteristics of their success is the commitment to openness, and that decisions are made based on consensus and ability to demonstrate technical merit and functional pragmatism, rather than the needs or business plans of any particular participant.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

The key venue of decision making in research communities is not necessarily national boundaries but disciplinary communities. Agencies could make it possible for data experts (with a deep understanding of disciplinary needs) to attend disciplinary conferences, to seed and support standards discussions and possibly to staff standards development efforts. Agencies could look to how (and why) private sector companies

have supported development of open source software and open standards through groups like W3C, Apache, and Mozilla, and follow a similar model.

As they have with the Federal Agencies Digitization Guidelines Initiative http://www.digitizationguidelines.gov/, Federal agencies are well-placed to develop leadership roles in establishing best practices. While deeply engaged with disciplines, Federal agencies typically stand outside of them, giving the agencies the opportunity to provide a unique perspective.

Funding joint international demonstration or infrastructure development projects would also be helpful.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

The answer to this is similar to the answer to #9 above – it's a combination of development and support for standards that meet this goal and encouraging broad adoption of the standard by embedding in in the publication, citation, and recognition norms of research communities. Ultimately, it will have to be something that will be easy for researchers to use - highly complex approaches may be technically superior but will likely not be used. The standard should be something researchers are already familiar with – the DataCite project is going with DOI, since these are already well understood in academia and actionable in many publication and discovery systems. Agencies could set expectations that authors should include with publications data citations and DOI links to supporting data (whether it is data they created or from another source) and could encourage publishers to make these links prominent in publishing systems and use the links in reporting and analysis.

A valuable analysis of many of this issue and many of the issues discussed above can be found in this blog: http://opencitations.wordpress.com/

These comments are submitted on behalf of the Duke University Libraries by:

Deborah Jakubs
Rita DiGiallonardo Holloway University Librarian & Vice Provost for Library Affairs

Kevin L. Smith
Director of Scholarly Communications

Paolo Mangiafico
Director of Digital Information Strategy

Myles Axton PhD
Editor, Nature Genetics
Nature Publishing Group
New York NY

Preservation, Discoverability, and Access

(1) Metrics and reporting reduction. If public data can be shown to produce measurable results, then scale up the promotion of data access policies and implementation. If trusting people to maximize their gain from data generated with federal funds works better, let them do it and get government out of the way.

(2) Make the granularity of attribution of data to individuals, grants and teams easily discoverable. Datasets are often submitted by one informatics professional or grantholder but are the product of a team effort. Perhaps patenting could be streamlined to protect data registered in publicly accessible repositories (either as IP or precompetitive).

(3) Researchers in each of the fields will always play up the differences. In practice the most generic data storage with the fullest and most standardized metadata will always be best. Re-use of data from one field to the next sometimes requires handshaking workshops to identify formatting and quality issues and to set standards and formats: (http://www.nature.com/ng/journal/v42/n1/full/ng0110-1.html)

(4) The effort to develop metrics should precede any attempts to enforce data access. The game may not be worth the candle. Public datasets may be only trial runs or burn-in for the datasets and technologies that follow.

(5) Use unique citable identifiers (UUID, DOI, URI, ORCID) for individuals, roles, grants, datasets, samples, departments, institutes, funders and projects. Have unique identifiers for everything that can be shared and an access plan for everything. Where possible the unique identifier should be citable whether or not the data are accessible. Use successful simple templates for data management plans in a journal, database or public repository rather than reinvent them from scratch in a private repository.

(6) There is a metrics requirement here, especially in reallocating resources for IT infrastructure and curation to those projects that justify it. Funding for consortia and for the publishers that present their data and publications to the public would represent a step in funding open access in business models that differ from the current one where the author pays article charges and funders subsidise the publisher. Transparently accounted curation services operated by publishers are a possible alternative to publicly-funded bodies such as NLM (http://www.nature.com/ng/journal/v43/n5/pdf/ng.827.pdf).

(7) Funding agencies will always try compliance approaches first. These are deadening and turn research reporting into a cheating game. Standardization (10, 11) and metrics (1,2,4,5,6) may be more helpful. Rewarding data sharing consortia or defined communties with extra funding for existing grants that are still live - in response to high re-use in substantial secondary publications by other data users – should be tested to see if it will encourage pre-publication data sharing. Minimization of reporting for grantees can be done by engaging publishers who already deposit papers in PubMed Central to help with reporting standards.

(8) The Million Veterans Project should be given help to overcome institutional barriers to become a national cohort for healthcare research and translational improvement via the Veterans' administration. An open interface converting self-reported experiences to medical ontology modeled on Patientslikeme.com would help with recruitment and coordination.

(9) Microattribution (http://www.nature.com/ng/journal/v41/n10/full/ng1009-1045.html) based on ORCID is a central tenet of the drive to data citation via attribution credit. The Datacite inititative is another example that may be useful. I think it is a mistake to have PubMed as the central reputation server

(http://www.nature.com/ng/journal/v41/n4/full/ng0409-383.html), rather standard attribution formats should be used openly with each provider (journal, database, institute, researcher) offering to display attribution credit for the items it holds.

(10) Format datasets in a restricted set of interoperable formats (http://www.nature.com/ng/journal/v43/n1/full/ng0111-1.html)and standardize metadata that contains field-specific reporting standards (Example: http://isatab.sourceforge.net/tools.html).

(11) MIAME: GEO made deposition easy, ArrayExpress made formatting and compliance part of deposition at the price of deterring submissions. The existence of standards and their enforcement does not have the desired result and other incentives are needed (http://www.nature.com/ng/journal/v41/n2/full/ng0209-135.html).
GWAS: A user community, funder (NHGRI) and Nature Genetics decided that replication and correction for multiple testing and stratification would make the technique mor robust to false positives (http://www.nature.com/nature/journal/v447/n7145/full/447655a.html).
ORCID: Thomson Reuter was persuaded Researcher ID would work better if shared
PDF: a proprietary format from Adobe can be replaced by HTML5

(12) Engage with Datacite and with international publishing intitatives (CrossMark from CrossRef) and publishers who get the point (Nature, BMC and PLoS).

(13) Universal versioned DOIs or other persistent granular electronic identifiers. We also need a convention on bidirectional linking as well as technology to make it easy.

Myles Axton, Ph.D.
Editor
Nature Genetics
http://www.nature.com/naturegenetics

ED_ng.1099.0212.SO_PR

nature genetics

It's not about the data

Researchers, funders and journals are in broad agreement that data must be accessible to support the conclusions of scientific publications and for the research to have impact. What is lacking is agreement on timing, formatting and attribution.

In December 2011, the US National Science Board (NSB) presented its report Digital Research Data Sharing and Management, which makes recommendations for the US National Science Foundation (NSF) to implement with its associated scientific and engineering communities. The report acknowledges that there are a broad range of challenges inherent in sharing research data and a need to provide instructions, support and trained professionals to enable data management. The report warns that "one-size-fits-all solutions cannot adequately address most digital research data policy issues because each research community is best suited to address the nuances of its own data." We agree that some communities have more sophisticated approaches to data access than others and that both the style of data presentation and the deal-breaking issues preventing access may differ somewhat by field. However, presenting many different solutions will do nothing to promote interdisciplinary data access. So, our recommendation is that we learn from each field's best examples and then all concentrate on the three crucial issues of timing, formatting and attribution. Each party can then bring what it does best to bear on solving these problems, whether that is funding research, teaching, programming, generating data or publishing.

While keeping up pressure for access to data resources ("No second thoughts about data access"; *Nat. Genet.* **43**, 389, 2011; http://www.nature.com/ng/journal/v43/n5/full/ng.827.html), we have been advocating the use of citable data management plans in line with the proposals of major funding agencies. Like the US National Institutes of Health, the NSF wants a formal declaration of the data resources in each large resource project and their use conditions, whereby "using the Data Management Plan to determine the timeline for initiating the data sharing process recognizes the rights and responsibilities of investigators." The report also recommends that "data should be shared using persistent electronic identifiers, which enable automatic attribution of authors and award funding." As an example of excellent practice in integrative data management, we laud the International Cancer Genome Consortium (ICGC; http://dcc.icgc.org/), which laid out its data policy for its 34 constituent studies in a marker paper (*Nature* **464**, 993–998, 2010). We particularly like the way in which a data management plan written at the grant stage evolves from an explanation of the project and the resources it will generate. As the project progresses, the plan is versioned to detail the databases and data fields that will be generated, with a detailed timeline for data use. The plan finally matures into a 'data descriptor', which we define as a user guide to the resources, accession codes and use conditions accompanying a completed project or publication. One ICGC study currently has a data descriptor in the database of Genotypes and Phenotypes (dbGAP), with accession code phs000370.v1.p1 linking the associated publication (*Science* **333**, 1157–1160, 2011; doi:10.1126/science.1208130) and 883 sequence data depositions in the Short Read Archive (SRA) database. We note that all versions of data plans and descriptors can be citable by digital object identifier (DOI) and can reside online in databases, project websites or journals.

Reformatting data is a full-time job for many researchers, even before the minimum reporting guidelines, terminologies and formats of each field are taken into consideration. In this issue, we present a Commentary and a Perspective suggesting solutions to these problems that have been developed by a process of community consultation and open review to which the journal was a party. In the Commentary, Susanna-Assunta Sansone and colleagues identify one central problem, namely that "most repositories are designed for specific assay types, necessitating the fragmentation of complex datasets," and they offer a unified view of the metadata formatting that will be needed to ensure that biomedical research datasets become interoperable. This solution is the overarching ISA framework, where the acronym stands for 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement) (p 121). This proposal shifts the sets of reporting standards agreed upon by each community into the infrastructure and formatting of the data files themselves. Sansone and colleagues also list a set of participant communities that can pioneer the approach and teach by example. In the Perspective, Jonathan Derry, Stephen Friend and colleagues lay out the infrastructure requirements for a data commons in which all of the data depositors, curators and users become participants who engage with each other and the data by sharing tools and datasets. Their common uniting purpose would be improving preclinical drug design via multidimensional molecular modeling of human disease (p 127).

Within a data commons, attribution for scholarly contributions can be tracked and acknowledged. So, too, in the market of peer citation, and in this issue, the web of coauthorship during the recent years of genome-wide association studies (GWAS) is discussed by Brendan Bulik-Sullivan and Patrick Sullivan (p 113). Recognition of coauthor groups as well as formally declared consortia is the first step to establishing responsibilities for stewardship over complex datasets spanning multiple institutions, journals, databases and funders. Recognizing this need, a complementary approach is being taken by Neil Caporaso and Siiri Bennett (http://hdl.handle.net/10101/npre.2011.6680.1), who sent a survey to the participants in at least 110 of the named consortia in the GWAS field. Consortium information can be sent to these authors or updated by participants via the survey on the WikiGenes site (http://www.wikigenes.org/GWAS/consortia.html), to be published in a future issue. We anticipate that the more complete and granular information about the people who generated knowledge in this field will contribute to sustainable access to the datasets in perpetuity. ■

Hello Ted Wackler,

I am writing to the OSTP office concerning the "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research" that is available at http://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific

I will put in my comments after the numbered sections below.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

> I would like to see of PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/) to include more data as well as journal articles. With the new NSF data management plan requirements, research done with NSF funds could copy the data to an NSF repository. I would also like to see expanded roles for NTIS and the DOE Information Bridge in holding more data from research. I know that NTIS often sells their reports, but it would be better if the reports and data were freely available to the general public. Astronomical data could be held at the NASA ADS with greater Federal support, http://adsabs.harvard.edu/index.html

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

> Where applicable, I would recommend that Federally funded research license their material with a CC by license (http://creativecommons.org/licenses/by/3.0/) or CC0 (http://creativecommons.org/publicdomain/zero/1.0/). This will provide the widest reach to readers throughout the whole world. This will also have the most benefit for scientists, federal agencies, the readers and the citizens of the United States. It may not be as beneficial for commercial publishers, but they have plenty of other non-government sponsored material they can publish.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

> There are many different data types. The Global Change Master Directory provides recommendations to scientists who deposit data to the directory. They provide guides to their metadata writers (Directory Interchange Format (DIF) Writer's Guide). See http://gcmd.nasa.gov/User/difguide/WRITEADIF.pdf and http://gcmd.nasa.gov/User/difguide/difman.html. This guide could be used as a template to help data management writers describe datasets in other disciplines.

The Digital Curation Centre is another good resource to consult, http://www.dcc.ac.uk/resources/data-management-plans. This is another good resource, "National initiatives for promoting data management strategies: an overview," http://sonexworkgroup.blogspot.com/2011/04/national-initiatives-for-promoting-data.html

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

It depends on who needs to use that data, and the intended audience of the research.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

There are many librarians who are getting to be a lot more familiar with data management plans and e-science. I would recommend that the government work with university programs such as those listed at http://www.arl.org/rtl/eresearch/escien/nsf/nsfresources.shtml.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

I am not sure.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Scientists need positive reinforcement for depositing and describing their data. If they received more grant funding for cooperating in projects, or if they received greater recognition by university administrators, then that would be some positive rewards for compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There are always more mashups that could be done with GIS data and social science data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data sets could be given a permanent citation link, such as a DOI. http://www.doi.org/ I would recommend that you read some of the papers presented at this conference, http://sites.nationalacademies.org/PGA/brdi/PGA_064019 "Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop"

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

This chapter might be of use to you. http://www.ncbi.nlm.nih.gov/books/NBK45678/ "The Current State of Data Integration in Science" found in the book, Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop. National Research Council (US) Committee on Applied and Theoretical Statistics. http://www.nap.edu/catalog.php?record_id=12916

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

I can't find any right now.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Start with one country, and then start working with other countries. I'd recommend that you take a look at the policies of the United Kingdom. Consider looking at http://www.dcc.ac.uk/resources/policy-and-legal/policy-tools-and-guidance and http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/establishing-a-digital-preservation-policy/.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

I would recommend that you take a look at this article, http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101 for some practices that are used.

Response to this RFI is voluntary. Responders are free to address any or all the above items, as well as provide additional information that they think is relevant to developing policies consistent with increased preservation and dissemination of broadly useful digital data resulting from federally funded research. Please note that the Government will not pay for response preparation or for the use of any information contained in the response.

Name/Email - Joseph R. Kraus

Affiliation/Organization – Science & Engineering Librarian, University of Denver, Penrose Library, 2150 E. Evans Ave.

City, State – Denver, CO 80208

In addition, please identify any other items the Working Group might consider for Federal policies related to public access to peer-reviewed scholarly publications resulting from federally supported research.

Please attach any documents that support your comments to the questions.

.

To: Office of Science and Technology Policy (OSTP) [*digitaldata@ostp.gov*]

Fr: Samuel Helfaer, Bachelor of Science, Physics, Yale University [*samuel.helfaer@yale.edu*]

Re: Proceed with Public Access to Digital Data From Publicly Funded Scientific Publications

EXECUTIVE SUMMARY

- The United States Government spends in excess of $145 Billion a year to support research and development to produce new knowledge.
- This new knowledge fuels innovation in many diverse fields such as medicine, technology, climate science, defense and agriculture. This improves the American economy through both higher productivity and new sources of productivity.
- Open access to non-confidential data and publications produced using government funds would increase the government's return on investment for research, improve our economy by increasing productivity and decreasing unemployment, and, in turn, provide further incentive to invest more in science.
- An open access protocol will allow more collaboration and less unnecessarily repeated science as well as the opportunity to improve academic rigor and repurpose data.
- This effort will also enable common people to educate themselves and attempt to distill data, thus encouraging STEM education and creating an opportunity for research with low barriers to entry. This will catalyze an improvement in STEM research as well as STEM literacy in the American population, further improving our economy.
- Open data will make peer review more effective and promote a more collaborative model of discovery while maintaining the incentives provided by competition.
- While there are significant logistical challenges to opening access to all data produced through federally funded research, the benefits far outweigh the costs.
- The Human Genome Project as well as the NIH and NSF data sharing policies demonstrate the feasibility and promise of open data in federally funded research.
- Open access to data from federally funded research will spur innovation by other federal researchers as well as those in private industry.
- Open data will contribute to better policy and decision-making by individuals, corporations and government agencies and improve the American economy through job growth, productivity increases and happier, healthier lives for all American citizens.
- Opening data to the taxpayers that fund federal research will ensure American economic prosperity, strengthen national security and propel the pursuit of knowledge forward.

BACKGROUND: FEDERALLY FUNDED RESEARCH

Every year the United States government invests more than $145 billion in funding for research and development (R&D)[1]. Countless additional dollars are spent elsewhere (or not collected in taxation) in order to incentivize innovation and improve the welfare of the American people and people throughout the world. This funding has yielded incredible advances in virtually every field of science and technology from defense and agriculture to artificial intelligence and medicine. These research dollars come from taxpayers, including students, other researchers and scientists, as well as many hard-working men and women. While all of these constituencies serve to benefit, they generally do not have access to the resultant data.

Remarkable advances have been made in both basic and applied research and spurred inventions that have improved every facet of our lives. Still, we, the American people, are not realizing the full potential benefits of this research. Much of the data collected, programs and algorithms created and insights gleaned remain in the hands of a precious few. Even when an academic paper is published with open access, the clean, peer-reviewed, edited academic paper often belies reams of data and years of work that don't make it to the final draft. The academic paper has become a vital means of communicating research and serves a valuable purpose, however, there is much to be gained from sharing all that went into its creation.

Allowing others to utilize all of the data produced from federally funded researchers will promote innovation through both collaboration and competition. It will enable other scientists to avoid conducting the same experiments as well as confirm the rigor and validity of studies

[1] Sargent Jr., John F. "Federal Research and Development Funding: FY2011."*Congressional Research Service Report for Congress* R41098 7.5700 (2011). <http://www.fas.org/sgp/crs/misc/R41098.pdf>.

conducted on the taxpayer's dime. Through open data, research dollars will be used more effectively and efficiently. This will yield a higher return-on-investment and, in turn, higher impact research will trigger more research dollars and more innovation. Similarly, through open access to data, STEM students will have an unprecedented ability to engage with real research data and make a contribution from an early stage in their education. As the world economy becomes increasingly globalized, strong domestic research and development efforts coupled with strong STEM education will keep the American economy competitive in the global economy.

AMERICA COMPETES REAUTHORIZATION ACT OF 2010: DIGITAL DATA

The America COMPETES Reauthorization Act of 2010 established a working group under the National Science and Technology Council (NSTC) to "coordinate federal science agency research and long-term stewardship of the results of unclassified research, including digital data and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the federal science agencies." It also stipulates that "[OSTP] shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, access, including online access, and long-term preservation of such collections for the benefit of the scientific enterprise."[2]

This brief intends to outline the benefits and implementation challenges of open data specifically related to the Request for Information (RFI) pertaining to "Public Access to Digital Data Resulting From Federally Funded Scientific Research"[3] released in accordance with

[2] 1861, 111 Cong., America COMPETES Reauthorization Act of 2011 3985-3987 (2011) (enacted). <http://www.gpo.gov/fdsys/pkg/PLAW-111publ358/pdf/PLAW-111publ358.pdf>.
[3] Wackler, Ted. "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." *Federal Register* 76.214 (2011): 68517-8518. <http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/pdf/2011-28621.pdf>.

Section 103(b)(6) of the America COMPETES Reauthorization Act of 2010 (ACRA; Pub. L. 111-358).

HIGHER IMPACT, HIGHER RETURN ON INVESTMENT

Measuring the impact of scientific research across diverse fields can be difficult and highly subjective. Many metrics have been developed to assess the impact or value of a given study. These metrics often include the number of times other papers cite a given paper as well as the impact of those papers. Consistent with this, open access articles are found to be cited more frequently even when controlling for article quality, indicating that opening access to articles contributes to their impact.[4] A similar effect is seen with open data as other researchers can build on the research already conducted. By releasing more data, the opportunities for usage increase proportionally.

When more researchers use data gained through federally funded research, the research dollars go further. The opportunities for progress are larger and more numerous and represent a significant increase in return-on-investment (ROI) for taxpayers and government agencies. Higher impact, higher ROI scientific research means more funding which will directly benefit researchers who compete for research dollars as well as the American people who benefit from the new knowledge and technologies.

CULTURAL CHALLENGES: THE REPUTATION CONUNDRUM

At present, primary external motivators for researchers include money, reputation, and time. Money and time, along with space and other resources allow researchers to pay themselves and

[4] Gargouri Y, Hajjem C, Larivière V, Gingras Y, Carr L, et al. 2010 Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE 5(10): e13636. doi:10.1371 <http://www.plosone.org/article/fetchObjectAttachment.action;jsessionid=55B7DC69A50002ADB4E7B4F02FF2FBA0?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0013636&representation=PDF>.

others in their group, provide the necessary equipment for experimentation, and have the time to think and do research. Reputation, attained through successful research, ensures continued access to time, money and other resources. Given the importance of reputation, much of what drives scientists is derived from how reputation is measured. In our current construct, reputation is often measured by citations, h-indices, number of papers published, and download counts.[5] This incentive structure places the primary focus on the publishing phase of research and paints this as an ultimate step, downplaying continued engagement with published research. In this "publish or perish" environment, any time spent away from publishing and grant-writing serves as essentially wasted time.

CULTURAL CHALLENGES: REALIGNING INCENTIVES

While academic papers play a valuable role in the spread of information, they lack much of the information gleaned during the research project. Time spent by researchers cleaning up and publishing data, meta-data and code and engaging with other researchers on a blog or other communication interface is time spent furthering human knowledge and science. By not incentivizing or requiring this information to be published, we are sacrificing millions of dollars in potential innovation and unnecessarily repeated research and creating a lack of transparency that can allow academic dishonesty and poorly conducted research.

This incentive structure has become deeply ingrained and institutionalized in many academic disciplines. Changing this culture requires a realignment of incentives to favor publishing of, and engagement with, data. This can be achieved through open data stipulations accompanying federal research funds as well as consideration of the open access contributions of researchers (both previous and in the context of a proposed project) when evaluating federal grant

[5] Siemens, George. "What, Exactly, Is Open Science? | The OpenScience Project." *The OpenScience Project | Open Source Scientific Software*. Web. 05 Dec. 2011. <http://www.openscience.org/blog/?p=269>.

applications. Through enhanced options and requirements of open data, tenure committees and hiring decisions as well as other reputational measures will increasingly hinge on a researcher's participation and impact in open science.

Many models are currently being used to catalyze this shift toward open access. ORCHID (Open linguistic Resources CHannelled toward InterDisciplinary research), for example, explicitly incentivizes blog post and other less formal peer review while the NIH and NSF both have data sharing stipulations in their research grants. This is not enough. Without requiring data to be published across all federal funding sources, we are missing out on a substantial portion of the potential research impact and wasting money on redundant research. We are not optimizing the performance of our research and development minds and resources if they do not have access to the most current and complete datasets possible.

CONFIDENTIALITY CONCERNS

As with all experiments, it is vital to ensure that all privacy and confidentiality concerns are considered with the utmost care. Any information that could compromise personal privacy, national security or intellectual property rights should not be released. These concerns are especially relevant in the medical[6] and defense spaces and, as such, all grants that could include proprietary or confidential data should be granted in accordance with an explanation of what data will likely need to be withheld from open access publishing.

COMPETITION CONCERNS: INTELLECTUAL PROPERTY

Both collaboration and competition are essential to successful innovation and groundbreaking research. With open access to data and experiment code, many other researchers will have the

[6] See HIPAA, the Health Insurance Portability and Accountability Act.

potential to benefit from the hard work of those who have already published their data. While this will allow more minds to collaborate, it also has the potential to diminish the incentives of developing proprietary algorithms and methodologies as well as collecting proprietary data. By allowing researchers to maintain control of their data and code, we provide them an opportunity to conduct as much analysis and algorithmic optimization as they desire with little acute time pressure. In order to preserve the incentive that proprietary data and code provides, it is essential to allow scientists to request more time with their proprietary data and code before release (similar to a patent). This should be able to be included in their initial grant application or in a subsequent application before they have published their papers or data. Allowing a patent-like system for researchers will maintain the incentive to make large research advances.

This extended proprietary-data period should not apply to sectors with well-defined databases and methodologies in which the researcher likely received substantial benefit from the already established database and has merely made an incremental advance. While an extension-of-proprietary-usage provision incentivizes large innovation, it has the potential for abuse. This policy should be used sparingly and continually monitored to ensure it is fulfilling its intended role and not being overused.

CONCERN: PEER-REVIEW AND THE COST OF DATA-SCRUBBING

One of the tenets of modern research is peer-review. This review process is meant to keep scientists conducting honest, rigorous and important research through the oversight of a community of knowledgeable colleagues. While much peer-review is currently undertaken in conjunction with journals, with a realignment of incentives toward valuing community review and collaboration, open review can become a major component of research. By opening up

data to the greater research community and beyond, the opportunities for peer-review become significantly larger and more powerful.

Many arguments against requiring data and in-depth methodology publication hinge on the challenges of cleaning up data and study specifications. This process of data (and methodology) scrubbing can be time-consuming, however, it is worthwhile. This process provides this data for many other potential researchers. More importantly, any study that is being peer-reviewed should have the data and meta-data available to be audited and reviewed by a reviewer. A peer-reviewer is not merely supposed to read over the academic paper and check for grammatical mistakes and guarantee stylistic writing. It is their responsibility and goal to ensure the academic rigor and honesty of the study. While data is often made available in peer-review processes, it should be made generally available with low barriers to entry so that all stakeholders can evaluate the methodologies, data and conclusions without merely looking through the small window into what the researchers felt to be important or conclusive.

ACADEMIC RIGOR THROUGH OPEN SCIENCE

The dangers of closed science have come to light in a variety of cases recently.[7] In one highly publicized case, a famous social psychologist, Diederik Stapel, fabricated immense amounts of data over 15-20 years and 100 publications are now under investigation.[8] Similarly, a recent study entitled "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" demonstrates how easy it is to

[7] Fanelli, Daniele. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." Ed. Tom Tregenza. *PLoS ONE* 4.5 (2009): E5738. <http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0005738&representation=PDF>.

[8] Crocker, Jennifer, and M. Lynn Cooper. "Addressing Scientific Fraud." *Science Magazine* 334.6060 (2011): 1182. <http://www.sciencemag.org/content/334/6060/1182.full>.

accumulate and report statistically significant evidence for a false hypothesis.[9] In their article, they point out that leaving out certain data can be as important as keeping certain data and that current practices encourage reporting exclusively "what worked": that is, those results that were statistically significant. Decisions about data manipulation are not made in advance of the data collection. Simmons et al. propose guidelines that would require researchers to, for example, disclose every question asked on a survey not merely the questions with statistically significant responses. This same goal can be achieved through requiring federally funded researchers to release all of their collected data. Similarly, releasing all collected data is a much more effective and powerful mechanism for preventing undue corporate (or other conflict of interest) influence than merely requiring authors' disclosure in small print.

As they explain "Our goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth…We should embrace these [proposed rules about disclosing research methods] as if the credibility of our profession depended on them. Because it does." By requiring the release of data for federally funded research, the process of peer-review will become more effective and academic honesty and rigor will improve. This cultural shift will also catalyze the release of data in research not funded by federal dollars yielding a research multiplier effect. This will increase both the quality and value of research done in the United States and propel our pursuit of knowledge forward.

THE CURSE OF EXPERTISE: AMATEUR SCIENTISTS

Wikipedia provides a case study of the amazing advances that can be made through open access. Another oft-cited example of an open access success is Tim Gowers' Polymath

[9] Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11 (2011): 1359-366. <http://pss.sagepub.com/content/22/11/1359.full.pdf>.

project[10], which allowed 27 mathematicians to prove a theorem that had stumped many famous mathematicians using an online cooperative model. Still, academic research should not become Wikipedia. There is a valuable place for expertise. Data that is released should not need to meet an arbitrary standard of understandability to a layman (as Wikipedia does). Likewise, while the data should be accessible to everyone, there will certainly be means of stratifying the research conducted with the data. A tenured professor should not be required to directly reply to a high school student who thinks the researcher has made a mistake or if the high-schooler does not understand their methodology – this is the role for an online community forum. A confused young person or amateur scientist with their own theory can initiate a conversation in an online forum. This conversation could be seen by a colleague of the researcher who, in turn, may validate its legitimacy and ask the author to explain that portion of the study.

By allowing authors' colleagues to repeat their experiment or build on their experiment without having to recreate and back-engineer every method, open access to data and methodology will accelerate the pace of research. In this process, researchers may have to explain their data or process to some of the others in their field but those colleagues will ultimately reciprocate and a dynamic, collaborative model of research will flourish.

KEY COMPONENTS OF DATABASE IMPLEMENTATION

Given the current databases in place and rapidly evolving database landscape, an attempt to explicitly design and implement a single, all-inclusive database for federally funded research would be both logistically challenging and of questionable long-term value. Still, there are guiding principles that should steer our vision of successful database creation. These databases

[10] Parker, Matt. "Welcome to WikiMaths: Home of Hard Sums." *Shortcuts Blog*. The Guardian, 8 May 2011. Web. 05 Dec. 2011. <http://www.guardian.co.uk/science/2011/may/08/welcome-to-wikimaths>.

should allow anyone to access and search through the data and meta-data readily and without buying or downloading any proprietary software. In short, databases should be easy enough to use that an average taxpayer could use them (even if they do not necessarily understand all of the data, methodologies, etc.). In this way, we can democratize the data that taxpayers fund.

Databases should:

1) Be machine searchable and indexable by other databases (such as google).
2) Maintain data in non-proprietary formats (such as csv, xls, txt, etc.).
3) Aggregate and organize data based on usage, researchers and other criteria to ensure proper secondary accreditation and citation as well as ease-of-use.
4) Attempt to include all available information within its purview, as it is hard to anticipate the needs and desires of other researchers (and other users).
5) Automatically update when new data is released/published or be easily updateable to adapt to the new model of dynamic scientific research.

LOGISTICAL CHALLENGES: DATA STORAGE AND EXPENSE

Federally funded research produces an enormous amount of data and meta-data each year. While much of this data may seem unimportant or excessive, especially as compared to the data presented within and alongside the final academic paper, it has the potential to benefit other researchers and citizens. As the cost of data storage and aggregation continue to fall, the cost to value ratio of storage and maintenance of data continues to improve.

Many databases are currently being developed both in connection with the government and through private endeavors. Given this current fragmented landscape, the key objectives going forward should be 1) increased access to data, 2) aggregation and organization of that data, and

3) interoperability between platforms. The government can play a direct role in goal 1 and a more indirect role in goals 2 and 3.

The Human Genome Project, the NIH Data Sharing Policy, and the NSF Data Management Policy all provide frameworks for successful implementation of government data sharing programs. Other smaller ventures outside of the government's direct purview include PLoS (Public Library of Science), SPIRES/INSPIRE (High-Energy Physics), and JASPAR (DNA transcription factors binding preferences) and VIVO (an academic-institution-only research network. All of these models are slightly different and reflect the realities in their given academic discipline(s), however, they would all benefit from more open data. By creating more open data, the value of aggregation, organization, and interoperability increases exponentially[11], incentivizing these goals by others. The government can also pursue goals 2 and 3 by providing grants to fund the creation of new databases and to improve those currently in operation.

STEM EDUCATION: ENABLING AMATEUR SCIENCE

The primary purpose of open data is not to allow amateur science or to improve STEM education. However, this will likely be one of its primary long-term positive effects. Open data will promote a collaborative model that allows amateurs and students access to much more data and resources than they do under the current model as well as opportunities to interact more readily with the greatest minds in the fields they are interested in through their research. This positive "collateral damage" will create future generations of scientists who are more interested in helping young people develop their analytical and experimental skills, further improving STEM education and literacy in America.

[11] Similar to Metcalfe's Law of Network Value, which says that network value is proportional to n^2, where n is the number of users in the network. In the research case, there is an additional positive feedback mechanism.

By opening access to data produced with federal dollars, we will promote a more scientifically inclined, academically rigorous society. This will help improve the productivity of the American economy and maintain our impressive track record as the leaders in global R&D and education.

SUCCESS WITHIN INDUSTRY

Another positive of open data is allowing private industry to benefit and innovate more directly from basic and applied research conducted through federal grants. This will, in turn, benefit both the United States economy as well as the American people who will have access to improved medical care, new technologies and less expensive manufactured products. Many companies already operate through a combined collaborative and competitive model in which data is shared even as various divisions compete for funding. This model has provided many corporations with record profits and rapid innovation. Adapting this model to the public sector will allow federal dollars to produce more knowledge, innovation and economic benefit.

INFORMED AND DYNAMIC POLICY-MAKING

One of the key benefits of open access to data outside the realm of academia is that it enables individuals and organizations to make informed decisions based on the best and most-recent data available. This allows the actors in our economy to perform more efficiently and improves the overall productivity of our economy. Remarkably, many of our government bodies do not have access to relevant research or data that could allow them to make better policy-related decisions. While economic data can be easily extracted from the bureau of labors statistics

online[12] (as one example), attempts to access scientific data are often met with pay-walls or simply academic papers containing a small percentage of the actual data collected.

The lack of accessible, organized, trustworthy data in our economy prevents individuals, corporations and government agencies from making the best fact-based decisions about vital issues such as agricultural policy, national defense, climate change and healthcare. This promotes inefficiencies in our economy and can have disastrous consequences as many of these actors (especially large corporations and large government agencies) have the ability to affect large changes.

OPEN DATA: FOR A BETTER AMERICAN FUTURE

"Improving the way that science is done means speeding us along in curing cancer, solving the problem of climate change and launching humanity permanently into space. It means fundamental insights into the human condition, into how the universe works and what it's made of. It means discoveries not yet dreamt of. In the years ahead, we have an astonishing opportunity to reinvent discovery itself. But to do so, we must first choose to create a scientific culture that embraces the open sharing of knowledge."[13] As Michael Nielsen, a pioneer in quantum computing, aptly describes, we stand at a turning point in human knowledge. Never before has the pace of innovation and acceleration of discovery been as rapid. We must seize this opportunity to use what we are learning to its fullest and propel discovery forward.

The possibilities for open data extend beyond science. By requiring that all federally funded research and development release data and meta-data into the public domain, we will improve

[12] http://www.bls.gov/data/
[13] Nielsen, Michael. "The New Einsteins Will Be Scientists Who Share." *The Wall Street Journal*. 29 Oct. 2011. Web. 05 Dec. 2011. <http://online.wsj.com/article/SB10001424052970204644504576653573191370088.html>.

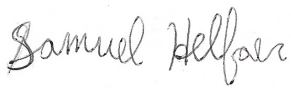
the American economy and our national security and allow us to make vital data-driven decisions about how we will face the seemingly insurmountable challenges that lie before us. Open data will allow us to take full advantage of our rapidly growing and evolving scientific knowledge to guide humanity forward on a successful trajectory.

No one can fully predict how open data will evolve. I am certain, however, that open access to our data will lead to a better tomorrow.

Thank you for your time and consideration.

If you have any questions for me, please do not hesitate to contact me at Samuel.Helfaer@yale.edu.

Sincerely,



Samuel Helfaer

Constituent, Pennsylvania's 6th Congressional District

Bachelor of Science Candidate, Physics, Yale University

References

1861, 111 Cong., America COMPETES Reauthorization Act of 2011 3985-3987 (2011) (enacted). <http://www.gpo.gov/fdsys/pkg/PLAW111publ358/pdf/PLAW111publ358.pdf>.

Crocker, Jennifer, and M. Lynn Cooper. "Addressing Scientific Fraud." *Science Magazine* 334.6060 (2011): 1182. <http://www.sciencemag.org/content/334/6060/1182.full>.

Fanelli, Daniele. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." Ed. Tom Tregenza. *PLoS ONE* 4.5 (2009): E5738. <http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0005738&representation=PDF>.

Gargouri Y, Hajjem C, Larivière V, Gingras Y, Carr L, et al. 2010 Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLoS ONE 5(10): e13636. doi:10.1371 <http://www.plosone.org/article/fetchObjectAttachment.action;jsessionid=55B7DC6A50002ADB4E7B4F02FF2FBA0?uri=info%3Adoi%2F10.1371%2Fjournal.pone.001636&representation=PDF>.

Larson, Phil. "Request for Information on Public Access to Digital Data and Scientific Publications." *The White House*. 7 Nov. 2011. Web. 05 Dec. 2011. <http://www.whitehouse.gov/blog/2011/11/07/request-information-public-access digital-data-and-scientific-publications>.

Nielsen, Michael. "The New Einsteins Will Be Scientists Who Share." *The Wall Street Journal*. 29 Oct. 2011. Web. 05 Dec. 2011. <http://online.wsj.com/article/SB10001424052970204644504576653573191370088.html>.

Parker, Matt. "Welcome to WikiMaths: Home of Hard Sums." *Shortcuts Blog*. The Guardian, 8 May 2011. Web. 05 Dec. 2011. <http://www.guardian.co.uk/science/2011/may/08/welcome-to-wikimaths>.

Sargent Jr., John F. "Federal Research and Development Funding: FY2011."*Congressional Research Service Report for Congress* R41098 7.5700 (2011). <http://www.fas.org/sgp/crs/misc/R41098.pdf>.

Siemens, George. "What, Exactly, Is Open Science? | The OpenScience Project." *The OpenScience Project | Open Source Scientific Software*. Web. 05 Dec. 2011. <http://www.openscience.org/blog/?p=269>.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11 (2011): 1359-366. <http://pss.sagepub.com/content/22/11/1359.full.pdf>.

Wackler, Ted. "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." *Federal Register* 76.214 (2011): 68517-8518. <http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/pdf/2011-28621.pdf>.

Attn:
Office of Science and Technology Policy
725 17th Street, Washington, DC 20501

RE:
OSTP RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Massachusetts Institute of Technology's comments on Federal Register Document 2011-28621

Claude R. Canizares, Vice President for Research and Associate Provost;
Ann J. Wolpert, Director, MIT Libraries /
Massachusetts Institute of Technology
Cambridge, MA

The Massachusetts Institute of Technology (MIT) appreciates the opportunity to comment on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research. The comments below, in concert with our comments specific to Federal Register Document 2011-28623 (OSTP RFI: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research), affirm MIT's belief that public access to unclassified research and the data collected as part of research funded by Federal science and technology agencies is a topic of substantial significance to this institution because MIT's mission includes a commitment to generate, disseminate, and preserve knowledge. This commitment carries particular weight when the new knowledge generated at MIT flows from federally funded research. We address each question from this RFI in turn:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Comment 1: As articulated in MIT's response to the OSTP RFI on Public Access to Federally Funded Research last January, MIT believes that a key first step in providing access to the research results of federally funded research is to expand the goals of NIH's public access policy to other federal funding agencies. Providing access to the data without the context of the corresponding, peer-reviewed research results would be short-sighted and create an unnecessary barrier to other researchers and interested parties in being able to interpret and re-purpose the data in productive ways. In addition, the research data resulting from federally funded research should be subject to a data management and sharing policy, similar to either NIH's current policy, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html, or the NSF's, http://www.nsf.gov/eng/general/dmp.jsp. To generate the most benefit it will be important to avoid unnecessary proliferation of varying requirements. Critical to the success of the NIH's policy has been the infrastructure provided by NIH through the National Library of Medicine's National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov/. It's also important to note that there are other efforts currently supported through collaborations between federal agencies and research institutions for access to other important disciplinary domains in the sciences, e.g. http://cdp.ucar.edu/, and social sciences, http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/fifty/factsheet.html. Leveraging the infrastructure created by the NCBI and others to include other research domains rather than suggesting that each federal agency funding research create their own infrastructure should be seriously considered.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2: Existing copyright laws already provide a framework for protecting the IP interests of all. Deploying open licensing and/or waiver protocols for data made available under public access policies would expand the benefits of openness to allow for more innovation, by allowing for data mining and creating new works from existing works. However, because the IP landscape is so complicated, significant educational efforts are needed and will need to continue to ensure these stakeholders understand these complex issues particularly as they relate to data. Useful background on this topic can be found at http://sciencecommons.org/projects/publishing/open-access-data-protocol/, http://creativecommons.org/about/cc0, http://www.dcc.ac.uk/webfm_send/332, and http://www.nature.com/nature/journal/v461/n7261/full/461171a.html.

Of particular concern is a scenario where data is transferred to publishers who then assert copyright due to value added services, e.g., extended and/or normalized descriptive information. While services of this type should not be limited, it will be important to insure that the original data resulting from unclassified federally funded research is still publicly available. It is important to note that MIT, like most research institutions, has explicit rules regarding compliance with HIPAA, http://web.mit.edu/committees/couhes/procedures_healthcare.shtml.

Whatever policies are developed, consistency of requirements is the key element that will allow federal agencies to maximize the benefits of their public access policies. Based on our experience supporting the NIH Public Access Policy and the MIT Faculty Open Access Policy, compliance will rise directly with convenience to the author. For this reason, common procedures, requirements, and processes should be established across all funding agencies whenever possible.

Another key factor in protecting the IP interests of stakeholders is adopting an agreed upon standard for citing data. This will enable the easy reuse and verification of data, allow the impact of data to be tracked, and create a scholarly structure that recognizes and rewards data producers, http://datacite.org/whatisdatacite.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3: Working with disparate stakeholders to develop a minimum set of core metadata for all datasets along with an API for standards based data exchange will help ensure a level of interoperability and discovery across all disciplines. Also, these inherent differences mean that there is a need for flexibility in funding amounts for data curation, and a commitment by agencies to provide the necessary funds for the data curation. Again, consistency of requirements as much as possible is the key element that will allow federal agencies to maximize the benefits of their public access policies.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Comment 4: Flexibility in cost models for data curation will be necessary. The importance of having a data management plan that is adequately resourced prior to the beginning of the research is paramount to prevent unnecessary costs and the possibility of data loss during the life cycle of the research. In addition, there needs to be a clear understanding that the long-term stewardship of the research data, when appropriate, is a much deeper commitment than that of costs incurred during the life of the research project. For such cases the intent will be for the data to be preserved and disseminated well beyond the life of the particular project.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Comment 5: The current landscape is complicated by the heterogeneity of practice across disciplines, a lack of common standards, rapidly developing (and changing) tools for data management, and confusion regarding roles and responsibilities. Developing standard practices for data attribution and citation will be critical. Collaboration among all stakeholders will also be necessary to minimize costs and maximize data sharing. Raising awareness for all those involved in the enterprise is necessary. Most research libraries and institutions are now involved in advising researchers when needed on best practices for data management, and many are working to develop tools to support the long-term preservation and dissemination of research data within the parameters of intellectual property concerns. Examining how existing federal infrastructure, e.g. NCBI, can be leveraged to support long-term stewardship and dissemination of different types of data resulting from federally funded research will also be important to prevent a scenario where proprietary solutions are developed which might be counterproductive to the goals of public access over the long term.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Comment 6: Recognition that the costs associated with preserving and making digital data accessible beyond the life time of the research project is vital, and providing options – whether it is providing funds within the research project to cover this long-term cost and/or providing infrastructure to manage the preservation and accessibility to the data, e.g. NCBI – is an absolute requirement if this effort is to be successful.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Federal granting agencies will need to develop roles, responsibilities, and procedures to review grants awarded to ensure compliance. While requiring that individual responsibility for the management of the research data be assigned and made publicly known might be considered, it will be paramount to develop approaches that both make it easy for researchers to comply and add value to their research efforts. The successful implementation of data management plans from previous awards might also be considered when examining new grant proposals.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8: Encourage data creators to use the Creative Commons CC0 license whenever possible, http://creativecommons.org/about/cc0. Promote "success stories" that demonstrate the successful use of secondary data in advancing research and productivity. Establishing a new granting program to develop innovative tools for mining scientific research data should be considered.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9: Adopt standard practices for data attribution and citation, and require that these practices be required for funding and publication. Relevant initiatives currently underway are DataCite, http://datacite.org/, and ORCID, http://orcid.org/.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Comment 10: The Data Documentation Initiative, http://www.ddialliance.org/, is an example of an initiative that works collaboratively to develop and adopt standards across different disciplines and stakeholders. These standards enable machine usability of the data, as well as facilitate data documentation throughout the life cycle. Balancing discipline specific needs against the desire to have easy interoperability and repurposing of data will be challenging, and may require the adoption of a simplified core set of metadata standards for all data types. Also key will be the implementation an API for standards based data exchange.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Comment 11: Relevant examples are the Internet Society and its IETF, http://www.ietf.org/, the W3G, http://www.w3.org/standards/, the Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/, and the DOI, http://www.doi.org/index.html. Characteristics that have made these successful are that they are completely open and transparent, and that they typically require a proof of concept.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Comment 12: Work proactively with national and international standard bodies.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13: Persistent identifiers for data sets are critical. Also important will be enabling relationships between different versions of data sets to be made visible, something similar to the CrossMark service being developed by CrossRef, http://www.crossref.org/crossmark/index.html. Again, the DataCite initiative, http://www.datacite.org/, is a useful forum and resource to further explore these issues.

Ann J Wolpert
Director
MIT Libraries
http://libraries.mit.edu

Public Access to Digital Data Resulting from Federally Funded Scientific Research

http://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific#p-32

RFI #2011-28621
Released 11/04/2011

Responders:

Daniel Crichton
Daniel.J.Crichton@jpl.nasa.gov
Planetary Data System, Engineering Node
NASA Jet Propulsion Laboratory
Pasadena, California 91109

Faith Vilas
fvilas@psi.edu
Planetary Data System, Chief Scientist
Planetary Science Institute
Tucson, AZ

J. Steven Hughes
Steve.Hughes@jpl.nasa.gov
Planetary Data System, Engineering Node
NASA Jet Propulsion Laboratory
Pasadena, California 91109

Susan Slavney
Planetary Data System, Geosciences Node
slavney@wunder.wustl.edu
Washington University
St. Louis, Missouri

Reta Beebe
rbeebe@nmsu.edu
Planetary Data System, Atmospheres Node
New Mexico State University
Las Cruces, NM

Raymond Arvidson
Planetary Data System, Geosciences Node
arvidson@wunder.wustl.edu
Washington University
St. Louis, Missouri

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

There are a number of policies that would encourage the treatment of data as a national asset in combination with encouraging the use of the data to enhance scientific discovery. These include:

1. Agency requirements for sustainable infrastructures: Agencies should be required to invest in and maintain public archives. These archives should be considered essential "facilities" and provided sustained funding.

2. Requirement for funding: The delivery of data, within a specified amount of time, to national, public archives, should be a requirement for funding public scientific research.

3. Release after a period of time: Scientists and data providers should have a period of time in which they have exclusive access and use of the data prior to delivery to pubic archives.

4. Capture data in reliable formats: Data should be captured in long-term, sustainable data structures that limit the use of proprietary data formats. Ample descriptive information (e.g., metadata) should be provided to support interpretation of the data long-term.

5. Auditing and certification of "official" archives: The U.S. should establish core guidelines for public archives and perform regular auditing for compliance against those guidelines.

6. Ensure active participation of discipline experts: The involvement of discipline scientists in sustaining the data is critical to ensure proper preservation and usability of the data.

7. Policies that encourage the use of data from public archives: Funding should be made available from agencies supporting scientific research that requires analysis of data from public, scientific archives.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Encouraging the public release of data is a critical step in sharing scientific research results. Often the incentive to share data is the result of a funding requirement. As such, scientists and data providers will do the minimum to prepare and share the data. If sharing the data affects their ability to publish and/or receive proper recognition for the research results, the incentive is diminished further. Incentives need to be in place whereby researchers are rewarded for sharing data. Furthermore, a specific set of practices can be put in place to protect researchers' IP interests better and encourage data sharing if it can be considered on the caliber of a peer-

reviewed publication. These include:

1. Release after a period of time: Scientists and investigators who have acquired the data,should have exclusive access to the data for a period of time. This should be for the purposes of improving the reliability of the discovery and results as well as corresponding publications.

2. Separate Intellectual Property (IP) from data: Specific algorithms and methodologies used to support the acquisition and generation of the data may be considered intellectual property of the investigator. While capture of the provenance information that produced those data is critical to understanding the heritage, there are opportunities for releasing public data while protecting the specific techniques used by the investigator.

3. Citations: Develop citation of data in public archives as a standard, scientific practice. A citation of high quality data should be considered equivalent to the citation of a scientific publication. Public archives should provide support for data citations.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Scientific involvement and oversight in establishing and maintaining usable scientific archives is essential.

1. Discipline leaders need to be involved: As prominent experts in their fields, discipline leaders should create a bridge between the scientific community and the public archives ensuring that the archive is scientifically useful and that the local policies are consistent with the scientific needs.
2. Peer review: Scientific review of the data is important for ensuring usability of both the metadata and data itself. Peer review ensures that the metadata can effectively used to annotate and understand the data. For the data itself, it helps in ensuring scientific usability.
3. Local standards: Proper annotation of the data is dependent on having discipline-specific descriptions that can be used to describe the data fully.
4. Sustainable infrastructures (establishment of national archives): Sustainable infrastructures are critical to ensuring long-term stewardship. These infrastructures must address use of the data for specific disciplines and therefore should involve the discipline experts in their implementation and operations.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Agencies should take a long-term, agency-wide view of preserving data in formats that support long-term analysis and use. Rather than fund each individual research project, agencies should develop sustainable infrastructures that are separately funded. This is essential for treating these infrastructures as facilities.

5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data

management plans?

The community should participate in the implementation and operations. This can be accomplished in multiple ways including:

1. Open source development: Development of the necessary computing infrastructure and publication as open source software. This is an effective approach for fostering collaboration.
2. Peer review: As mentioned, involvement of the community in review of the data is an effective way to foster collaboration.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Cost models need to be developed that take into account the entire data lifecycle including the supporting infrastructure. Funding is essential for ensuring data are captured and preserved for long-term usability through an infrastructure that supports effective access to the data. As a result, recommendations include:

1. Develop cost models: Cost models should be developed for disciplines that address both the cost of preparing and ingesting data as well as the long-term preservation. These costs should identify both the costs for the data supplier as well as the archive itself.
2. Archives as facilities: Archives should be treated as facilities. Their funding should originate from a sustained, operational allocation rather than be funded per experiment or investigation. Individual investigators should have a portion of their funding allocated towards preparing and submitting data to national archives.
3. Audit of archive systems: Archives should be regularly audited to ensure they meet Federal requirements for preserving their data and guaranteeing access.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Agencies should provide ample documentation and support to investigators to help them plan for archiving their data into public archives. Archive centers should work with the investigators to help them plan for capture of their data early in the lifecycle of a project to ensure the burden is minimized and adequate tool support exists. In addition, both peer review of the data and auditing of the archive itself should be performed. Peer review should assess whether the data meet the necessary requirements for compliance against standards and usability for scientific research. Agencies should consider auditing by an independent ISO-approved auditing body that follows a process by which digital repositories can be formally evaluated in terms of their ability to preserve the digitally-encoded information with which they have been entrusted.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Funding should be made available from agencies supporting scientific research that requires analysis of data from public, scientific archives.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Develop citation of data in public archives as a standard, scientific practice. A citation of high quality data should be considered equivalent to the citation of a scientific publication. Public archives should provide support for data citations.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Data standards need to be defined at different levels. Discipline-specific data standards are critical and must be funded as a community effort. These data standards need to be defined and managed as "open standards" that can also allow for international adoption. The NASA Planetary Data System, for example, has developed standards for annotating planetary science data that are used world wide. They are developed as part of a community-wide effort with cross-disciplinary experts from both planetary science and computer science.

Furthermore, standards that support the development and long-term preservation of national archives must also be funded. Examples such as the Open Archive Information System (OAIS) [ISO 14721] are important for defining reusable standards that cross disciplines. These types of discipline independent standards on both the data and computing infrastructure are important to define.

In addition, best practices for the construction of long-term data and archive systems are also important to develop. These should link to agency and national priorities and requirements for preserving data and be used as a basis for auditing the implementation and operations of such systems.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Planetary Data System (PDS) is a good example of a community-driven effort that has been effective for use in capturing and managing the scientific results from NASA solar system missions. The PDS has assembled a data standards design team that includes discipline-specific experts in planetary science who are responsible for the capture and curation of data specific to their area of planetary science. Each of these representatives works with an expert team who designs, implements and operates a set of integrated standards that span the entire discipline. These standards are integrated and published as the basis for capturing planetary science data for long-term archive for assembling, documenting and preserving data in stable formats. NASA Announcements of Opportunity require the use of these standards by investigators who generate

data during a mission or investigation.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Working internationally can be a challenge due to different priorities and limited budgets. However, working in an "open standards" environment is essential. In addition, creating alignment with the international scientific community is a must. The NASA Planetary Data System worked to establish the International Planetary Data Alliance (IPDA) as a body for distributing its data standards internationally. The IPDA has become a vehicle for distributing the data standards. Many agencies have not had the budgets necessary to develop their own standards and therefore have been open to adopting the standards. Rather than distributing the U.S. standards directly from the U.S., the Planetary Data System is working to distribute these through an international organization for which the agencies are stakeholders. This has helped to pave the way toward international adoption. In addition, the IPDA aligned itself with the international science community through the Committee of Space Research that passed a resolution recognizing the effort. The critical goal was ensuring that there is "one" data standards effort for capturing planetary science data archives, rather than several individual and independent efforts.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Data are often not treated as publications. Publications generally have a well-defined structure for data annotation and standard practices for citing data. Data need to be treated as a publication going through a peer review and getting cataloged with well-defined annotations. The registration of data should be done using a standard set of metadata for describing the data. That standard set of metadata should be independent of any one discipline defining a universal common set of data elements/attributes (such as Dublin Core) which can be adopted by agencies implementing data archive systems. In addition to defining a standards set of attributes, a standard scheme should be defined which references the data much like that of a journal paper identifying the authors, title, dates, and location of the data. Researchers that use the data should be required to cite it in their papers as a key reference.

Wed 1/11/2012 12:24 PM
Response to RFI: "Public Access to Digital Data..."

Dear National Science and Technology Council's Interagency Working Group on Digital Data :

I am responding to your RFI as a US federally-funded structural biologist and a member of the International Union of Crystallography Diffraction Data Deposition Working Group and the International Union of Crystallography Commission on Biological Macromolecules.

The International Union of Crystallography Diffraction Data Deposition Working Group is actively working on the issue of archiving the raw data (diffraction images) in crystallography (in addition to summary data and crystallographic models; see http://forums.iucr.org/viewforum.php?f=21).

I give some comments on your requested information items below.

Sincerely,
Tom Terwilliger
Los Alamos National Laboratory

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Long-term Federal support for:
A. International databases that preserve this data (example: The Protein Data Bank) B. Development of international agreements on data and metadata formats C. Development of software and procedures to make it rapid and easy for researchers to deposit their data and to retrieve data from these databases

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?
—-

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

As much as possible, the policies should focus on the desired outcome , not on the mechanism of achieving it.
Policies should be minimized and general. Instead , the focus should be on support for achieving the goals (making deposition and extraction of information easy for the researchers)

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

In the structural biology community there was a most lively discussion on this point on the CCP4 bulletin board (see
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1110&L=CCP4BB&F=&S=&P=323904
for example). The discussion showed that even within a community there is a great difference of opinion on what needs to be archived. It also showed that these costs and benefits can be discussed in a thoughtful way within a community. The CCP4bb discussion revealed some variations of opinion but >50% voted to archive raw data rather than processed data. I expect more will support raw data archiving via local repositories.
It is less clear how this can be done between communities, however.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

The international structural biology community has shown that a research community can work with publishers to develop standardized formats for presentation of data (Rapid structural reports were developed for Protein Science, Acta Crystallographica Section F, and Journal of Structural and Functional Genomics, among others to respond to the need for short reports on macromolecular structures.) This is being extended to working together on coordinating deposition of data with publication.
Research institutions and universities can contribute by providing local repositories for data archiving and retrieval. As some raw datasets can be very large, transferring them via the network can be slow so that local data storage may be most effective.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Repositories (such as the Protein DataBank) that store data in perpetuity and that show a clear benefit for the community should (continue to be) funded at a high level.
Research institutions and universities that store data locally and make it available should also receive funding to help cover the cost of doing this.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

With clear policies and with procedures to walk researchers through any compliance, the burden can be minimized. However it simply will take some effort on the part of researchers to provide their data in a way that others can use.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?
--

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

A basic requirement is that a simple and clear path to those who produced
the data should always be available. The user of a piece of data should
always have an easy way to know who produced it.

Secondarily it would be helpful to have general policies that indicate
that credit should be given. For example a granting agency could require
its grantees to check off a box saying that they will give attribution to primary data producers in their publications that use this primary data.
Third, journals that use Supplementary Materials sections should be expected to include these sections in indexing and citation analyses.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

In structural biology, the mmCIF (macromolecular Crystallographic Information File) is the community-driven standardized data format. This format is used by the PDB to represent and transfer data and it is likely to soon be the standard for communication between researchers as well.
Crucially, the mmCIF file contains metadata about the experiment carried
out. It uses a standardized extendable dictionary of terms. This data
structure and associated software is sufficiently developed so that many aspects of the experiment can be re-analyzed automatically , though it is not yet possible to automatically fully interpret an experiment.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Protein Data Bank (www.pdb.org) developed standards along with the International Union of Crystallography (IUCR).
The key characteristics were (1) international involvement (IUCR) in development of standards (2) a funded central repository (the PDB), (3) a very active structural biology community using the central repository

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

The NIH and NSF should continue strong financial support the PDB, which is a highly effective organization that carries out international coordination of digital data standards for structural biology. Existing nternational committees working on this issue should be brought into the discussion. For example the International Union of Crystallography and The International Council for Science Ad -hoc Strategic Coordinating Committee on Information and Data on these issues (see http://forums.iucr.org/viewtopic.php?f=21&t=63#p175)

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

This is a highly important goal. One practical implementation is the general use of the Digital Object Identifier (http://www.doi.org) as a reference for data items and publications. The structural biology community is considering this for archiving of large data items (TB size) so that they can be stored locally.

James Jacobs
Free Government Information
San Francisco, CA

I am writing as the co-founder of Free Government Information (http://freegovinfo.info) to wholeheartedly endorse the National Digital Stewardship Alliance (NDSA) response to the OSTP RFI on public access to digital data resulting from federally funded scientific research (FR Doc No: 2011-28621), dated January 2, 2012 (http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf).

Best,

James Jacobs
Free Government Information

--
James R. Jacobs
Government Information Librarian
123D Green Library,
Stanford University

"The art of research is the ability to look at the details, and see the passion."
-- Daryl Zero, "The Zero Effect" (1998)

(\
{|||8-
(/

Response to RFI: "Public Access to Digital Data Resulting From Federally Funded Scientific Research" Office of Science and Technology Policy

From: Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan

Contact: George Alter, Director

January 11, 2012

About ICPSR

The Inter-university Consortium for Political and Social Research (ICPSR), a research center in the Institute for Social Research at the University of Michigan, is the world's largest archive of social science data. More than 100,000 users download data from ICPSR every year. Since our creation in 1962, we have expanded to provide quantitative data across all social science disciplines. The Consortium includes more than 700 universities and research organizations located around the world, and we disseminate data for a range of government agencies and other groups, including the Bureau of Justice Statistics, the National Institute on Aging, the Substance Abuse and Mental Health Services Administration, the Bill & Melinda Gates Foundation, and the National Collegiate Athletic Association. Our archive has more than 8000 research collections, some of which include hundreds of datasets. The American Educational Research Association (AERA) and ICPSR are currently working together to encourage broader use of NSF-funded data on education. AERA is offering small grants to young scholars for re-analyzing existing data, and data producers are being assisted in making their data publicly available through ICPSR. The highly regarded ICPSR Summer Program in Quantitative Methods offers more than fifty courses every summer, and almost 900 participants attended in 2011. ICPSR was also one of the founding members of the Data Documentation Initiative (DDI), which has become an international standard for metadata in the social sciences, and we provide the home office for the DDI Alliance.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

ICPSR advocates Federal policies in these areas to improve the access and preservation of scientific data:

1. Require deposit of all scientific data resulting from funded scientific research in an appropriate repository
2. Long-term funding for specialized domain-specific repositories to distribute and preserve data
3. Consistent citation of data in scientific publications
4. Encouragement of standards for data and metadata
5. Including data re-use as a criterion in evaluating research designs

We explain these recommendations briefly here and include additional detail in our responses to subsequent questions.

1. Require deposit of all scientific data resulting from funded scientific research in an appropriate repository

   A general Federal mandate requiring grantees to archive scientific data for secondary analysis would promote re-use of scientific data, maximize the return on investments in data collection, and prevent the loss of thousands of potentially valuable datasets. We have surveyed NSF and NIH grantees in the social sciences to learn what happened to data created on their projects. One quarter of these grantees reported that the data are now lost, and only 14% archived their data at an established repository (Pienta, Gutmann, and Lyle 2009). Our research also shows that sharing data increases scientific productivity: twice as many scientific publications resulted when data were shared (Pienta, Alter, and Lyle 2010; see also Piwowar 2011; Piwowar et al. 2007).

   In our experience, broader access to scientific data is found in research communities that have developed a culture of data sharing. This occurs when leading scientists share their own data, funding agencies commission datasets for general use, and younger scholars can establish their careers analyzing data produced by others. In contrast, domains that condone secrecy create a culture in which researchers seek a competitive advantage by hoarding data and resist scrutiny of their work. Although researchers in these fields sometimes say that they fear being "scooped" with their own data, we consider such concerns unfounded. In our study of NIH and NSF grants, researchers who shared their data had more publications of their own than those who did not share (Pienta, Alter, and Lyle 2010).

   Precedents for a data archiving requirement are available in both the U.S. and abroad. The National Institute of Justice requires archiving of all data resulting from their funding (see http://www.nij.gov/funding/data-resources-program/applying/data-archiving-strategies.htm). In the United Kingdom, grantees

of the Economic and Social Research Council must offer any data resulting from an award to the UK Data Archive (http://www.esrc.ac.uk/funding-and-guidance/guidance/grant-holders/open-access.aspx). The UK Data Archive operates a self-archiving system providing all ESRC award-holders with a way to meet the data archiving requirement. Deposits in this system are reviewed, and high value datasets may receive additional processing to improve accessibility to the research community.

A data archiving requirement does not imply that all data must be preserved in perpetuity. Repositories can offer a limited preservation commitment, perhaps five to ten years depending upon the scientific domain. During this time, datasets can be selected for long-term preservation based upon use by other researchers and judgments of experts in the field.

2. Long-term funding for specialized domain-specific repositories to distribute and preserve scientific data

We advocate the creation of long-lived, sustainable institutions for archiving, preserving, and disseminating data in each specialized scientific domain. Domain specific repositories are needed to solve both technical challenges related to data preservation and re-use and to champion data sharing within their disciplines. For fifty years, ICPSR has been performing these functions for the social science community, and the relatively high level of sharing and re-use of data in our research community would not be possible without the decades of leadership by ICPSR and our peer institutions in the U.S. and abroad. In addition to ICPSR and our peer institutions in the social sciences (see http://www.data-pass.org/), domain repositories exist in a few other domains (e.g., the Protein Data Bank, Dryad), but wide areas of science lack basic long-term infrastructure. New digital repositories need not be free-standing organizations, like ICPSR. They can also be formed within the framework of existing repositories that have strong, long-term institutional commitments. It is essential, however, for these institutions to have governance structures that make them responsible to the communities that they serve.

Domain repositories are needed to mediate between the specific needs of scientific disciplines and the rapidly developing world of digital preservation. The distribution and preservation of digital assets is a complex and rapidly developing area, and each type of scientific data presents its own problems. The requirements for social science data are very different from those for large-scale experiments in the physical sciences. The development of the Data Documentation Initiative (DDI), an XML standard for social science metadata in wide use around the world, is an example of a domain-specific initiative that was promoted primarily by a coalition of domain

repositories. Specialized repositories can monitor and focus attention on issues relevant to their communities.

Our focus on domain repositories is not meant to exclude other institutions from playing a role in distributing and preserving scientific data. We believe that libraries, archives, and other memory institutions have an important part to play, and a number of universities have created institutional repositories for digital objects. These institutions are in a position to provide general services (such as expertise in data management) and personalized assistance to researchers. The main weakness in the institutional repository model is their lack of experience with data. Most institutional repositories developed out of libraries, and their core competence is in the management of digitized text. We believe a partnership between institutional repositories and domain repositories is needed (Green and Gutmann 2007). For this reason, ICPSR has been actively developing ways to work with institutional repositories under a grant from the Institute for Museum and Library Services (see http://www.icpsr.umich.edu/icpsrweb/IR/).

We are very concerned, however, about the role that scientific journals are playing in distributing data within some disciplines. Some journals have a longstanding practice of accepting data as a supplement to published articles. We see several problems with this model. Journal publishers have neither expertise nor financial incentives to redistribute scientific data in forms that will be most useful to the research community. Data are sometimes published in a very limited format like pdf, which is not intended for extraction of numeric data. Publishers also have no obligation to preserve data to provide long-term access for future researchers. Preservation requires accurate and complete documentation and attention to formats, which become obsolete and inoperable. The enormous volume of data being generated in some fields also raises questions about how long publishers will be willing to pay rapidly rising storage costs to make data available.

3. Consistent citation of data in scientific publications

Scientists who create and share data have a right to expect credit for their efforts. Today, merit for academic advancement is measured by citation counts and "impact factors," and the contributions of scientists who create important datasets should be counted. We strongly believe that datasets should be cited in scholarly publications in the same way that other scholarly products are cited. Unfortunately, citation of data in most scientific publications has been incomplete, inconsistent, and unreliable. With our partners in the Data Preservation Alliance for the Social Sciences (Data-PASS), ICPSR has been urging professional associations to adopt and enforce standards for citing data in their journals. The response of these

associations has been positive, and we note that the American Sociological Review revised their guidelines to authors to require citing data in the reference list of every article. Their new guidelines also require a persistent digital identifier, such as a digital object identifier (DOI), which is an important step in facilitating the capture of these citations by indexing services.

4. Encouragement of standards for data and metadata

   Data access is meaningless without documentation (metadata) describing the contents, context, and origin of each digital object. Standards for data and metadata allow developers to create tools for discovery, access, and analysis of shared digital resources. Standards are especially important for long term digital preservation to assure that data will be accessible and comprehensible ten, twenty, or fifty years from now. As mentioned above, ICPSR has been an active participant in the Data Documentation Initiative Alliance, and we are now beginning to realize the benefits of DDI for facilitating data discovery, providing more detailed documentation, and the standardization of access and analytical tools.

5. Including data reuse as a criterion in evaluating research design.

   Federal agencies that support the collection of scientific data can increase access and availability of data for re-analysis and re-purposing by including re-use as a criterion in evaluating research designs in grant and contract proposals. Scientific review panels should be encouraged to consider whether design features (such as the sample size, representativeness, compatibility with earlier studies for meta-analysis) will affect access to data for secondary analysis. For example, samples drawn from one or two locations are much more difficult to share than national samples, because it is much easier to re-identify subjects when the location is known. It is clearly less expensive to collect data in only one location, but the evaluation of a research proposal should consider potential for future analysis of the data. Public-use datasets are much more likely to be re-analyzed than data only available under a data-use agreement. Consequently, the benefit to cost ratio (e.g., publications per dollar invested) may be much higher for a national sample than for a sample based in a single location. In evaluating the overall scientific value of a proposed project, scientific review committees should consider potential for secondary analysis by future researchers.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Scientists who create digital data have a right to expect their contributions to be recognized through citations in publications based on those data. Citation has been the standard way of recognizing original scholarship for hundreds of years. As we noted above, academic careers are measured by citations, and proper citation of data would credit data producers for the impact of their work on science. Citations can also be linked to funding sources (e.g., grant numbers) in ways that can be captured to measure the impact of Federal investments on scientific productivity.

Researchers often desire time to complete their own publications before releasing data to others, and a short delay in the public release of data is consistent with an open data policy. ICPSR sometimes defers the release of data for a limited time (usually 6 to 12 months).

Embedding scientific data in publications is not necessary to make data available to other researchers. Datasets in online repositories are assigned unique persistent digital identifiers, which can be cited in publications. As we argued above, repositories are in a much better position to assure access and preservation of data than publishers.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

We agree that scientific data are becoming more diverse. Important differences include:

- Size. Storage requirements have become problems in some disciplines with massive instrument arrays, video, transactional data.
- Confidentiality. Protecting the privacy of subjects is an essential consideration in biomedical, behavioral, and social research. (See National Research Council 2003 and 2005.)

- Obsolescence. Many types of data remain valuable to researchers for a long time, but in some disciplines improvements in instrumentation make data obsolete in a few years.

We believe that a network of domain-specific repositories would be valuable in creating policies to serve the needs of different disciplines. Repositories in constant contact with their communities are in the best position to understand the unique requirements of their disciplines.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Scientists are not trained in data management, and they often think about the process too narrowly. Few scientists understand the difference between "backup" and "digital preservation," which have very different meanings in the community responsible for digital libraries and repositories. Research communities can benefit greatly from the expertise of librarians and information scientists, and we have noticed the rapid expansion of positions in "data curation" and "data stewardship" in university libraries and research centers. As noted above, we believe that partnerships between organizations with domain-specific and institution-specific mandates are the best way to provide services to diverse and dispersed scientists.

There is a broad need for training in data science to educate stakeholders about the importance of sound data management across the data life cycle and emerging best practices. ICPSR is developing a course on this topic for inclusion in its 2012 Summer Program in Quantitative Methods of Social Research.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The central problem in funding of digital repositories is that preservation requires a long-term commitment and most Federal funding agencies provide only short-term funding. It is not possible for a repository to assure long-term preservation if funding is provided only in the form of short-term grants.

ICPSR, which is celebrating its 50th anniversary, has developed a sustainable business model based on two sources of funding. First, we have a base of 700 member

institutions that pay for access to data. Second, we distribute data under grants and contracts for twenty different Federal and private funding agencies. Data archived with member dues is only available at member institutions, but access to data supported by external sources is usually open. This model works for ICPSR, because we have a large collection of data that is only available to member institutions, and because we have a diversified portfolio of other funding sources. However, ICPSR cannot provide unlimited open distribution to non-members. If a data distribution agreement ends, the long term preservation of that data is supported by the ICPSR membership, and data access is limited to members.

Two changes in Federal funding models would help to sustain access and preservation of digital data. First, in addition to grants and contracts for data distribution, Federal agencies should be able to pay data archives and institutional repositories for long-term preservation. This could involve a single payment for the estimated present value of future distribution and preservation, which repositories could annuitize in some way.

Second, Federal agencies should make commitments to long-term funding of necessary digital repositories. A number of other countries consider data archiving an essential aspect of their research infrastructure and have made long-term commitments to digital repositories for scientific data. A Federal program to establish and support long-lived institutions is needed to create repositories capable of providing preservation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Federal agencies can assure that data from funded research are accessible and preserved by requiring grantees to report a persistent digital identifier pointing to the data in an established digital repository. Most repositories already assign persistent digital identifiers to objects, and these identifiers can be included in citations. Compliance will be easy and inexpensive to verify, because a persistent digital identifier works like a URL pointing directly to a digital object.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

In our opinion, the most important barrier to broader use of data is lack of standardization in data formats and metadata (documentation). By reducing development costs and broadening the range of compatible data sources, standardization will stimulate innovation.

It is particularly important to develop robust, machine-actionable standards for metadata. We are very concerned that inadequate documentation will result in misinterpretation of important policy-relevant data. Modern surveys involve complex "skip patterns" so that respondents only answer relevant questions. For example, married subjects answer different questions than unmarried people. It is very easy to reach incorrect conclusions if the "universe" of each question is not available. Standards for metadata (such as DDI and SDMX) provide ways for data producers to specify background information that is critical to data users.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Assigning proper citations and persistent identifiers to data resources is critical to enabling reuse and verification of data, understanding and tracking the impact of research data, and creating a structure that recognizes and rewards data producers for their contributions to the scientific record. Many data archives and repositories now provide citations that should be used in publications based on the data, and many are also registering persistent identifiers for the data they manage. Data citations permit data to be integrated into the system of scholarly communications and to be picked up by the electronic citation services so that data usage can be tracked.

Federal agencies should be assigning citations and persistent identifiers to the data they distribute across the federal statistical system. This would ensure proper attribution and credit for data producers and would also help agencies track data reuse to better understand the impact of their funding decisions and data programs. Appropriate attribution language that can be easily inserted into manuscripts should be included with all documentation. This language will make giving appropriate credit easier.

Publication authors should acknowledge original data producers by including citations to the data in the references section of their papers. Treating data citations as first-class references provides attribution and recognition of the importance of data as an intellectual product. Journals and other publishers should require data citation and persistent identifiers as part of their submission criteria.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

In the social sciences, many data producers and data archives have converged on the Data Documentation Initiative (DDI) metadata standard – see www.ddialliance.org. Currently expressed in XML, the DDI specification provides a mechanism to document data in a structured, machine-actionable way. This structure enables metadata-driven survey design and processes along the entire life cycle of research data generation.

In the DDI model, metadata needs to be entered only once and then can be referenced and reused later, resulting in greater efficiency. Metadata creation should ideally begin at the conceptualization stage, when survey questions are being designed. Moving this step "upstream" in the data production process leads to greater cost savings for data producers as metadata can be reused.

DDI is being taken up in many countries (see map at http://www.ddialliance.org/community) and by many projects (see a sampling of projects at http://www.ddialliance.org/ddi-at-work/projects), including the National Children's Study and other large-scale efforts.

A Federal commitment to DDI and emerging standards for other types of data would go a long way toward lowering the costs of data management by promoting convergence on these standards and encouraging the development of tools.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Statistical Data and Metadata Exchange (SDMX) standard for aggregate time series is another example of a community-driven standard. Eurostat, the European Central Bank, and other partners have developed the standard, also expressed in XML, to share and exchange data.

Making such standards efforts effective and successful requires a defined community of practice whose members are engaged and invested in the outcome. Seed funding can be very important to these efforts. In the case of DDI, the National Science Foundation provided initial funding that supported meetings of the DDI committee developing the specification and beta-testing. This was key to developing momentum. In 2003 the DDI committee reorganized itself as a self-sustaining membership organization to provide modest ongoing funding for standards development.

Federal agencies might consider investing in standards development with the goal of interoperability.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Initiatives like DDI and SDMX are already international in nature. Federal agencies generating data in related disciplines should become members of these efforts in order to have a say in shaping the standards and in coordinating their development internationally. (The Bureau of Labor Statistics is already an associate member of the DDI Alliance.)

The UNECE High-Level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS) -- http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+Strategic+Developments+in+Business+Architecture+in+Statistics+%28HLG-BAS%29 -- oversees various groups that help to coordinate the development of interoperable metadata across agencies and countries. Federal agencies should encourage and participate in these efforts and the Generic Statistical Business Process Model (GSBPM).

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Federal agencies should craft policies requiring that their data have citations and persistent digital identifiers and that publications based on them use these citations properly. Disciplines tend to develop their own standards for the elements that belong in a data citation, but in general this is a small set of items. Organizations like ICPSR and DataCite can consult in this area.

Once persistent digital identifiers are part of citations and are integrated into the scholarly publication process, it becomes much easier to automate the harvesting of citations for online indexes and to understand the links between data and publications.

ICPSR has a Bibliography of Data-Related Literature that contains over 60,000 citations to publications based on data in the ICPSR data holdings. This permits two-way linking from the publication to the data and from the data to the publications. Most of the work in associating data and publications for the Bibliography has been manual in nature, but greater use of data citations and unique persistent identifiers should make automated harvesting of this information easier.

Agencies could consider providing such linkages for the data they fund. They currently require acknowledgment through grant numbers in publications. Using data citations could become another such requirement. It would also be welcomed if large publication databases like PubMed would integrate links to the underlying data in their systems.

References

Ann G. Green, Myron P. Gutmann. 2007. "Building partnerships among social science researchers, institution-based repositories and domain specific data archives", *OCLC Systems & Services*, Vol. 23 Iss: 1, pp.35 – 53.

National Research Council. 2003. *Protecting participants and facilitating social and behavioral sciences research*. Washington, D.C.: National Academies Press.

National Research Council. 2005. *Expanding access to research data: reconciling risks and opportunities*. Washington, DC: National Academies Press.

Pienta, Amy M., George Alter, and Jared Lyle. 2010. "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data." Presented at the BRICK, DIME, STRIKE Workshop, The Organisation, Economics, and Policy of Scientific Research, Turin, Italy, April 23-24, 2010 (http://hdl.handle.net/2027.42/78307)

Pienta, Amy, Myron Gutmann, & Jared Lyle. 2009. "Research Data in The Social Sciences: How Much is Being Shared?" Research Conference on Research Integrity, Niagara Falls, NY.

H. A. Piwowar. 2011. "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data." PLoS ONE 6: e18657.

H. A. Piwowar, R. S. Day and D. B. Fridsma. 2007. "Sharing Detailed Research Data Is Associated with Increased Citation Rate." PLoS ONE 2: e308.

Timothy Bogart
Wed 1/11/2012 2:39 PM
Federally funded research data

What tax payers pay for is by definition owned by the taxpayer. Legislation limiting access to the data beyond that logically required by legitimate privacy or security concerns is dangerous and and worse.
--
"Don't tell me the sky is the limit when there are footprints on the moon" -- *Sawyer Rosenstein*

Next steps for digital data from federally funded research

Todd Vision,
Associate Professor of Biology
University of North Carolina at Chapel Hill

Heather Piwowar
Postdoctoral Research Associate
DataONE and Duke University

Submitted to the U.S. White House Office of Science and Technology Policy in response to the Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research, and available for redistribution and reuse under the terms of a CC-BY 3.0 License.

Citation: Vision TJ, Piwowar HA (2011) Next steps for digital data from federally funded research. http://bit.ly/sPyZrz

We would first like to commend the recommendations and broad strategic vision outlined by the Interagency Working Group on Digital Data in "Harnessing the Power of Digital Data for Science and Society" (IWGDD 2009). While there are many challenges in realizing the vision articulated in that document, **we think the greatest risk is inaction**. We currently tolerate a high level of wasted investment in data from Federally Funded Scientific Research (FFSR) that cannot be verified or reused, and are paying tangible opportunity costs as a result. Leaving behind us what has been called the "digital dark age" (Kuny 1998) as it applies to research outputs should be one of the top priorities for the US science policy.

We are responding as individual scientists, though we are affiliated with data archiving initiatives in the biosciences, namely the Dryad Digital Repository (http://datadryad.org) and the Data Observation Network for Earth (DataONE, http://dataone.org), and have been active for a number of years in both research and implementation aspects of data archiving. Our response will be primarily concerned with basic research data, which is where we have relevant experience.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Data archiving mandate. Our first recommendation is based on the experience of Dryad, which grew out of a grassroots initiative among a number of biology journal editors to craft a Joint Data Archiving Policy (JDAP) and to ensure that a suitable repository infrastructure existed to support the specifics of the policy before it came into effect (Moore et al. 2010). Thus, a policy mandate acceptable to the community came first and the repository infrastructure to support that specific policy was developed as a result. We recommend a similar approach at the federal agency level, first applying a strong and common data archiving policy for data arising from

basic research funding investments across all agencies, and promoting the development of technical solutions to support the policy as a second step. We offer a template for such a policy as it pertains to data associated with publications:

"This agency requires, as a condition for funding, that data supporting the results in research publications must be archived in an approved public repository. Grantees may elect to have data publicly available either prior to or at the time of publication, or may opt to embargo access to the data for a period up to a year after publication. Exceptions to this policy may be granted if justified in the Data Management Plan for data that meet certain exemption criteria [to be enumerated for each agency or program]."

We feel that a policy of this nature is broadly applicable across agencies and disciplines, which are free to make specific guidelines regarding suitable repositories, what qualifies as an exemption criterion, and the presence/length of the embargo period.

Timely Archiving. It is important that data be archived in a trusted repository at the time the research concludes (in the case of JDAP, at the time of publication), rather than shared upon request after the fact. Multiple studies have shown that disseminating data upon request does not work: researchers or data can't be found, investigators share data selectively with certain colleagues, impose unreasonable conditions on reuse, and are more likely to decline requests when there are quality issues with the data or analysis in question (Campbell 2000, Wicherts et al. 2011). Repositories can offer limited-term embargoes on data release (as discussed above) in order to protect researchers from competitive pressure where this is deemed appropriate. It is important that embargoes not be longer than necessary. We have observed that investigators publish almost all associated papers within two years of archiving their data; in contrast, published reuses of data by third party investigators continue to accumulate for years beyond that timeframe (Piwowar 2011c).

Repository Oversight. To ensure the responsible stewardship of public assets, federal agencies should coordinate policy regarding certification of trusted repositories. This would help ensure that repositories meet agency expectations for preservation processes, metadata standards, governance, financial sustainability, and so on. One lightweight model for such certification is the Data Seal of Approval (http://www.datasealofapproval.org/).

Peer Review for Data. For data associated with publications, an increasing number of journals require that data be made available at the time of peer review (for instance, those published by the Public Library of Science). This is a useful model for funders to promote, in that enlists expert reviewers and editors in ensuring data availability and re-usability. Funders (and research institutions) pay considerable sums for the service of peer review provided by publishers, and so have the right to expect a high level of service from it. The capacity to support secure, anonymous access of peer reviewers to data may be included among the expectations for trusted repositories of publication-related data.

Recognizing the Scientific Impact of Data. The research community need to be confident that publicly archived datasets are valued as first-class scholarly objects by funders and grant reviewers. Specifically, producing a highly valued dataset should contribute more to success in obtaining future funding than producing an insignificant article. We recommend the following:

1. Federal agencies should explicitly encourage the inclusion of publicly archived datasets in the credentials of grant applicants. As an example of current practice, instructions for NSF biosketches mention only that that "patents, copyrights, and software systems developed may be substituted for publications" and that Synergistic Activities may include the "development of databases to support research and education" [GPG Chapter II]. These guidelines inadvertantly imply that datasets are not scholarly products of value.
2. Agencies should systematically collect information on the datasets that have been produced by each grant through the annual and final report mechanisms.
3. Funding agencies should work to promote the infrastructure needed to support impact tracking of datasets (see Question 7). For instance, funders may require the assignment of DataCite IDs (http://datacite.org) as part of the certification criteria for a trusted data repository.

Take Data Management Plans Seriously. We recommend that all federal agencies ensure that data management plans are rigorously reviewed during evaluation grant proposals, and ensure that grant budgets include funds for the execution of the plan. Following the lead of funders such as the Wellcome Trust and the UK Research Councils, US federal agencies should issue a common statement that the costs for curation, preservation and access of research data are integral to the costs of doing research, and thus must be explicitly budgeted.

Filling Repository Gaps. Many disciplines lack appropriate repositories for data, code, mathematical models and other digital research outputs. Research funders should provide seed funding for such infrastructure. Funders should ensure that new infrastructure efforts are not chosen on the basis of technical innovation alone, but will have the capacity to be trustworthy stewards of public assets.

Research For More Effective Research. The effectiveness of data policy and infrastructure must be systematically monitored so that future decisions may be informed by evidence. Federal agencies should issue specific solicitations to researchers to collect the relevant, actionable evidence they need to make such decisions.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

We agree with the response from Cameron Neylon to the "sister" Publications RFI that intellectual property is to be treated as of as a means of incentivizing investments in research rather than as an end in itself (Neylon 2012). FFSR may at times require access to data that is confidential for legitimate commercial reasons (being trade secrets or of relevance to an undisclosed patent), but agencies supporting basic science should not fund the original

acquisition of such data, except - under current law - as it relates to an invention under the terms of the Bayh-Dole Act. Other data are kept confidential for reasons of national security, protection of personal privacy, or protection of sensitive assets (endangered species, cultural artifacts) and may legitimately be produced with federal funds. Protecting the confidentiality of data for commercial exploitation should require significant value-added investment. This is consistent with the position of the International Association of Science, Technical and Medical Publishers in the so-called Brussels Declaration, which states that "raw research data should be made freely available to all researchers" (STM 2007).

In the absence of the above reasons for confidentiality, intellectual property policy should protect the driving incentive for ongoing research, which is the availability of public funds for the conduct of science. **Researchers and universities do not require further IP as incentive to conduct FFSR, as it is already the nature of what they do.** Rather, the continuation of generous public support for FFSR is endangered by policies that allow researchers, universities, publishers or others to place unnecessary restrictions on the exploitation of outputs from public investments in research. Thus, it is in the interests of maintaining a healthly FFSR enterprise, and the corresponding commercial innovation sector that it spawns, that federal agencies ensure restrictions not be imposed where they serve no legitimate public purpose.

Furthermore, since most scientific data, being facts and not creative works, are generally not subject to copyright, a Creative Commons Zero waiver is the most suitable instrument for providing clear and nonrestrictive terms of reuse for data. (See Question 9 for a discussion of rewarding credit to data authors). **Funders should not permit restrictions on commercial use or derivative works for the outputs of FFSR**, as such restrictions stifle innovation without providing incentive for research investment.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Every discipline perceives itself to be unique. However, it is appropriate for federal agencies to articulate strong *general* principles and policies with regard to the management and dissemination of research data, while allowing for discipline-specific implementations that are sensitive to inherent differences such as data volume, machine format, complexity of human curation, long-term value, the applicability of particular metadata standards, etc. In truth, **many of the sociotechnical challenges in data management, standardization and dissemination are shared across disciplines,** particularly for the high-value portion comprising the “long-tail” (Heidorn 2008) of “small science” (Onsrud and Campbell 2007) data associated with publications.

A strong interdisciplinary ‘information community’ (in the parlance of the IDWGG) of data librarians, data scientists and educators should be cultivated. Development of such a workforce should be modeled on exemplar efforts such as the NSF DataNets, the Digital Curation Center in the UK, and the Australian National Data Service. This community is needed to help shape and support general policy and infrastructure within and among agencies, and to help spread data expertise into the educational and research communities.

At the same time, grass-roots 'communities of practice', *sensu* IDWGG, must engage disciplinary scientists in order to determine how to implement general agency policies. Such communities would be in the best position to develop the discipline-specific standards that govern the reporting of data, as well as other research products (*e.g.*, software code).

Individual disciplines and communities may wish to opt-out of general policies (*e.g.*, data archiving). This should be permitted only where the community makes a strong public case that the principles and goals are not applicable to their area, or that the same goals may be effectively achieved in a different way. **Funding agencies are the only stakeholder that can be relied upon to speak for the public interest** in the dissemination of data from FFSR when it is in conflict with the short-term competitive interests of other stakeholders in the research enterprise, and taxpayers expect their government to exercise that responsibility.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

It is frequently impossible to accurately determine the reuse value of a dataset at the time of initial reporting. Many reuses -- indeed, perhaps the most valuable ones -- are for unanticipated applications. Furthermore, we have seen in biomedical data archives that data reuses are not confined to just a few "hot" datasets but spread broadly among them (Piwowar et al. 2011).

Best-practice data archiving is less expensive than many assume. On the basis of our projections for Dryad, the marginal cost of data publication is only a small fraction (< 2%) of the cost of scientific article publication (Beagrie *et al.* 2010a, Vision 2010). For Dryad, it turns out to be much less expensive to accept all the data deposited, and to hold it indefinitely, than to make decisions regarding what to ingest or remove. By comparing the number of published articles generated by a typical grant with that enabled by typical patterns of data reuse, we have found that the modest amount of funding needed to maintain a repository like Dryad is almost certain to generate a **comparatively large scientific return on investment** (Piwowar *et al.* 2011b).

Curation at the time of ingest is a much more significant expense for many repositories than long term storage (Beagrie *et al.* 2010a) and much of the most valuable data (*e.g.,* that associated with publications) is relatively small. For example, the average dataset in Dryad is less than 5MB in size. Furthermore, cost-effective models for the publication of very large datasets are emerging, such as the recently launched BMC journal *GigaScience* (http://www.gigasciencejournal.com).

Finally, the burden of archiving on individual investigators should not be overestimated. Although new practices invariably generate anxiety, Whitlock (2010) and others have demonstrated that **basic guidelines for good data archiving and reuse can be made simple and intuitive**.

(5) How can stakeholders (*e.g.*, research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

While many stakeholders must play a role, we wish to emphasize the crucial role of funders in monitoring the effectiveness of agency policy and individual adherence to data management policy..

Funders should recognize the depth of the need to raise awareness about expectations regarding data management. As part of an ongoing study (Piwowar 2010b), we asked corresponding authors of biology articles about their funders' policies on data archiving: **27% of the investigators responded that they didn't know if their funder had a data archiving policy** (*n*=1500; 39% said their funder had no policy, 10% said their funder required online public archiving). At the same time, consistent with other studies, respondents overwhelmingly believed that mandatory public online data archiving is the "right thing to do." It appears that funders are missing an opportunity to reinforce the best instincts of their funded researchers.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

A variety of funding mechanisms will be needed to provision support for data services (curation, dissemination, migration, replication, *etc.*), given the heterogeneity of all FFSR data. The desired model specifically for long-tail small-science data will (a) provide for some direct investments in repository research and development, (b) scale with the volume of service provided, (c) facilitate the operation of an efficient market for data services, and (d) enable investments in shared international infrastructure.

Investment in repository infrastructure. There will be an ongoing need for direct investment both to support research and development needs of existing repositories and to fuel the development of new resources for datatypes or disciplines lacking existing solutions. When it is necessary for the funding model of data services to be dependent on grants, these should be evaluated based on criteria relevant to infrastructure, rather than solely innovation.

Scalability. Scalability of finances for data services can be achieved by including the costs for data management within research budgets, and allowing individual awardees to direct those costs as needed for their project.

Market for data services. Similarly, if funds are allocated to services on a project-by-project basis, that establishes a competitive market for data services within which those of greatest value receive the most support.

International coordination. Insofar as direct funding from agencies is required for certain datatypes, the greatest challenge will be to develop mechanisms for multinationational investment in shared resources (*e.g.*, such as that used by ELIXIR, http://www.elixir-europe.org/).

While the costs of supporting data infrastructure are tangible, funding agencies should also attempt to understand the hidden economic costs of not having infrastructure to support investments in FFSR data, so that the cost and benefits of investment can be fairly compared.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Evaluating plans and tracking compliance are important. Evidence suggests the NIH requirement for data management plans is generally considered toothless (Tucker 2009) and has made little difference to data availability (Piwowar et al. 2010, Piwowar 2011f).

Disseminating research results is the responsibility of the funded researcher. In the short term, better mechanisms for tying outputs to funding are required. As mentioned in response to Question 1, we recommend that annual reporting require researchers to list publicly available datasets derived from FFSR. Research results that have not been disseminated in accordance with policy should not be acknowledged as output of the grant for the purposes of evaluation.

Federal agencies should enthusiastically collaborate with publishers, libraries, universities, and other stakeholders in promoting technological solutions that will promote trackability of research data products and the reuse of those products, such as DataCite (http://datacite.org), ORCID (http://orcid.org) and VIVO (http://www.vivoweb.org).

In the longer term, there is great potential in moving beyond compliance monitoring to fostering enthusiastic reporting through incentives. The impact of both traditional and non-traditional research products (articles, datasets, code, blogs, preprints, slidedecks, etc.) can be collected for investigators, research groups, institutions, grants, and even whole grant programs using traditional and non-traditional metrics (citations, views, downloads, bookmarks, tweets, etc.). These statistics can then be used to demonstrate the impact of individuals and organizations during evaluations, **providing an incentive for products other than only publications to be reported**. We have been working, with others, on a prototype project to demonstrate this potential (http://total.impact.org); an example showing the "impact report" for one of us (HP), including download metrics for archived datasets, is shown here: http://total-impact.org/report.php?id=SIIysw

Achieving compliance through incentives is currently hampered by our closed scholarly communication infrastructure. Existing citation indexing systems do not index datasets -- even when cited in a paper's reference list (Piwowar 2011d), do not make citation data available for innovative impact mashups, and can not be improved through open source contributions. Barriers to text mining the scientific literature are also significant because the context of a citation contains important information about the nature of the attribution. Future funder initiatives could help address these barriers.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

As we mentioned in Question 7, opening up our scholarly communication infrastructure would make the output of funding more generative - more able to produce innovation (Zittrain 2008).

We recommend licence terms for data or other research outputs that do not exclude commercial and derivative products; this will ensure that outputs from FFSR are available for innovative scientific applications and the creation of new business opportunities. Specifically, **nonrestrictive access to all research outputs** (papers, data, code, etc) would permit machine access, text- and data-mining, data integration, third-party curation, and other value-added services.

We recommend that access and preservation of software from FFSR be given the same policy attention as data. Almost all digital data is collected, and statistics are computed, through the execution of software code. Access to the code associated with a dataset increases the comprehension, re-usability, and replicability of that dataset and its analysis.

The accessibility of the scientific literature is also key to fully leveraging associated datasets. The most valuable piece of metadata about a dataset is the publication that describes its original collection and analysis. When this metadata is not available without restrictions on copying and reuse, it limits the reusability of that dataset.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Citation formatting flavors notwithstanding, the scholarly community has fairly efficient and effective norms for citing published papers. Because community norms for citing datasets have been lacking, investigators have adopted a variety of conventions for providing attribution to the authors of datasets (Weber et al. 2010, Enrique et. al 2010, Piwowar et al 2011e). Few stakeholders provide guidance on data citation (Weber et al. 2010); journals, unsurprisingly, are leading the way whereas funders have provided very little guidance thusfar. This diversity of citation practice makes it difficult to track data reuse.

Nonetheless, **even in the current chaotic environment, investigators receive benefit for archiving data**. Several analyses, in diverse disciplines, have found that studies which make their data publicly available receive more citations than similar studies which keep their data private (e.g. Piwowar et al. 2007). In an survey of 1500 corresponding authors in biology, 45% of authors reported that their datasets have been used and formally cited; only 21% said their datasets had been used without citation (Piwowar 2010b).

Data citations standards, coupled with the impact tools we discussed in Question 7, will make collection and interpretation of data attribution simpler, quicker, and more accurate. Various initiatives are underway to establish and promote standard practices for attributing data. The predominant approach, and our recommendation, is to cite datasets very similarly to how we cite papers. This has the obvious advantages of familiarity and easy integration into scholarly communication tools (we hope!). It also provides a distinct author list to fully recognize the names of the people responsible for the data product. Others are working on “data publications” that combine data archives with a data-centric article wrapper.

Standards for Interoperability, Reuse and Repurposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma *et al.,* 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The array of existing data standards is already very large, and since it evolves with advances in both experimental technology and information technology, it will never be complete. Community interest in the development of a new standard can be stoked by the existence of a critical mass of previously unavailable data with potential for reuse. Thus, funding agencies can indirectly promote new standards by their support for data repositories in particular fields.

It is the role of communities of practice to determine when such standards are applicable to a particular datatype or repository, while it is the role of funders to ensure these are appropriately applied to each project. Rigorous evaluation of data management plans during proposal evaluation can help to ensure this outcome.

It is also the role of the funding agency to ensure that communities of practice are inclusive, transparent, responsibly governed, non-duplicative, and follow best practices in the development of standards. An excellent model of a community of practice working to promote dialog between funders and stakeholders regarding data standards is the Biosharing project (http://www.biosharing.org/standards). Umbrella coordination projects such as this help to rationalize the bewildering diversity of data standards and ultimately help facilitate adoption of appropriate standards by the research community and its repositories.

Two publication outlets for communication of digital data standards in biology include the Open Data Standards section of BMC Research Notes (http://www.biomedcentral.com/1756-0500/3/235/) and the curator collection at PLoS (http://www.ploscollections.org/article/browseIssue.action?issue=info%3Adoi%2F10.1371%2Fissue.pcol.v03.i05). Both outlets, unfortunately, suffer from a relatively low rate of submission.

To help guide public investment in standards efforts, **we recommend federal agencies encourage research into the economic tradeoffs inherent in standard development.** Standards have benefits in ease of data reuse, but also incur costs in development, maintainenance, and compliance. We need to understand better how to balance these costs and benefits.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

See response to Question 10.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

In individual investigator-driven science, international coordination occurs fairly naturally within disciplines provided that funding resources are available for research networks, and that these networks allow international participation. Coordinated calls for funding relevant to digital data standards among agencies in different countries would facilitate the international distribution of effort, and increase the efficiency of investments within each country. One useful model for international cooperation among funding agencies that share a vision of 'making a layer of scholarly and scientific content openly available on the Internet' is the Knowledge Exchange (http://www.knowledge-exchange.info), which includes partners in Denmark, Germany, the UK, and the Netherlands.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Dryad specializes in the linkage between publications (from bioscience and biomedical journals) and associated data, and currently hosts data from over 100 different journals. The model for how to achieve this linkage was developed based on extensive consultation with Dryad's and board and stakeholder community (including journal editors, society officers, publishers, researchers, librarians, and technologists). Some of the elements of Dryad's approach which we think are generally applicable include:

1. Providing a forum for development of shared data policy among journals, including policies regarding embargoes, data citation, terms of use, metadata standards, *etc.*
2. Participation of journal editors, scientific societies, and publishers in governance of the repository and developing its feature road-map.This enables responsiveness to the particulars of journal policies and procedures.
3. Direct involvement of journals, societies and publishers in repository sustainability through payment of membership and deposit fees.
4. Technical integration between article submission and data submission, including direct communication of metadata between repository and publisher.
5. Assignment of DOIs to data through DataCite. Inclusion of article DOIs in online data records and data DOIs within articles.
6. Professional curation to ensure quality metadata, file validity, and preservation actions such as format migration.
7. Exposing metadata through multiple standards, including pushing to third party indexers to enable discovery through standard bibliographic services.
8. Support for secure and anonymous peer review of data associated with articles prior to acceptance.
9. Clear policies regarding data citation, and promotion of technologies to support data citation tracking.
10. Leveraging the preservation infrastructure used by traditional publishers (e.g. CLOCKSS http://www.clockss.org)

References

- Beagrie N, Eakin-Richards L, Vision TJ (2010a). Business Models and Cost Estimation: Dryad Repository Case Study, Proceedings of the 7th International Conference on Preservation of Digital Objects (iPres) 365-370.
- Beagrie N, Lavoie B, Woollard M (2010b), Keeping Research Data Safe 2, Final Report
- Campbell E (2000) Data withholding in academic medicine: characteristics of faculty denied access to research results and biomaterials. Research Policy 29, 303-312.
- Enriquez V, Judson S, Weber N, Allard S, Cook R, Piwowar H, Sandusky R, Vision TJ, Wilson B (2010 Data citation in the wild. Nature Precedings.
- Heidorn, PB (2008) Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2).
- Kuny T (1998) A Digital Dark Ages? Challenges in the Preservation of Electronic Information. International Preservation News, No. 17.
- Michener W, Vieglais D, Vision TJ, Kunze J, Cruse P, Janeé G (2011) DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. D-Lib, doi:10.1045/january2011-michener
- Moore AJ, McPeek MA, Rausher MD, Rieseberg L, Whitlock MC (2010) The need for archiving data in evolutionary biology. J Evol Biol. 23, 659-60.
- National Science Foundation, Grant Proposal Guide, Chapter II
- Neylon C (2012) Response to Request for Information - FR Doc. 2011-2862
- Onsrud, HJ, Campbell J (2007) Big Opportunities in Access to "Small Science" Data, CODATA Data Science Journal 6, OD58-OD66.
- Piwowar H, Chapman W (2010) Public sharing of research datasets: A pilot study of associations. Journal of Informetrics 4, 148-156.
- Piwowar H (2010b). Study on Impact Of Journal Data Policies: Towards understanding the impact of journal data archiving policies on attitudes, experiences, and practices of authors. Recruitement ongoing.
- Piwowar H, Vision TJ, Whitlock M (2011b) Data archiving is a good investment. Nature 473, 285.
- Piwowar H (2011c) A New Task For NSF Reviewers: Recognizing The Value Of Data Reuse. Research Remix blog.
- Piwowar H (2011d) Tracking Dataset Citations Using Common Citation Tracking Tools Doesn't Work. Research Remix blog.
- Piwowar H, Carlson J, Vision TJ (2011e). Beginning to Track 1000 Datasets from Public Repositories into the Published Literature. ASIS&T 2011.
- Piwowar H (2011f) Who shares? Who doesn't? Bibliometric factors associated with open data archiving. PLoS ONE 6(7): e18657.
- STM, (2007) Brussels Declaration on STM Publishing http://www.stm-assoc.org/public_affairs_brussels_declaration.php (STM 2007) http://ftp3.dns-systems.net/~stm/2007_11_01_Brussels_Declaration.pdf?PHPSESSID=5cd58816ea0a9087be865c6cf046626f
- STM (2007) Brussels Declaration on STM Publishing
- Tucker J. (2009) Motivating Subjects: Data Sharing in Cancer Research. PhD Dissertation, Science and Technology Studies, Virginia Tech.
- IWGDD (2009) Harnessing the Power of Digital Data for Science and Society.

- Vision TJ (2010) Open Data and the Social Contract of Scientific Publishing. BioScience, 60(5):330-330.
- Weber N, Piwowar H, Vision TJ (2010) Evaluating Data Citation and Sharing Policies in the Earth Sciences. ASIS&T 2010.
- Whitlock M (2011) Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution, 26 (2): 61-65.
- Wicherts JM, Bakker M, Molenaar D (2011) Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. PLoS ONE 6(11): e26828.
- Zittrain JL (2008). The Future of the Internet And How to Stop It. Yale University Press.

Wed 1/11/2012 5:13 PM
In re: Public Access to Digital Data Resulting From Federally Funded Scientific Research

To whom it may concern,

We are a private biotech company. Lack of access to scientific literature is a severe impediment to our business. Given the extraordinary expenses associated with purchasing publications, we must beg information from academic colleagues. This is particularly frustrating as our tax dollars pay for the research, pay for publication of scientific data and pay private companies for publication reprints. This is a paradigm that serves less than no purpose.

Please register our support for complete open access to any and all research supported in part or in full by institutions such as the NIH and NSF.

Many Thanks for this Opportunity,

Dr. Ray Perkins
President
New Liberty Proteomics Corp.

Association of American Publishers, Inc.
www.publishers.org

455 Massachusetts Avenue, NW 7th Floor
Washington, DC 20001-2777
Telephone: (202) 220-4544
Fax: (202) 347-3690
adler@publishers.org

Allan R. Adler
Vice President, Legal & Governmental Affairs

January 12, 2012

Re: FR Doc. 2011–28621

On behalf of the Professional and Scholarly Publishing Division of the Association of American Publishers ("AAP/PSP"), I am pleased to respond to the Office of Science and Technology Policy's ("OSTP") November 3, 2011 Request for Information ("RFI") regarding "Public Access to Digital Data Resulting from Federally Funded Research."

AAP/PSP members publish the vast majority of materials used in the U.S. by scholars and professionals in science, medicine, technology, business, law, reference, social science and the humanities, and they include the worldwide disseminators, archivists and shapers of the public record on scientific research via print and electronic means. They include non-profit professional societies, commercial publishers and university presses that create books, journals, computer software, databases and electronic products in virtually all areas of human inquiry and activity.

Collectively, AAP/PSP members represent tens of thousands of publishing employees, professional individuals, editors and authors throughout the country who regularly contribute to the advancement of American science, learning, culture and innovation. They comprise the bulk of an $8 billion commercial and non-profit publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade by at least $3.5 billion annually.

For AAP/PSP members that publish scientific journals and other peer-reviewed scholarly publications, the primary goal of their publishing activity is to disseminate information and provide access by providing a high quality and user-friendly digital environment in which to discover, analyze and link to the latest breakthroughs and developments in scientific and other scholarly research. In particular, publishers of scientific journals have, for more than 100 years, played an integral role in building and documenting the unrivalled U.S. scientific research enterprise. In addition to their efforts to disseminate publications that report and analyze the

latest research, they also have considerable experience and investment in digital technology, metadata standards and tools to help users understand and manipulate data. This makes publishers uniquely positioned to help the Federal Government in expand public access to digital data, ensure the long-term stewardship and discoverability of data and support the innovation and economic development that is derived from scholarly advancements.

It is worth mentioning at the outset that, in contrast to peer-reviewed publications, which are not the "result" of federally-funded research, digital data does often directly result from research funded by the government. Research and publication are unique creative acts. Publishers support better discoverability and reuse of scholarly data and are pleased that OSTP has recognized the distinction between data and peer-reviewed publications in this RFI. The dissemination of information is an area of publishers' professional expertise, and data access policies potentially impact AAP/PSP members not only as publishers of peer-reviewed scientific journals, but also as disseminators of information whose innovative products and services enhance and add value to taxpayer-funded research activities and are expected to do so in the future.

It is with this view that the attached comments and recommendations have been submitted on behalf of AAP/PSP in the hope that they will help to facilitate the successful development of sustainable and effective policies on public access to data that are consistent with the Administration's "Open Government" framework[1] and embodies a spirit of collaboration in recognition of the intellectual property rights and private investments of publishers as key stakeholders in these matters.

General Recommendations

Scholarly publishers have long served as integral hubs of the America's research enterprise, validating research through the peer review process, producing a scientific record and facilitating scholarly communication through dissemination and preservation of scientific literature, including the presentation and long-term stewardship of digital data. The primary goal of publishing is to facilitate the widest possible dissemination of the information that publishers create. In the digital age, publishers have invested significantly to enhance the discoverability, public access to and the utility of research data, particularly for the scientific, technological, engineering, social science and medical communities: expanding accessibility,

---

[1] As articulated in Memorandum for the Heads of Executive Departments and Agencies on Transparency and Open Government (January 21, 2009), available at http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment and Memorandum for the Heads of Executive Departments and Agencies on Open Government Directive available at http://www.whitehouse.gov/open/documents/open-government-directive

improving interoperability and fuelling innovation. Publisher investments have created digital platforms with the latest and continually evolving Web capabilities, providing researchers with faster and more robust delivery of scholarly information, new ways to present data and research findings and links that enable information to be found and navigated with ease. Publishers have improved interoperability through new metadata standards and pilot projects, which are driving innovation and providing for better information discovery and expanded use of research results. As long as the government does not diminish incentives for creative publication, publishers will continue to provide tools that enhance innovative reuse and discovery of research information. Publishers look forward to continuing a positive collaboration to enhance science and innovation in the United States, and welcome any partnership with the Administration to harness the power and potential of technology and innovation to spur long-term economic growth and provide cutting-edge solutions to support domestic priorities.

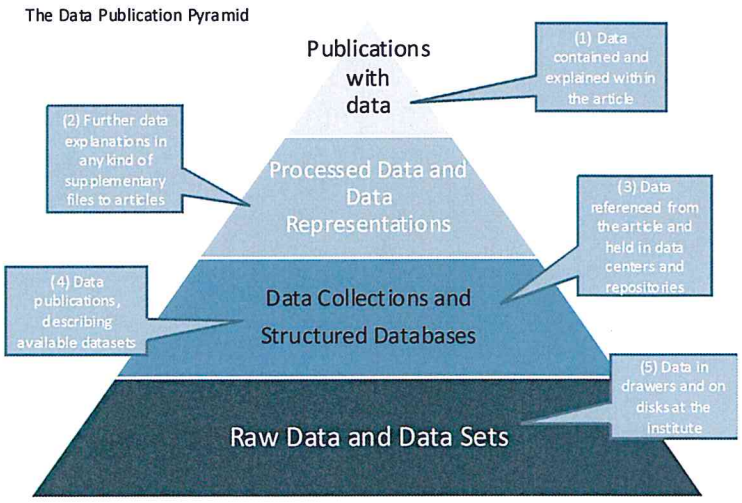
At the same time, the experience of publishers informs an understanding that significant challenges remain to wider public access to and long-term stewardship of digital data. Unlike publications which can be counted by the number of pages, datasets can be measured in mega-, giga-, or even terabytes of data. Unlike publications, which undergo peer-review and editing, extensive formatting and tagging, data files come in a variety of forms, formats and levels of quality and validation. It is difficult to control for quality and consistency. While publication incentives have been embedded in the academic and research process, incentives for complete tagging of individual datasets are limited. There is no consistent approach to presentation, standards for which data should be preserved or overall management and no consistent responsibility for data storage, tagging and dissemination. In addition, the significant costs of storage, distribution bandwidth and overall management and curation must be addressed. These inconsistencies and unknowns must be addressed in a collaborative manner among all stakeholders to promote the development of robust, sustainable and flexible standards that meet the needs of users at all levels. Publishers stand ready to lend their expertise to such a collaborative process to provide value to the research community and to the taxpayer.

The government has a responsibility to provide broad access to the digital data that results from federally-funded research, in contrast to the peer-reviewed publications that contain significant value-added beyond the federal investment. At the same time, the government should not invest funding or energy to recreate what is already being achieved by the private sector. The government's best approach is to leverage public-private collaborations, ensuring the continued innovations in publishing that contribute to the progress of science, allow innovation to flourish and help grow the American economy. A federal role in expanding access to and the preservation of digital data could include partnering with the scholarly community for the identification of standards and best practices for the interoperability of data repositories; creating clear rules for citation, modification and privacy; improving links between data, research grant reports and peer-reviewed publications; facilitating cyber infrastructure

and collaboration within and between federal agencies; and advancing policies and funding to ensure the long-term sustainability of data archives. Public access policies should be developed through voluntary collaborations with nongovernmental stakeholders, including researchers and publishers, university administrators, librarians and the public.

OSTP could learn from initiatives already underway to standardize metadata and provide links between sources of research information. In response to the several questions posed in the RFI asking for best practices or suggested approaches to expanding access, managing data, minimizing compliance costs and other policy questions, AAP/PSP encourages OSTP to review these voluntary projects, develop relationships with groups engaged with the issue and encourage the continued evolution of programs that are working to improve data stewardship and public access to data. These include CrossRef (www.crossref.org), DataCite (www.datacite.org), Opportunities for Data Exchange (www.ode-project.edu), APARSEN (www.alliancepermanentaccess.org/index.php/current-projects/aparsen/) and the NISO/NFAIS Working Group on Supplementary Journal Information (www.niso.org), among others. Such collaborative approaches provide the best way forward towards broad access to and preservation of digital data.

It is critical that the federal government continue to distinguish between data and various types of presentation of data and preserve and respect intellectual property protection and copyright ownership as appropriate. The Data Publications Pyramid displayed here,[2] derived from open science pioneer Jim Gray's e-science pyramid, provides a model for understanding how research data can be presented in a variety of ways with increasing levels of curation and analysis. Federal policies should take into account the differences between information products at different levels of the pyramid and work with all stakeholders, including primary researchers, secondary researchers, publishers, libraries and data centers, to create clear rules and protocols for the



[2] As appearing in the October 17, 2011 *Report on Integration of Data and Publications*, a report of Opportunities for Data Exchange which brings together stakeholders including researchers, publishers, libraries and data centers to support a more connected and integrated scholarly record. Full report available at http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

sharing of data. A collaborative approach will ensure that the needs of each stakeholder group are addressed and that the progress of science is not impeded. In particular, the need to expand incentives for providing broad and timely access to new data must be balanced with the need to preserve incentives for researchers to interpret and analyze their results through curation and peer-reviewed publication.

Rather than imposing an inflexible mandate, federal policies should focus on supporting and encouraging the development of cyber infrastructure, standards for the structure of data and metadata, navigation tools and applications to achieve discoverability and interoperability and ensuring appropriate and sustainable funding for innovation and long-term stewardship. These policies should be developed in collaboration with all key stakeholders involved in the presentation, analysis, deposit, storage and preservation of data.

Responses to specific questions:

Many of the questions posed in the RFI ask for the best approach to specific issues involved in access, interoperability and preservation of digital data. The general response above offers AAP/PSP's recommendation for the government's approach: encouraging OSTP to proceed with a sensible, flexible and careful approach, and to learn from and leverage the experience of successful pilot projects and public-private collaborations. Below I add additional comments in response to questions 2, 5, 8, 9, 11 and 13, where publishers, with their long experience in presenting, showcasing and archiving data, have additional information that may be of use to OSTP.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Intellectual property, copyright and patents are critical features of the scientific and scholarly research world that incentivize innovation and the development of value-added products related to scholarly discovery. Any access policy must comply with current law and provide clear rules for sharing, citation and acknowledgement of any modifications. Federal agencies should recognize that there are different levels of data publication, as described in the Data Publications Pyramid above and craft policies that appropriate protect the value-added in curated collections, processed data and publications. Additionally, appropriate policies, developed in consultation with all stakeholders, should be developed to protect the rights of the creator of data to publish analysis and interpretation of the data in peer-reviewed publishing.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

AAP/PSP agrees with the Interagency Working Group on Digital Data that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." A critical component of any policy needs to be collaboration with researchers, publishers, librarians, universities and research institutions in an interconnected system based on community needs, standards and best practices. Each stakeholder community can contribute their expertise and ensure the creation of data management policies that reflect the different practices of individual research communities.

The involvement of each stakeholder will ensure the preservation of incentives for innovation and help improve information sharing and training within each stakeholder community. Stakeholders can help develop clear standards and guidelines for the availability of research data, certification and auditing of data repositories and metadata standards, which respect each community's standards and practices, working together to create universal policies that work for all communities. Stakeholder input is also important for the integrity of the scholarly record, including the creation of links between datasets and the scholarly publications that analyze and interpret the data. Finally, stakeholder input is necessary to incentivize the deposit of datasets and minimize the administrative burden on researchers, libraries and publishers.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Over the past decade, publishers have invested significantly in providing new services and interactive tools to enable innovation and make information more useful and discoverable. Publishers will continue to innovate in ways to present data to advance science and grow the economy in partnership with researchers and industrial users of data.

The most important step that agencies could take to promote broader use of data is to promote a comprehensive framework for reliable digital data preservation, access and interoperability through the promotion of standards and clear rules developed by the scholarly community. Agencies could also support pilot projects, data curation programs and interpretation initiatives for the relevant scholarly disciplines. Finally, agencies could use their web presence to provide a clearinghouse to the data they hold or which is funded by their grants.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The scholarly community already has a robust attribution and credit system with respect to peer-reviewed publication. This could be leveraged in a bi-directional manner by linking between datasets and publications on the one hand, and a publishing requirement that all data which informs the analysis and conclusions of a peer-reviewed publication be cited according to community standards on the other.

The federal government could help by promoting those standards and provide clear rules for the citation of datasets and acknowledgement of modifications to source data. They should also promote unique and persistent identifiers for data and disambiguate researcher, institution and funder information in metadata. Over the past decade, publishers developed the Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, and similar identifiers are being developed by DataCite[3] for data (www.datacite.org). The work of DataCite, CrossRef and DOE's Data ID Service should be leveraged to ensure data is appropriately archived and recognized as primary research output.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, was developed and adopted through a multi-stakeholder, community-driven approach. It is successful because the standard evolved in response to a real problem in scholarly communication and is providing practical benefits to users of published research.

Digital data standards are newer and still evolving. OSTP should learn from ongoing initiatives that are working to address real problems in collaborative public-private partnerships with stakeholders, such as:

- NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental), which is working to address technical issues surrounding the definition, publication and linking of journal articles and

[3] DataCite is a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, they are currently working primarily with organizations that host data, such as data centers and libraries.

supplemental materials, including data, as well as archiving, preservation and migration of different file formats;

- DataCite (http://datacite.org/), which is working collaboratively address the challenges of making research data visible and accessible;
- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), which is working through a collaborative group of more thirty research institutes, national libraries, IT providers and research funders to create a Network-of-Excellence on digital preservation;
- Opportunities for Data Exchange (ODE, www.ode-project.eu), which is working to promote best practices around the way scientific data are treated;[4]
- PARSE.insight (http://www.parse-insight.eu/), which published a roadmap and recommendations[5] for long-term accessibility and usability of scientific digital information in Europe; and
- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science Unions (ICSU) currently working on an initiative for a World Data System.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

In its October 2011 report, *Federal Engagement in Standards Activities to Address National Priorities: Background and Proposed Policy Recommendations,* the Subcommittee on Standards of the National Science and Technology Council noted that "There was agreement among respondents that the US government should continue to play the role of participant in private sector standards setting processes. There was also general agreement that the effectiveness of government participation depends on the level and consistency of involvement and commitment of resources, both staff and budgetary, to the process. Lack of coordination among agencies...was cited by many respondents as having a negative impact on government effectiveness. " AAP/PSP fully endorses this role for OSTP and the federal agencies.

Strong incentives and guidance for coordinating and aligning policies among federal agencies are critical, but much of the infrastructure is already available to create these linkages. The Digital Object Identifier (DOI) is already used to provide links between publications. With appropriate policies (see question 9) and metadata standardization (see, for example, the CrossRef initiative), the DOI could enable links between publications and associated data.

---

[4] ODE's *Report on Integration of Data and Publications* is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/

[5] The *Insight into Digital Preservation of Research Output* report is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf and the *Science Data Infrastructure Roadmap* is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

As mentioned above, AAP/PSP recommends that agencies coordinate with and support ongoing initiatives to address the technical and practical issues involved in linking data and publications. For example, the NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental) is preparing draft recommendations on how publishers can best attach data as supplementary information to a peer-reviewed publication. As mentioned earlier, DataCite is similarly working to determine best practices for data archiving and metadata standards which would be critical to providing the links needed.

Sincerely,

Allan R Adler

Allan R. Adler
VP, Legal & Governmental Affairs

AMERICAN ASTRONOMICAL SOCIETY

Kevin B. Marvel
Executive Officer

11 January 2012

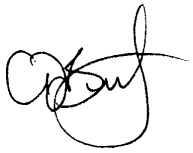
Submitted to digitaldata@ostp.gov

The American Astronomical Society (AAS) appreciates the opportunity to submit comments in response to the Request for Information concerning Public Access to Digital Data Resulting From Federally Funded Research [FR Doc. 2011-28621].

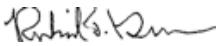
Sincerely yours,

Debra Elmegreen
President, AAS

Kevin B. Marvel
Executive Officer, AAS



Chris Biemesderfer
Director of Publishing, AAS



Richard F. Green
Chair, AAS Publications Board

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

[FR Doc. 2011-28621]

Submission from the American Astronomical Society

The mission of the American Astronomical Society is to enhance and share humanity's scientific understanding of the Universe.

The American Astronomical Society (AAS) is the major association for professional astronomers in the United States, with over 7500 members. One of its primary functions is the publication of the key North American scientific journals dedicated to the dissemination of peer-reviewed research in astronomy and astrophysics, the *Astrophysical Journal* and the *Astronomical Journal*. As a society of research and higher education professionals, we have made a concerted effort to conduct our scholarly publishing enterprise with sensitivity to and balance among the need for prompt and inexpensive access to new results, the pressures on the budgets of technical libraries, and the challenges of obtaining grant and institutional funding to support author fees.

The Society's mission has a broad public purpose, but its constituency is primarily professional research astronomers. Consequently, public access to data, while an attractive desideratum, is less of a concern than is ensuring access to data among research professionals engaged in on-going investigations. However, it is reasonable to assume that the mechanisms for sharing research data among professionals will also serve the needs of interested members of the public, much as is the case for access to the scholarly literature.

Questions from the RFI

Preservation, Discoverability, and Access

1. *What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

   Use of scientific data by the public is a less crucial concern than the accessibility of digital data by other scientists. It is access and re-use by other scientists that will improve the productivity of the American scientific enterprise. Most scientific data themselves are not easy to monetize, so public accessibility follows straightforwardly once data are available to professional researchers. The AAS is in general agreement with the Interagency Working Group on Digital Data (IWGDD) that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." [1] Agency policies should support and encourage a distributed system for both access and preservation. Once community-based repositories are in place and in use by a community, agencies and other entities such as learned societies and journals can insist on deposit of digital data. Deploying mandatory deposit policies in the absence of trustworthy repositories exacerbates challenges in communities already struggling with incompletely coordinated efforts to manage the increasing amount of data being produced. Community-based repositories need to be supported first, and soon.

2. *What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?*

   Policies need to be compliant with applicable copyright and patent laws. Most astronomical facilities guarantee a proprietary period for researchers who collect digital data, and this seems a sensible policy broadly. Astronomy, however, is not biomedicine, so there tend to be fairly few secondary IP issues in our discipline.

3. *How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?*

   We believe that disciplinary differences are real and important and can't be – and shouldn't be – homogenized away. To that end, the most critical thing for the government to do is to be aware of those differences and to respect them. That will require the maintenance of discipline-specific apparatuses for research prioritization, for reviewing research proposals, and for assessing facility and infrastructure effectiveness. This includes any committees and task forces empowered by the government to oversee data management infrastructure.

4. *How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?*

   It is neither practical nor intellectually desirable to keep everything; some data are not worth preserving. So an important element of cost-effectiveness is recognizing this fact, and allowing for the disappearance of insignificant data. Being able to distinguish data that matter from data that are ephemeral is critical: therefore a process of evaluation (by peers, technology experts, etc.) is appropriate. The distinction is not as simple as choosing one type of data over another. The size and complexity of both data sets and any necessary post-processing streams must be taken into account. The availability of high-quality descriptions of data sets is important.

5. *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

   Data management plans (DMPs) need to operate (minimally) at two levels. One level has to arrange for the near-term management of digital data from the current experiment, to ensure its quality and its availability to others investigating the research problem. Another level, a different set of considerations, is needed to address issues of long-term stewardship and preservation. On the question of long-term data management, it would seem that effective DMPs would subject data to a "publishing" process. A publishing perspective would seem quite sensible for considering the curation necessary for the long-term preservation of data sets.

6. *How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?*

   Support for (inevitable) deposit fees must be available for researchers whose data will be deposited in community-based repositories for long-term preservation.

7. *What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?*

   There are two aspects to issues of compliance: the quality of the data repositories themselves as "trustworthy" parties for the management of data sets, and whether or not individual researchers are complying with mandates for data deposit in trustworthy repositories. On the first concern, it seems to be appropriate for the government to allow the academy to manage and maintain these certifications. Organizations exist that either already perform these tasks, or could perform them with nominal broadening of scope and governance (the ICSU World Data System [2], e.g.). We anticipate that the trustworthy repositories will be oriented along disciplinary lines, and will necessarily be international, and for those reasons it makes sense for the US government to use a light touch at most in exercising control over these resources. We presume that individual researchers will be subjected to some level of mandatory deposition of data gathered in the course of federally-funded research. (See our comments to question 4.) The policies employed by the National Institutes of Health (NIH) for ensuring that researchers comply with mandated deposit of articles in the PubMedCentral repository could serve as a model for the policies and procedures that might be used for mandatory data deposit, with the proviso that they may need refinement on a disciplinary, and possibly a repository, basis.

8. *What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?*

   As the IWGDD has noted, there is a lack of a comprehensive framework for long-term data management in most disciplines. In astronomy, there have been efforts to resolve that shortcoming over the years, most recently in the form of "virtual observatories", and these have resulted in fairly effective channels of communication as well as a collection of standards and procedures for managing digital data across wide scales. National governments (not just the US) have a role to play in ensuring the development of the comprehensive framework envisioned by the IWGDD. This should take the form of continued support for efforts in broad disciplinary organization (like the virtual observatories and international alliances), and also support for the creation of trustworthy repositories in appropriate niches in the academy.

9. *What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?*

   In general, citation and attribution of data resources should follow the examples set in the academy for citing articles in the scholarly literature. There is a good deal of discussion in academic circles about the need for data to be regarded as "first-class objects", and it is important for that to happen. The feeling among many astronomers is that the astronomy community is already well down that road, with data set creation being considered (albeit on an ad hoc basis) by tenure and promotion committees. The broader interests of attribution are served in the community by an efficient and well-understood mechanism for data citation, akin to citing the literature. DataCite [3] is an international coalition whose purpose to build a framework for persistent identification of data sets and for the evolution of policies and practices for citing data so that appropriate credit can be assigned to data set "authors".

Standards for Interoperability, Re-Use and Re-Purposing

10. *What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.*

    There are good examples of data formats in many disciplines: in astronomy, the data format of choice is called "FITS", which stands for Flexible Image Transport System [4]. We agree that the purposes stated in the question are important, but rather than trying to name specific standards that are good for those purposes, it might be better to consider the *properties* of the formats, such as FITS, that work well. In our experience, those properties (certainly as they relate to FITS) are that: the standard is community-sourced (defined by the community and governed by on-going community efforts); the standard should be well-documented, and the definitive documentation should be openly and permanently available (the FITS standards are published in the astronomical literature); and the format needs to be widely adopted by both the researchers in the community and the groups in those communities that build the tools for managing and analyzing data. In addition to pure format considerations, for broad data interoperability it is also necessary to have agreed-upon metadata elements and semantics. Metadata semantics should be defined in a way that is nominally independent of specific data formats to permit multiple data formats to co-exist in research niches, and so that data formats can evolve over time.

11. *What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?*

    In addition to FITS, we would also cite the development of the Digital Object Identifier (DOI) [5] (mostly by the publishing industry) for persistent digital object identification, and the creation of the Dublin Core [6] (spearheaded by the library community) for core metadata semantics. Those standards by and large have the properties we described in our response to question 10.

12. *How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?*

    The committees and task forces that are assigned to oversee data management infrastructure (question 3) should be charged with maintaining awareness of standards efforts, and for participating in appropriate forums for standards development. The government's committees have to be well-enough informed so that the government can (credibly) endorse effective international programs.

13. *What policies, practices, and standards are needed to support linking between publications and associated data?*

    Conventions and mechanisms for these purposes are being investigated in the academy today. The most prominent coalition is DataCite (question 9); the AAS supports and participates in DataCite through an alliance with the California Digital Library. The technological standard being proposed by DataCite to support linking is the persistent

identifier, of which the DOI is an important example because of its use in the scholarly literature for these same purposes.

References

1. IWGDD report, *Harnessing the Power of Digital Data for Science and Society*, January 2009, http://www.nitrd.gov/About/Harnessing_Power_Web.pdf
2. http://www.icsu-wds.org/
3. http://www.datacite.org/
4. Wells, D. C., Greisen, E. W., and Harten, R. H. 1981, *Astronomy and Astrophysics Supplement*, vol.44, p.363.
5. http://www.doi.org/
6. http://dublincore.org/

January 11, 2012

White House Office of Science and Technology Policy
Comments on RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research Research.

Response from Arizona State University Libraries

There are four different areas in which the Federal government can make substantive contributions in the effort to ensure that publically financed research data are made widely available, and are preserved and curated for the long-term: technical (build robust, sustainable technical infrastructure); standards (establish national as well as discipline-specific standards for data and metadata); sustainability (determine what services should be supported as a basic part of a university research environment and what services need to be offered on a cost-recovery basis); and governance (ensure that solutions and initiatives undertaken by various agencies, universities and other stakeholder groups are coordinated and communicated widely, and that issues are prioritized and solved in ways that benefit the majority of stakeholders).

At present, most if not all research universities are struggling with similar issues: how to preserve research data and make it accessible for the long-term; how to support increasingly complex E-science research projects; and how to build and maintain a sustainable cyberinfrastructure in times of economic downturn and budget cuts. Universities have an important role to play in addressing these issues, as do the professional societies, the Federal government and private industry. While having grassroots solutions developed by several hundred universities has advantages, the concern is that the results will be diverse, incompatible solutions for the same problems. The Federal government can play a much needed role in coordinating the individual efforts of all stakeholders thus ensuring that the results are complementary, interoperable, and communicated broadly.

Preservation, Discoverability, and Access

1. *What specific Federal policies would encourage public access to, and the preservation of, broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

A necessary first step in preserving digital data, and making it accessible for the long-term, is to mandate that publically funded research data be deposited in open access or publicly accessible digital repositories, unless confidentiality or privacy concerns prevent this. Since Federally funded research data has been paid for by the public, it should be accessible to the public in as timely a manner as possible. Requests for short embargo periods (e.g., 6 months to 3 years) could be accommodated but recent experience with genomics data indicates that when research results are made public quickly, scientific advances proceed apace and new commercial applications and product development rapidly results (Williams 2010). The current norm, that individual researchers post their data on personal web sites, or respond to individual requests for

copies of their research data, is risky (from a preservation standpoint) and not sustainable over the long-term (what happens when the researcher retires?).

The second critical issue is the need for open access and not-for-profit digital repositories to be interoperable, perhaps linked by a Federal portal. Here, the important Federal role is to coordinate efforts across the various scientific disciplines to establish standard data and metadata formats, and mandate their use. At the very least, there should be a national clearinghouse for best practices for research data (data formats, metadata, data management using a lifecycle approach) from the various fields.

2. *What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?*

Balancing access and intellectual property rights in the digital age is complex. Texts and other digitally published works have the option to include license or access restrictions by utilizing a standardized infrastructure, such as that provided by Creative Commons. However, there is widespread concern that a Creative Commons model for research data is problematic, in part because facts cannot be copyright protected and in part because data – unlike a published text – is mutable (see, for example, de Cock Buning et al. 2009). A useful Federal role would be to coordinate existing efforts, like those of the Open Knowledge Foundation, to help establish best practices for public data access and licensing.

Intellectual property issues notwithstanding, a mandate that data be deposited in an appropriate digital repository is essential to ensure curation, preservation and to facilitate public access. Limited embargo periods, as mentioned in question #1, can be accommodated by digital repositories with the added benefit that the data are stored securely and preserved during the embargo period.

3. *How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?*

An existing Federal organizational structure is in place as a result of the disciplinary focused granting agencies (NSF directorates, NEH, NIH). Working with the professional and learned societies, this structure can be used to help ensure that disciplinary differences are recognized and accounted for. However, what is missing is a national coordinating body to facilitate communication among the various disciplines and representatives from the Federal funding agencies. Cross-disciplinary standards should still be a priority to help ensure that data are preserved and accessible for the long-term. These include minimum standards for descriptive and technical metadata, file formats, the use of disciplinary specific controlled vocabularies, and permanent digital identifiers (DOIs).

4. *How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?*

Ultimately, assessing the significance of research data sets is a task that should fall to the researcher, the repository, and be informed by standards and practices of the various professional societies. Federal agencies might usefully work with stakeholder groups to establish guidelines for evaluating research materials, and suggest best practices for archiving and retention schedules, similar to document retention guidelines for public documents and business records.

It is important to recognize that even after scientific data are no longer current, they retain important historical value. University institutional and digital repositories generally take the approach that all deposited materials should be preserved for the long-term. Like other cultural memory institutions (e.g., museums, libraries and archives) University digital repositories hold the scholarly and creative output from that institution and assume a stewardship role that lasts in perpetuity.

5. *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

Since January 18, 2011, when NSF required data management plans for grant proposals in all directorates, followed by a similar announcement from the NEH Office of Digital Humanities (June, 2011), research universities throughout the country have expended considerable effort in providing support and best practice information to faculty and researchers. As noted in the second paragraph on page one, a diversity of results from individual efforts is valuable, but more valuable is the active coordination among institutions so that successes and lessons learned can be shared. The Federal government can play a much needed role in coordinating the individual efforts of all stakeholders thus ensuring that results from the various efforts are complementary and communicated broadly.

Data management planning, throughout the active research period of the grant-funded project and beyond, involves a partnership between data producers, repositories and Federal funding agencies. The "stick," i.e., that a data management plan is required by funders and evaluated as part of the proposal review process, is certainly important. However, digital repositories need to develop services and mechanisms that researchers consider valuable, to serve as a "carrot" that encourages researchers to describe their data with adequate metadata and deposit it in sustainable formats in publically accessible repositories. An example of a basic repository service would be to assign permanent, unique identifiers (DOIs) so that data sets can be cited unambiguously. Tracking citations, as well as data set views and downloads, are value added services because they provide useful measures of research impact, measures that can be very important in tenure and promotion decisions.

6. *How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?*

Digital repositories are expensive, with costs primarily falling into three categories: repository staff salaries, short-term file access and storage, and long-term data curation and preservation services (Goldstein and Ratliff 2010; Rumsey 2010). Universities can reasonably view repositories as a new and essential component of the research infrastructure, on a par with libraries, museum collections, and university archives. As such, basic funding should come, at least in part, from grant indirect costs. However, as with other research costs that exceed the basic campus infrastructure, some digital curation and preservation costs should be eligible for inclusion as direct costs, particularly for projects where large amounts of data are generated (e.g., terabytes or petabytes) and/or where the data are complex and difficult to archive. Currently, there is no clear understanding of how to establish the line between what can be managed as a part of the "basic" university research infrastructure, and what exceeds that threshold. We need a national referendum where stakeholders with experience in this area can discuss the costs of preservation and possible solutions for establishing sustainable repositories.

7. *What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?*

Adopting a research data lifecycle approach, such as outlined by the Digital Curation Centre in the U.K., is helpful. At a minimum, there are three points in the lifecycle at which compliance could be systematically evaluated: grant proposal review (data management plan), active grant research phase (data production and storage), and data publication at the end of the research project (deposit in a public access repository).

Grant reviewers hold researchers accountable for proposing a reasonable and feasible data management plan as a normal part of the review process. Agencies could provide guidelines to proposal reviewers enumerating the essential elements of a well-developed data management plan tailored to disciplinary best practices (e.g., NSF could develop such reviewer guidelines by directorate).
University sponsored program offices typically monitor researcher compliance with agency mandates during the active research phase, though this is typically limited to audits of financial expenditures. A useful Federal role might be to coordinate stakeholder meetings to develop systematic and automated workflows that could become part of the University's monitoring process. Minimally this might involve requiring project principal investigators to complete an annual or semi-annual compliance form, stating that they have implemented the steps they outlined in their data management plan.

Finally, if research data are deposited in public, open access repositories, agencies as well as university sponsored program offices could systematically verify that appropriate files had been deposited and were accessible. Linking repository based research data sets with publications, something the ecological sciences achieve using the Dryad repository, is one approach that has worked well (Beagrie et al. 2009). A necessary activity, however, is to provide opportunities for disciplinary stakeholders to discuss these compliance measures, thereby ensuring that sufficient attention is given to differences between scientific disciplines and different types of digital data.

8. *What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?*

One of the primary goals of the America COMPETES Reauthorization Act of 2010 (Gonzalez et al. 2010) is to improve education in science, technology, engineering, and mathematics (STEM) in support of innovative research in the physical sciences and engineering. If we mandate that publically funded research data be deposited in open access or publicly accessible digital repositories, we build a valuable reservoir of educational materials that are easily available to K-12 teachers as well as college and university instructors. The next step is to encourage the widespread use of these resources, by encouraging teachers and students to explore repositories and incorporate digital collections into the curriculum. Purdue University has successfully pioneered this approach using their hubZERO digital repository platform (Magana 2010).

In partnership with researchers, new repository services could then be developed, such as providing sample homework assignments, curriculum modules, and learning objects for a variety of grade levels. This would help satisfy the broader impacts expectation mandated in NSF funding proposals.

9. *What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?*

As mentioned in our answer to question #5, repositories can build services (such as assigning DOIs) so that data sets can be cited unambiguously. At present scholarly disciplines have well – developed procedures for citing published articles, books and other texts. Coordinating disciplinary referenda with the professional societies, universities, federal agencies and other stakeholders would allow for the development of data citation systems that are appropriate for the various scientific fields.

Standards for Interoperability, Re-Use and Re-Purposing

10. *What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.*

With the growth of disciplinary repositories, and data repositories linked to journal publications, the various scientific fields have begun an active engagement with topic of metadata and other standards. As with other digital data management efforts, however, these tend to be conducted in isolation, without as much national or international input as needed. Coordination by agencies such as the Library of Congress and the National Archives and Records Administration would be helpful in enhancing communication and reducing duplication of effort.

11. What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

During the 1990s the Federal Geographic Data Committee (FGDC) successfully established the Content Standard for Digital Geospatial Metadata, building on nearly two decades of prior work by various federal agencies involved in geographic information systems. Several factors led to the widespread adoption of this standard, including early and persistent engagement by a variety of federal and industry stakeholders, simple tools incorporated into GIS software by commercial vendors that allowed users to easily enter metadata, and additional software tools that updated metadata automatically in response to file changes. The FGDC standard is currently being revised to a North American Profile, allowing for important updates and changes to be made to keep pace with changes within both the GIS and technological realms.

12. How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

The European Union has been quite successful in establishing standards among EU nations that result in applications that assemble disparate resources into a single web platform, easily accessible to users from around the world. For example, Europeana is a web portal that brings together digital cultural material from major European art museums, libraries, and other cultural heritage institutions. To leverage this experience, the U.S. should seek to establish working groups with European and Australian cyberinfrastructure partners. Increasingly, NSF has been coordinating grant programs with JISC (the Joint Information Systems Committee of the UK) and other others. Similarly, the recent National Endowment for the Humanities "Digging Into Data" challenge brought together a large group of international research funding agencies, representing Canada, the Netherlands, the United Kingdom, and the United States. Coordination at this level would seem to be helpful in fostering information exchange while still remaining sensitive to disciplinary differences and challenges.

13. What policies, practices, and standards are needed to support linking between publications and associated data?

As noted in our answer for question #7, a number of disciplinary repositories have successfully established links between repository data files (and other associated file types) and published journal articles, books and reports (e.g., Dryad). It is helpful if disciplinary repositories seek out partnerships with appropriate publishers and professional societies. There are mutual benefits from this kind of commercial / repository profit partnership. For example, tDAR (the Digital Archaeological Record), a not-for-profit repository for digital archaeological data, is able to link disparate information about an archaeological site, a research topic or a geographic area, by including metadata from commercial publishing firms with the metadata and documents in its repository (McManamon and Kintigh 2010). Publishers gain an inexpensive and easy way of advertizing their publications. Repositories gain additional digital resources that they can make available to users. The overall benefit is that available information is made more easily

discoverable, accessible, and usable. Users gain a "one-stop-shopping" experience that increases accessibility and expands the number of relevant search results for users.

These comments are submitted on behalf of Arizona State University Libraries by:

Mary Whelan
Geospatial Data Manager

Sherrie Schmidt
University Librarian

References Cited

Beagrie, Neil, Lorraine Eakin-Richards, and Todd Vision
2009. "Business Models and Cost Estimation: Dryad Repository Case Study." Society 1:1-6. Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf.

de Cock Buning, Madeleine, Allard Ringnalda and Tina van der Linden
2009. The Legal Status of Raw Data: a Guide for Research Practice. Report of the Centre for Intellectual Property Law, the Netherlands. Retrieved from www.surf.nl/publicaties.

Goldstein, Serge J. and Mark Ratliff
2010. DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data. Report for the Office of Information Technology, Princeton University. Retrieved from: http://arks.princeton.edu/ark:/88435/dsp01w6634361k.

Gonzalez, Heather B., John F. Sargent Jr., and Patricia Moloney Figliola
2010. America COMPETES Reauthorization Act of 2010 (H.R. 5116) and the America COMPETES Act (P.L. 110-69): Selected Policy Issues. Congressional Research Service Report for Congress. Retrieved from: http://www.ift.org/public-policy-and-regulations/~/media/Public%20Policy/0728AmericaCompetesAct.pdf.

Magana, Alejandra J.
2010. "How Engineering Instructors Use NanoHUB Simulations as Learning Tools?" Retrieved from: http://nanohub.org/resources/8742 .

McManamon, Francis P. and Keith W. Kintigh
2010. "Digital Antiquity: Transforming Archaeological Data into Knowledge." SAA Archaeological Record 10(2):37-40.

Rumsey, Abby Smith (editor)
2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Retrieved from: http://brtf.sdsc.edu/ .

Williams, Heidi L.
2010. Intellectual Property Rights and Innovation: Evidence from the Human Genome. National Bureau of Economic Research, Working Paper Series, Working Paper 16213. Retrieved from: http://t.co/9OYy9DjO.

TO: Office of Science and Technology Policy publicaccess@ostp.gov
FROM: Daniel Lee, Tucson, Arizona
RE: Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research
DATE: Wednesday, January 11, 2012

The following comments are in response to the request for information issued November 4, 2011, by the Office of Science and Technology Policy (OSTP) regarding recommendations on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research. I would like to thank OSTP for the opportunity to respond and contribute to the conversation. My comments follow.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

- Federal policies that require making appropriate digital data resulting from federally funded scientific research openly available in repositories committed to preservation and access would be a significant stimulant to the American scientific enterprise and to the economy as a whole. (Classified data and data that include personally identifiable information would certainly be inappropriate for inclusion in such policies.)
- On the one hand, such policies would allow for verification of findings and analyses by outside researchers. Verification and reproducibility are the very heart of the scientific enterprise. By facilitating this activity funding agencies would be contributing to the credibility of the funded research and thus promoting further progress by researchers who build on these results.
- On the other hand, in many cases these same policies would expedite further research by saving subsequent researchers from recreating the data that was already collected and produced. There is clear potential for saving both funds and time, thus allowing research funding to go further and accomplish more.
- Open data policies would also have the added benefit of creating resources where businesses large and small would have information available to them to use as they see fit to create new products, services, and markets that can drive economic growth. Further, by providing broad, ongoing access to data, a wide range of research could be promoted including interdisciplinary projects apart from the expected uses intended by the initial researchers.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

- First, it is important to recognize that copyright rests with the author/creator unless and until s/he transfers that right to another party such as a publisher. A research funder could reasonably require authors and creators to adjust these rights in some way or other. Research funders that are the author's employer might indeed be the rights holder themselves and thus impose even greater control over how rights are managed. Given that scientist/authors are often more interested in spreading results (while getting credit) and having impact than in controlling rights, protecting the intellectual property of publishers doesn't seem like a helpful place to start.

- This isn't to say that publishers don't add value and that their investment in that value doesn't need protection of some sort. They do and it does. If publishers invest in hosting and providing access to the data that supports the papers they publish, that investment does deserve protection. One form this protection might take is treating the data as part of the publication that makes up the version of record and is cited as such. The formal recognition that comes from citation and the granting of authority that comes with it is indeed a form of protection that drives future business their way.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

- One important way Federal agencies could account for disciplinary differences in regards to what data is usefully shared and how it should be presented is to build flexibility into an agency's policy that relies on the program directors and reviewers for each program area to largely define what data should be shared, in what time frame it should be shared, the mode it should be shared in, and where it should be shared within a broad mandate for making relevant data openly available.
- It is also important to recognize, though, that while allowing for differences in data types there also needs to be sufficient commonality to allow for and promote cross-disciplinary discovery and reuse. Much of the advantage of open data is creating the possibility of discovering data sets that were collected or created for one project in one field that serves useful for solving problems in a separate field unexpected and unintended by the original researcher.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

- As in comment #3, the value inherent in differences in costs and benefits of long-term stewardship and dissemination will depend on differences of disciplinary needs. Agency policies should allow for those differences and allow for those in the research areas to help define those needs.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

- Research communities can make a huge contribution to the implementation of data management plans by reaching consensus within each community on what makes up a high quality plan. Universities, research institutions, and their libraries can assist the communities in developing these best practices and share successful approaches in centralized web site linked to other research compliance support.
- Among the topics that would likely be included in support materials are guidelines for depositing the resulting files in the institutional repository where appropriate, promoting consideration of the issues around sharing data early on in project development to save time later having to re-work data into usable formats and forms at the end of the project, and advice on useful, practicable metadata templates and standards.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

- Once funding agencies begin requiring preserving and making digital data accessible grant announcements should clearly indicate that the costs required to achieve these ends are expected to be included in the budgets as part of proposals and awards. Working with the constituencies, agencies can also promote efforts to create best practices in these areas that will help researchers understand and define the real costs. Among the issues to be addressed is the need to develop business models for one-time payment out of grants that account for the ongoing costs of continued, persistent access.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

- This is a very important goal. Simplifying processes that funded scientists and their institutions go through to meet agency requirements for accountability would lower overhead costs and allow for more research to occur. Researchers want to do research and share the results, not satisfy bureaucracies.

- Simple procedures that fit into an existing workflow have the best chance of achieving desired ends with minimal additional burdens.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

- Besides requiring data resulting for federally funded research, agencies could stimulate innovative use of research data by facilitating standardization of the infrastructure that supports data modeling and data management. Within the context of differences of data types, such standardization promotes findability and reuse, thus allowing entrepreneurs eased access to research findings to create new products, services and markets and to enhance existing ones.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

- Cultures of citation and other forms of recognition develop and are enforced within disciplines. The greatest contribution agencies and others outside the specific fields could make would be to create and encourage standardized metadata schemas and templates that clearly indicate the responsible parties and that delineate the various roles in gathering and producing the data.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort

- I am confident other submitters will be more up to date and exhaustive than I can be on this issue.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

- The example of the Federal Geographic Data Committee (FGDC) is a model worth considering in other research areas. As a major factor in this research community, the FGDC was able to reach agreement on a standard that then was adopted by smaller venues in the field.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

- Federal agencies could promote effective coordination on digital data standards with other nations and international communities by working closely with equivalent agencies and other scientific organizations as full participants in ISO (http://www.iso.org/iso/home.htm) standards development.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

- This is a key issue to confront if we are to take full advantage of reusable data. Agencies could support linkage by requiring funded researchers to assign persistent unique identifiers to each data set resulting from the funded project and referring to the data through the unique identifier in all resulting publications. The California Digital Library has created one such tool call EZID (http://www.cdlib.org/uc3/ezid/). Such an identifier would link publications (ideally using Digital Object Identifiers) with uniquely identified authors using services such as ORCID (http://www.orcid.org/) with the data using EZID. Whether the data is archived by a publisher, a university, or a funding agencies wouldn't matter as long as the unique identifier stays with the associated file.

While the advent of data sharing plan submission requirements at the NIH and the NSF is a welcome development, encouraging the reuse of scientific data needs far more policy intervention.

First, Standards should be developed that can be used to grade data sharing plans, so that grant review panels can know both whether or not a specific data sharing plan is satisfactory and so that for any given call for submissions the reviewers have a sense of how important data sharing is versus the scientific goals of the project. Second, data sharing plans should be made public alongside the notices of awards and contact information for the principal investigators, so that both taxpayers and scientists know what promises were made and how to contact a scientist and ask for data under the plan approved.

Third, tracking should be possible to begin to estimate compliance: annual grant review forms should contain fields where the researcher is obliged to place URLs to data shared under the plan (or if left blank, explain why), for example. It should also be easy to create a data request system in which those asking for data send a copy of their request to the grants database, which can then be cross-referenced against the review forms to provide at least a rough estimate of compliance. And fourth, scientists with a record of subpar execution against data sharing plans should be downgraded in their applications for new funding. Taken together, these four elements create an incentive structure that would significantly increase the incentive for scientists to provide public access to the digital data resulting from federally funded research.

In tandem, the funding agencies might develop financial models for the preservation of these digital data in much the same way that models exist for estimating overhead and other baseline costs as a percentage of the grant. This could fund not only new library services and jobs in the research enterprise but also serve as a non dilutive funding source for a new breed of data science startup companies focused on preservation, governance, querying, integration, and access to digital data.

However, we should be careful not to treat data as property by default. Intellectual property is a useful frame through which to view creative works and inventions in science, as well as to protect valuable “marks” and secrets. But in the United States at least, data is typically in the public domain already, and therefore the extension of intellectual property rights to it would represent a vast expansion of rights in a space where there is zero empirical evidence that it is needed.

Typically data is treated more as a secret, which is at odds with the public nature of the idea of data access, and the obstacles to data sharing are less legal than they are professional and economic. The ugly reality is that sharing data represents a net economic loss in the eyes of many researchers: it takes time and effort to make the data useful to third parties (through annotation and metadata) and that is time that could be spent exploiting the data to make new discoveries. On top of this, there is a twin incentive problem. Scientists see no benefit to sharing data and are not punished if they fail to share data, while there is a pervasive fear that other scientists will “scoop” them if their data are available before being fully explored. This creates a collective action problem that can be overcome most easily by clear funder policy as enumerated above: data sharing plan mandates with transparency, accountability, tracking, and impact on future funding.

One policy action that would be very welcome would be an unambiguous signal that publicly funded science data is in the public domain worldwide, not just in the United States. This could be accomplished either through the use of a copyright waiver, such as the Creative Commons Zero tool, or through other means. But it is vital to make it unambiguous and clear when and where data are free to reuse, because applying conditions imported from creative works and inventions to a class of information that is fundamentally far less like "property" can have serious unintended consequences. Easily imaginable consequences include vast cascades of attribution requirements, so that a query to 40,000 data sets requires 40,000 attributions – every time – or worse, the poisoning of data for use in job creation by small companies who wish to build atop data as a platform or infrastructure.

The intellectual property status of data does differ across the scholarly disciplines and its own status in how far it's been processed. Some sciences rely on inherently copyrightable "containers" for data, from field books to recordings to photographs. And raw data converted to beautiful information by visualizations will touch on copyright. Policy should be flexible enough to account for this, but start with a default bias that public domain data is the most reusable, while providing "opt-out" capacity for data and disciplines where the public domain is simply not the best solution.

There is an obvious problem with this set of policy recommendations. They rely on money to work. We do not yet know the true costs of storing digital data over the same time frames that we store the scholarly literature. As our capacity to generate data explodes, we must invest at the same time in our capacity to steward it. Research projects into large data information science should be a priority, with specific attention paid to when and where it is possible to compress data, move data to secure "cold storage", jettison data (either because it is duplicative, or because it can be regenerated again later) , and more. We do not have the sociotechnical infrastructure required to answer questions of data stewardship with any authority, and we must create it on the fly at the same moment that the data creation burden is hitting exponential heights.

Solving these stewardship problems might be best achieved through a coalition of research institutions, the library community, publishers, and funders. Taken together these groups already heavily regulate the daily life of a federally funded scientist. It is a small extension to imagine leveraging that regulatory power to provide new services to the scientist – a university and its library might keep an archive of standard data sharing plans, standard budget items to implement, which together would take the guesswork out of filing and operating a data sharing plan. Even better would be a federal program to certify a small number of such plans for each discipline.

Missing from the set of stakeholders mentioned in the RFI is, notably, the business community, both the large scientific companies and the vast potential of startup firms. In an ideal world, the stewardship conversation will bring in actors from those industries, from pharma to venture capital, as we are missing an entire professional class of data stewards and data engineers (not just data scientists) who could serve the needs of the research enterprise while creating stable. Even better, because the data stewards must be close to the researchers to serve them, these jobs are less likely to move offshore. An investment in small business grants, job training (and retraining) vouchers, and the creation of

community college pedagogy for data stewardship functions could go a long way towards stimulating the emergence of this professional class.

In order to stimulate the interaction among these stakeholders and the emergence of a new class of data stewardship jobs, agencies could take additional steps to stimulate use of data. Contests are one obvious route, where a prize is posted in return for solving a problem (or simply for coming up with innovative ideas and/or applications that run on government data). Another route is the expansion of SBIR grants to create a track focused specifically on data startups, which lower the risk of company formation and job creation as well as creating non-dilutive funding sources for entrepreneurs.

A route that is vital, but less obvious, is investment in and commitment to the emergence of standards that enable interoperability of, and thus reuse of, digital data. Standards lie at the heart of the Internet and the World Wide Web, and together lower the cost of failure to such a low point that companies built on the web and the internet can begin in garages. Such is not the case in the sciences. And it will not spontaneously emerge, even if data flow onto the web. As long as those data are in a tower of babel of formats, incoherent names, and might move about every day, they will be a slippery surface on which to build value and create jobs. Federal policy could call for a standard method for providing names and descriptions both for digital data and for the entities represented in digital data, like the proposed standard of the Shared Names project at http://sharedname.org .

Standards also make it far easier to provide credit back to scientists who make data available, as well as increasing the odds that a user gets enough value from data to decide to give credit back. Embracing a standard identifier system for data posters will make it easier to link back unambiguously to a researcher as well as to make it easier for grant review committees and universities to receive a full picture of a scientist's impact, not just their publication list.

Standards for Interoperability, Re-Use and Re-Purposing

About me:

I am a Senior Fellow at the Kauffman Foundation, the Group D Commons Leader at Sage Bionetworks, and a Research Fellow at Lybba. I've worked at Harvard Law School, MIT's Computer Science and Artificial Intelligence Laboratory, the World Wide Web Consortium, the US House of Representatives, and Creative Commons. I also started a bioinformatics company called Incellico, which is now part of Selventa. I sit on the Board of Directors for Sage Bionetworks, iCommons, and 1DegreeBio, as well as the Advisory Board for Boundless Learning and Genomera. I have been creating and funding jobs since 1999.

The Alexandria Archive Institute
Comments on the Request for Information on Public Access to Digital Data
Office of Science and Technology Policy
(FR Doc No: 2011-28621)

January 12, 2012

Overview

The Office of Science and Technology Policy (OSTP) recently issued a Request for Information welcoming comments and recommendations for ensuring long-term stewardship of, and broad public access to, digital data resulting from federally funded research. The Alexandria Archive Institute (AAI) commends the OSTP for further exploring this topic.

The AAI (http://alexandriaarchive.org) is a non-profit organization that works to promote the dissemination and curation of digital scholarly resources. To this end, we developed Open Context (http://opencontext.org), a free, open access system for the publication of research content. Open Context demonstrates readily achievable ways to cultivate a distributed foundation for digital scholarship. Its methods for data portability enable researchers to work across silos and use a host of visualization, search and analysis tools. By leveraging archival and identity services offered by the University of California's California Digital Library (CDL), Open Context gains a strong institutional foundation for permanent citation and archiving.

We are delighted to have the opportunity to weigh in on the topic of "public access to digital data." Our responses to the questions posed in the RFI are based on ten years of exploration of issues around open access to digital data in the scholarly community. Below, we list our primary recommendations for encouraging public access to and preservation of digital data resulting from federally-funded research. Our responses to each of the RFI's questions follow the recommendations and provide more details to support each recommendation.

The OSTP request is the most recent development in broad moves to foster improved access, transparency, and stewardship of scientific data. The National Science Foundation (NSF) and private foundations have invested in developing technologies, standards, and datasets to support research. While we applaud recent developments promoting scientific data integrity and accessibility, policy provisions need to be strengthened. Data sharing remains at the margins of professional practice (*Nature* Editors 2009). The scientific community needs to put greater emphasis on data access and reuse to promote more robust, analytically rigorous, and more replicable scientific inquiry. To do so, the OSTP should

adopt a number of policies to clarify key requirements for maximizing the value of scientific data.

Summary Recommendations

Our recommendations are as follows:

- Cultivate a distributed information ecosystem: Integration, synthesis, analysis, and visualization of scientific data can foster tremendous opportunities across the commercial, not-for-profit and academic sectors. Agencies should foster an "open playing field" encouraging innovation in scientific data management and fresh ideas to advance new workflows, organizational forms, and technologies. To cultivate an open playing field, agencies need to promote the free flow of scientific data across multiple platforms and applications employing widely-used open and non-proprietary standards and formats.

- Cultivate a robust preservation infrastructure: Qualified digital libraries and digital archives are needed to maintain the integrity and longevity of scientific data. But not every participant in science data sharing needs to be a repository. To encourage innovation and experimentation, "sustainability" should not be required of every dissemination, visualization, analysis or aggregation platform. Rather, sustainability efforts should focus on digital libraries and archives. Since our understanding of how to best preserve digital data continually evolves, policymakers need to encourage innovation and collaboration across a broad spectrum of public interest organizations, particularly libraries and museums dedicated to playing stewardship roles. Multiple models, approaches, and organizations should play a role in scientific data stewardship to encourage continual learning and innovation in data longevity practices.

- Encourage data professionalism: Federally-funded research both creates and reuses data. Scientific integrity requires proper publication (including documentation) of data, and proper attribution and sourcing of reused, reanalyzed datasets. Data publication (including various models of peer-review and disciplinary archiving) and citation practices need to be mandated for federally funded research.

- Require non-proprietary data: The purpose of public support of science is to expand human understanding, not to subsidize particular commercial publishing models. In general, primary scientific data should be as free as possible from intellectual property and proprietary encumbrances. Such encumbrances create legal risk and complexities

that inhibit innovation around scientific data. Datasets should be in the public domain or under an open copyright license (such as the Creative Commons Attribution License) to widely encourage innovative approaches to data preservation and reuse.

- Data ethics: At the same time, the general need for minimized legal encumbrances should be balanced with data privacy and sensitivity issues. Privacy, research ethics, environmental and public health security concerns, and cultural property and indigenous rights needs, all require consideration. Defining ethical practices for data preservation, dissemination, and reuse will require broad-based, multi-stakeholder negotiations for different types of data in different scientific domains.

Responses to Questions

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Recent moves by NSF and NEH requiring Data Management Plans of all grant-seekers demonstrate how data sharing is becoming an expected outcome of the research process. These new requirements have the potential for improving transparency in research. Shared data also opens the door to new research programs that bring together results from multiple projects.

The downside to these new requirements is that grant-seekers may lack expertise and technical support in making data accessible. Thus, the new data management requirements will initially represent something of a burden, and many grant seekers may be confused about how to proceed. However, we expect the benefits of greater data accessibility, quality, and longevity will greatly outweigh any costs as expertise, support services (including the "Data Management Plan" tool offered by the California Digital Library and partners), and infrastructure mature.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

It is imperative that private intellectual property interests do not come at the expense of the public interest in promoting open and reusable scientific data. The public interest is best

served if scientific data can freely flow within a rich and competitive ecosystem that includes commercial publishers, aggregators and others, including nonprofits and academic institutions that contribute valuable services. Commercial, institutional, or other private interests should not have exclusive rights over scientific data created through public financing, since exclusive rights would only impede scientific inquiry and public transparency by imposing higher costs and legal risks. Public support for science should not serve as a subsidy for commercial (or not-for-profit) publishing interests.

The information ecosystem needed for scientific data management should make provenance and attribution of datasets easy to establish. This will promote citation and credit, both of original data creators and down-stream agents (commercial or non-commercial) that further enhance value. Citation is absolutely integral to scholarly practice. It enables and expresses collaborative knowledge production across space and time, and it is the foundation on which evidence and arguments are identified, assembled, reused, and critiqued. A major element on which careers are made and judged, citation is a key aspect of bringing digital communications into professional reward and incentive systems. But citation should not require complicated licensing or other encumbrances that would only make data expensive and legally risky to reuse.

Reliable and robust citation systems are key requirements for publishing data (Altman and King 2007). The DataCite project is establishing dataset citation standards and systems. DataCite promotes simple and readily adopted metadata requirements. To ensure persistence in citation, DataCite also promotes institutionally backed persistent identifiers, such as DOIs (Document Object Identifiers) and ARKs (Archival Resource Keys). The DataCite consortium can help establish the policy and technical requirements needed for efficient processes related to data citation, provenance, and credit.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The field sciences (archeology, field biology, and other areas of the environmental sciences) can be highly heterogeneous in terms of institutional context, methodology, and disciplinary perspectives. Because of this heterogeneity, it is often hard to establish pre-planned or top-down "cyberinfrastructure" that can meet researcher needs. To address this gap, policy efforts should encourage more bottom-up "Web-style" approaches as appropriate. Effective data management needs wide and open participation among diverse domain researchers to develop standards, data dissemination systems and work practices, as well as funding streams to support research that reuses existing datasets.

Data sharing advocates agree that data sharing requires more than "dumps" of raw and undocumented data on the Web. To be useful and used, data need adequate documentation to facilitate discovery and intelligibility. Data must also be disseminated through trusted

channels with clear versioning, quality control, and preservation mechanisms. Offering data with sufficient quality and levels of documentation requires expertise and effort. Doing so implies greater professionalism than encompassed by the term "sharing."

The term "publication" often better captures the effort, thought, and professionalism needed to make data dissemination meaningful (Kansa 2010). Publication models can align professional and career interests with public and scientific interests (see Costello 2009; Griffiths 2009; Piwowar et al. 2007). Policy efforts should promote innovative forms of professionally edited and reviewed data publication. Different data publication outlets can play an important role in shaping and communicating standards and expectations of data quality. The effort of cleaning and documenting data can be spread across multiple dissemination venues, each with editorial processes or other quality control mechanisms to improve quality, documentation, and alignment to expected standards.

In terms of technology, systems for reusing and analyzing shared data should be open for a variety of approaches. Experimentation can explore different mixes of "Linked Data" (Semantic Web) and more widely used "Plain Web" (Wilde 2008) approaches (favored by many commercial Web and mash-up developers) that may be appropriate in different circumstances. But all of these experiments must be supported by a distributed preservation infrastructure that safeguards data for the long term (Kansa 2011).

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

It is difficult to anticipate the impact of datasets over time. We need to experiment and develop a much richer body of experience to better understand which datasets to prioritize for dissemination and archiving. Data dissemination practices should experiment with ways of tracking the impact and use of different types of data, both with citation impact measures and alternative (Web, social media) impact metrics. Thus, different research communities will learn from experience about which types of data to prioritize for dissemination and archiving.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

All aspects of the data lifecycle must be understood and supported. For example, AAI is now working on a data publication workflow tools and practices in order to help shepherd datasets from the hands of the author, through a data clean-up, editorial, and documentation process, to a university-backed digital repository where the datasets receive permanent identifiers and can be discovered and used in different ways. In addition to our

approach, many other models of scientific data dissemination should be explored, since optimal approaches will be highly context dependent and will no doubt evolve as methods, expectations, and technologies change.

Demonstrated scholarly outcomes will help make a compelling case for sharing datasets. Graduate students, undergraduates, high-school students, and informal-education learners need training opportunities that offer Web data skills so that the next generation of researchers can best make use of emerging data resources.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Funding agencies should require a line item in proposed project budgets that commits a certain proportion of the project budget to publication (ideally in open access venues), as well as data dissemination and preservation. Dissemination and data preservation services and costs should be outlined in a project's data management plan and justified in budget justifications. Reviewers should be asked to carefully consider whether a project's budget seems appropriate to the proposed data management plan. Reviewers must understand key issues in data management so that they can better evaluate data management plans in proposals. Finally, funded projects should be required by the funding agencies to provide details in their interim and final reports about how project data have been disseminated and archived according to the data management plan.

In general, scientific knowledge and its underlying foundation of data can be expensive to produce. It requires expertise and often great effort. But to make data integration and reuse feasible, data need to flow freely and interchangeably. Unless there are strong overriding ethical or security concerns (chiefly privacy), access and use should be free-of-charge and free of legal encumbrances (especially proprietary IP interests). Therefore data dissemination and archiving services should focus cost recovery on accession fees (budgeted in grant proposals), not on fees for later access or use.

The key point is that publicly funded research should create public information goods. The outputs of publicly funded science should be available in a robust, expanding, and open public commons. Commercial interests can and emphatically should build upon this commons of public information goods. But, particular commercial or even nonprofit interests should not be able to monopolize the public commons or exclude others from freely drawing upon the fruits of publicly financed research.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Participation in suitable, disciplinary dissemination outlets (data publication venues) and digital repositories should be easy to verify. In order to serve any scientific purpose, these systems will provide datasets with enough metadata documentation to make it easy for officials to verify compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Granting agencies should develop funding streams specifically targeted toward encouraging graduate student use of publicly accessible data. These funding streams will help cultivate the needed Web data and analytic skills required to use public datasets effectively. They will also help change scientific cultures in ways that encourage greater openness, transparency, and participation in scientific data sharing systems.

In addition, to cultivate commercial innovation in this space, agencies can develop SBIR-type granting programs that fund innovative commercial projects that draw upon and add value to publicly accessible research data without monopolizing or excluding others from those data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The best protection from "data theft" is clear dissemination in widely accessible, easily searched, professionally recognized venues. Professional social norms of citation need to be promoted rather than encumbering and legally complex forms of licensing. Clear citation practices, supported by appropriate technical infrastructure, will promote proper attribution and professional rewards. Agencies can also encourage participation in open-access "data publication venues" with review and other editorial processes recognized by the researcher community. Publication of data, in recognized forums, where citation impacts can be tracked (like article impact metrics) can provide appropriate professional recognition for data creators.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Standards continually evolve and will vary by research community. Rather than dictating specific standards, agencies should set policies that promote adoption of "good practices" while recognizing that such practices continually evolve.

A key requirement for data dissemination should be data portability. Data should not be trapped in a given system or repository. Rather, data need to freely flow into different applications and systems. This requires that data have open licensing or lie in the public domain. Also, various machine-readable representations of data need to be available, though the specific formats and standards may vary across and between different research communities. In some cases, more elaborate standards may be required. In other cases, overly complex standards requirements may inhibit adoption. Simple and lightweight technical and semantic standards that yield immediate and tangible benefits may be most suitable for scientific domains with little funding or technical support.

The World Wide Web is an obvious choice for dissemination. As much as possible, data dissemination systems should adopt best practices in Web architecture. Systems need to adopt non-proprietary open standards and offer data in multiple machine and human readable representations. In many cases, Web-based dissemination should also emphasize and support "Linked Open Data" methods.

In our own efforts with Open Context, we have emphasized low-barrier to entry approaches to sharing machine-readable data. We offer data in widely used and recognized formats, including JSON, the Atom Syndication Format and other XML vocabularies. We are also incrementally adopting more Linked Data standards and services for interacting with RDF data. Our experience shows the need to continually adapt to expose data in new formats and services as expectations and needs evolve over time.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Some of the best examples of standards development come from the public Web. The Atom Syndication Format is a major example, but other successes include GeoRSS and GeoJSON. Development of these standards was largely "bottom up" where software developers (the stakeholders that actually implement the standard) play a key role in shaping standards development. Keeping complexity and barriers to entry at a minimum is vital in any standards building effort.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies can offer funding support for efforts that help build ties between international collaborators. The "Digging into Data" challenge organized by NSF, IMLS, and NEH represents a good, though under-funded model (grants awards were quite small relative to the costs and complexity of multinational collaborations).

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

The Web is the most successful distributed computing platform ever developed, and principles of Web architecture need to be used when linking between scientific datasets and other scholarly publications. Unfortunately, many scientific publications are merely electronic analogs of "paper," typically with little linking to other resources published on the Web. We need to see far more innovation in scientific publication processes, but this innovation is hamstrung by professional reward and evaluation structures and by the fact that scientific publishing is dominated by a few monopolistic commercial publishing houses.

In addition, the divide between a scientific "dataset" and "publications" is not always hard and fast. Many important scientific inferences and analyses could be conducted through large-scale text analyses and mining of scientific literature. However, access to this literature is highly restricted through very expensive and tightly guarded commercial channels. These restrictions make it difficult for the scientific community to explore ways to better use and understand vast bodies of scientific literature.

The NIH requires public access to peer-reviewed outcomes of NIH funded research (usually delayed by one year, so commercial publishers temporarily enjoy exclusive dissemination rights). Similar requirements should be made by other federal agencies. In this way, publications can be used to generate additional datasets, and datasets can be fully understood and contextualized by accessible publications.

References

Altman, M., and King, G. 2007. A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). Retrieved from http://dx.doi.org/10.1045/march2007-altman

Kansa, E.C. 2011. New Directions for the Digital Past. In *Archaeology 2.0: New Tools for Communication and Collaboration*, edited by E.C. Kansa, S.W. Kansa and E. Watrall. Los Angeles: Cotsen Institute of Archaeology Press, pp. 1-25. Retrieved from http://escholarship.org/uc/item/1r6137tb#page-17.

Kansa, E.C. 2010. Open Context in Context: Cyberinfrastructure and Distributed Approaches to Publish and Preserve Archaeological Data. *The SAA Archaeological Record* 10(5), 12-16.

Costello, M. J. 2009. Motivating online publication of data. *BioScience* 59, 418–427.

Griffiths, A. 2009. The Publication of Research Data: Researcher Attitudes and Behaviour. International Journal of Digital Curation 4. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/101.

Nature Editors. 2009. "Data's shameful neglect." *Nature* 461, 145 (10 September 2009).

Piwowar H.A., Day R.S., Fridsma D.B. 2007. *Sharing Detailed Research Data is Associated with Increased Citation Rate*. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308 Retrieved from http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308

Wilde, E. 2008. *The Plain Web.* pp. 79-83. In: Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008), 22 Apr 2008, Beijing, China. ISBN 978 085432885 7.

Carnegie Mellon

Office of the Dean
University Libraries
Carnegie Mellon University
4909 Frew Street
Pittsburgh, Pennsylvania 15213-3890

(412) 268-2447
Fax: (412) 268-2793
gstclair@andrew.cmu.edu
http://www.library.cmu.edu

January 12, 2012

OFFICE OF SCIENCE AND TECHNOLOGY POLICY (OSTP) – Request for Information: Public Access to Digital Data Resulting from Federally Funded Research

On behalf of Carnegie Mellon University and the roughly 4,000 faculty and staff we represent, I write to thank you for this opportunity and to share our perspective on public access to digital data resulting from federally funded research. Carnegie Mellon is a small, private university with over 11,000 students and 86,500 alumni. Recognized for our world-class programs in technology and the arts, interdisciplinary collaborations, and leadership in research and education, we are innovative and entrepreneurial at our core.[1]

Since 2007, the Association of University Technology Managers has ranked Carnegie Mellon first among U.S. universities without a medical school in the number of startup companies created per research dollar spent. Our 118 research institutes and centers create 15 to 20 new companies each year. Over the past 15 years, we helped start 300 companies, creating 9,000 jobs. Because of our success in research, innovation, and entrepreneurship, other universities have adopted our Greenlighting Startups approach to fostering commercial enterprises,[2] and Google, Apple, Disney, Intel, and Lockheed Martin have opened space on or near campus.

Carnegie Mellon University's 2011 financial statement reports that 38.4% of our total revenue was from sponsored projects, totaling $360.9 million. Federally funded projects account for $317.59 million (88%) of this revenue.[3] Our community creates a large quantity of federally funded research data. We strongly support public access to these datasets because open data will increase productivity, innovation, and commercialization. Developing an open data policy is in the national interest and warrants careful examination. We thank the Office of Science and Technology Policy for the opportunity to respond to its Request for Information. The rationale for our comments is provided at the end of this document.

Preservation, Discoverability, and Access

COMMENT 1

To maximize return on taxpayer investment, grow the economy, and improve the productivity of science, federal agencies must mandate that all data gathered in federally funded research projects be made available to the public – open – for use under appropriate licenses. Digital datasets should be

[1] See http://www.cmu.edu/about/index.shtml.

[2] For more information, see *Five Percent, Go in Peace*, available at http://www.cmu.edu/startups/go/index.html.

[3] *Carnegie Mellon University Consolidated Financial Statements June 30, 2011 and 2010*. See http://www.cmu.edu/finance/reporting-and-incoming-funds/financial-reporting/files/2011-annual-report.pdf.

promptly archived and accessible in trusted repositories committed to open data and preservation. Acknowledging that different disciplines operate under different constraints, federal agencies should work with their research communities to specify the conditions and timeframe within which data must be made open. (Trusted repositories are discussed in comment #5. Constraints are discussed in comment #2.)

Ideally, licenses for open data should be human- and machine-readable. Appropriate licenses for open data include[4]:

- Open Data Commons Attribution License, which requires only attribution and grants full use rights.
- Open Data Commons Open Database License (ODbL), which requires attribution and share-alike, meaning any derivative work must also be open.
- Open Data Commons Public Domain Dedication and License (PDDL), which waives all rights and places the data in the public domain.
- Creative Commons CC Zero, which waives all rights and places the data or content in the public domain.[5]

Additional licenses might need to be developed. An appropriate license will preserve rights and provide incentives for researchers to make their data publicly accessible.[6]

Attempts to restrict public use of federally funded research data to non-commercial purposes will stifle innovation and commercialization, unnecessarily limiting the return on taxpayer investment in research. In regard to whether products and services developed using open data must themselves be open (i.e., must the initial data be licensed under a share-alike license), this might effectively be addressed by requiring openness if and only if the subsequent use were federally funded. In all cases, however, subsequent use of open data should require attribution to the scientists and federal agency.

In addition, federal agencies mandating public access to digital data should:

- Require data management plans and budgets to be included in grant proposals submitted for peer review. Plans should describe the data to be gathered, the applicable standards or best practices (for data and metadata), the repository where the data will be deposited for access and preservation, and the license to be applied granting use rights. Plans should also address any key concerns or constraints, such as privacy and confidentiality, contractual obligations, and the timing of public access.[7] (See comment #2.)
- Prohibit researchers from spending money allocated for data management on anything other than data management.

---

[4] For details, see *Open Data Commons Licenses FAQ*, available at http://opendatacommons.org/faq/licenses/.

[5] Data placed in the public domain (with a PDDL or CC Zero license) can be hosted for free at the Talis Connected Commons. See http://blogs.talis.com/n2/cc.

[6] See *Digital Research Data Sharing and Management* (December 2011), p. 7.

[7] According to the National Science Foundation, "Using the Data Management Plan to determine the timeline for initiating the data sharing process recognizes the rights and responsibilities of investigators." See *Digital Research Data Sharing and Management* (December 2011), p. 9.

- Allow money allocated for data management to be spent to support data management infrastructure, including equipment and personnel at the institution receiving the funds and the trusted repository where datasets are deposited.[8]
- Promote and maintain open access copies of relevant existing or emerging standards and best practices for data management, including metadata.[9]
- Require research communities that do not have standards and best practices to develop them within a specified time frame.[10] Federal agencies should work with their communities (researchers and institutions receiving grant funds) and repository developers to ensure that this happens. (See comment #12.)
- Work with their research communities to promote the value of openness and to understand the inhibiting factors so that appropriate concessions can be made without unnecessarily slowing progress towards the goal. (See comment #2.)
- Promote public access policies and monitor compliance as a *quid pro quo* for future funding.
- Encourage the development of a standard, persistent identifier (something comparable to the PMC ID) for digital datasets. Policies should require this ID to be included in reports to the agency, publications that reference research findings associated with the dataset, and (if appropriate) subsequent data management plans. (See comment #7.)
- Maintain a registry of publicly accessible datasets funded with taxpayer dollars. Registry records should include the dataset ID and a link to the dataset location, and be discoverable in an Internet search. (Further details on the registry are provided in comments #7 and #8.)

To date, the National Institutes of Health (NIH), National Science Foundation (NSF), National Endowment for the Humanities (NEH), and the Institute for Museum and Library Services (IMLS) have adopted data management requirements for some or all of their granting activities. Their leadership is commendable and will demonstrate whether or not a requirement is sufficient to attain the goals of open data. As the National Institutes of Health (NIH) experienced with its public access policy, a legislative mandate might be necessary to accomplish the goals and reap the benefits of open data.

COMMENT 2

Carnegie Mellon University is heavily invested in and supportive of its research programs. We are proud of the intellectual output of our researchers, and want to protect their rights to use their intellectual output, including data, to its fullest. While many datasets are not protected by copyright and are not, in the legal sense, "owned" by the researchers, their de facto rights to the data cannot be denied.[11] Federal policies on open data must recognize these rights and the complex and often highly competitive environment in which they exist. The use of data to advance researcher careers, develop

[8] The National Science Foundation acknowledged that maintenance of trusted digital data repositories should be considered in data management plans to ensure sustained access to the data. See *Digital Research Data Sharing and Management* (December 2011), p. 6.

[9] If providing open access copies is not feasible, federal agencies should at minimum provide a list of relevant standards and best practices with links to where researchers can get the documents.

[10] We at Carnegie Mellon share the National Science Foundation's position that given the increasing scale, scope, and complexity of data, each research community should take the responsibility for developing standards for data stewardship that are accepted across fields of science and engineering. *Digital Research Data Sharing and Management* (December 2011), pp. 3-4. Available at: http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf.

[11] In the United States, some types of data are not protected by copyright. For example, numeric data are treated as facts, and therefore are not copyright protected. They are, however, proprietary. In any case, the owner of federally funded research data is either the funding agency or the institution funded to do the research, not the principal investigator(s).

patents, and contribute scholarship is a top concern for faculty and students at Carnegie Mellon and at research institutions across the nation.

Publishers have no claim to federally funded research data and no stake in how the data are licensed for distribution or use, though they may provide links to the datasets underlying their publications.[12] The stakeholders are the federal agencies and taxpayers who underwrite the research, the scientists who conduct it, and the institutions that manage the grants, provide laboratory space, and pay researcher salaries. Stakeholder interests can be protected by appropriate licenses and timelines for deposit in a trusted repository. (Appropriate licenses are discussed in comment #1. Trusted repositories are discussed in comment #5.)

Within this framework, critical concerns and constraints must be acknowledged, including

- The need to protect privacy and maintain confidentiality
- Data protocols required by international consortia or federal government science and technology agreements
- Contractual obligations (for projects with multiple funders)
- Scientists' concerns about competitive advantage

Research shows that among both academia- and industry-based scientists, as the competitive value of the requested information increases, the likelihood of sharing the information decreases.[13] Competition reduces openness and sharing, but it can also drive science and grow the economy. Competitive advantage must be preserved.

To address the issues of researcher rights, scooping, competition, and potential commercial value, the federal government should, in collaboration with the research community, specify a timeframe within which data must be made publicly accessible. Depending on the discipline, this may be before or after peer-reviewed publication of research findings. (See comment #3.)

While the ideal is prompt public access, in some disciplines the goals of growing the economy and increasing the productivity of science might be achieved more effectively by granting the researcher(s) control of the data for some finite time, after which the data becomes open and competitors can use it for commercial or non-commercial purposes. This could be accomplished by requiring prompt deposit in a trusted repository, but allowing the data to reside in a dark archive until it is licensed for public access – something akin to an embargo on public access to scholarly publications.[14] The point at which the dataset will become open should be specified in the administrative metadata.

Federal agencies should establish check lists for their research communities to address constraints applicable to maintaining and sharing data. Guidance on acceptable constraints and manuals of steps to be followed would greatly assist scientists writing data management plans. Peer reviewers should consider whether data management plans effectively address key concerns and constraints.

---

[12] Authors who publish their research findings may be required to transfer the copyright in their written expression to the publisher, but ownership of the data is not, in most cases, transferred to the publisher.

[13] C. Haeussler (February 2011), "Information-sharing in academia and the industry: A comparative study," *Research Policy* 40 (1): 105-122.

[14] The National Science Foundation acknowledges that an embargo period for open data may be necessary in some cases. See *Digital Research Data Sharing and Management* (December 2011), p. 6.

COMMENT 3

Disciplinary differences in data types and formats must be addressed through standards and best practices. Federal agencies should facilitate the development and dissemination of standards and best practices for digital data and its attribution. They can do this by

- Maintaining open access copies of relevant standards and best practices for data management (including metadata).[15]
- Requiring research communities that do not yet have relevant standards and best practices to develop them within a specified time frame. Federal agencies can identify disciplines that are poorly prepared to comply with open data policies and encourage them to collaborate with experienced and trusted partners, e.g., university libraries.
- Participating in and funding standards development activities. (See comment #12.)

In addition to differences in data types and formats and preparedness to manage them, disciplines have different levels of understanding of the benefits of openness and different pragmatic needs (e.g., to preserve competitive advantage). Federal agencies need to understand their research communities, take steps to remove unnecessary barriers, and make appropriate concessions that facilitate science as well as openness.

Federal agencies can work with scholarly societies to ensure that researchers understand the benefits of open data. They can establish minimal levels of service to be provided by trusted open data repositories. (See comment #5.) And they can endeavor to understand and address the most intractable environmental factor impeding openness: competition. (See comment #2.)

COMMENT 4

Admittedly the cost of digital data management will vary significantly across disciplines and projects. A relatively new endeavor, much remains to be learned about the various associated costs, for example, the cost of creating metadata, the cost of converting data to an open format, and the cost of long-term storage and migration. Grant proposal budgets and budget justifications should include the projected costs of data management. Federal policies should prohibit researchers from spending money allocated for data management on anything other than data management. Over time, the costs associated with managing public access to different types of data will be better understood, as will the optimum allocation for sharing in different disciplines. Those concerned about the high cost of data management should be made aware of the Knowledge Investment Curve that graphically conceptualizes the advance of science as a function of conducting research and of sharing the results.[16]

[15] If providing open access copies is not feasible, federal agencies should at minimum provide a list of relevant standards and best practices with links to where researchers can get the documents.

[16] W. Warnick and D. Wojick (August 2009), "The Knowledge Investment Curve," *OSTIBLOG*. Available at: http://www.osti.gov/ostiblog/home/entry/the_knowledge_investment_curve. The Rationale provided at the end of this document provides further information on the Knowledge Investment Curve.

COMMENT 5

Successful implementation of a data management plan requires standards or best practices and a trusted repository for open data. Research communities are at different levels of preparedness in both areas.

Federal agencies should require research communities that do not have standards or best practices for data management to develop them in a specified timeframe and encourage them to work with universities[17] and other trusted institutions to ensure that this happens. They can assist with this work by identifying potential collaborators and funding research and development. (See comment #12.)

The federal government should establish minimal service criteria to be met by trusted partners, for example:

- Support for appropriate open data licenses. (See comment #1.) Trusted repositories must be prohibited from converting data deposited in an open format into a proprietary format upon retrieval or download.
- Support for relevant standards and best practices for access, interoperability, and preservation, including metadata, protocols, hardware, software, and unique persistent identifiers for datasets, researchers, and organizations.[18]
- Searchable descriptive metadata that includes the licensing terms and, if attribution is required, a list of those requiring attribution. (See comment #9.)
- Verification of data integrity at ingest and retrieval / download.
- Security, redundancy, migration, disaster preparedness, and other preservation strategies, including the rights and technical metadata needed to preserve digital data.
- A mechanism for reporting problems.
- A mechanism for determining storage and preservation costs and a commitment to containing costs through cooperative agreements and economies of scale.
- Licensing agreements (between the repository and the owner of the dataset) that grant the rights necessary to preserve open data.[19]

Trusted repositories will have not only a commitment to long-term maintenance of digital datasets documented in a service-level agreement, but the financial resources and knowhow to sustain the operation.[20] If publishers meet the minimal service criteria, they may provide data management services. If not, they may only provide links from their publications to the underlying datasets deposited in a trusted repository.

Federal agencies should maintain a list of trusted repositories for various types of data.[21] To facilitate the development of trusted repositories for open data, they should work with university libraries,

---

[17] Within universities, data management and preservation services should be centralized within an administrative unit (for example, the library), not decentralized within academic departments, to take advantage of economies of scale and institutional commitment.

[18] See *Digital Research Data Sharing and Management* (December 2011), pp. 4-5.

[19] *Trusted Digital Repositories: Attributes and Responsibilities* (May 2002). An RLG-OCLC Report. Mountain View, CA, pp. 18-19. Available at: http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf.

[20] *Trusted Digital Repositories: Attributes and Responsibilities* (May 2002), p. 26.

[21] This list could be generated from registry records such as DataCite. See http://www.datacite.org/repolist.

disciplinary societies, research consortia, and other stakeholders to distribute the many responsibilities associated with establishing and maintaining a trusted repository for digital data.[22]

COMMENT 6

Federal agencies must allocate funding that can only be used for data management and preservation. The requirements for these funds should be amended to allow them to be used to support data management infrastructure, specifically:

- Equipment and personnel at the institution receiving the grant, thereby providing resources that can carry over from one project to the next.
- Trusted repositories, providing financial support to sustain these initiatives and guarantee long-term public access to the data.

Researchers must be required to include a data management plan in their grant proposals. Data management costs must be included in detailed budgets and budget justifications.

In addition, federal agencies should fund research aimed at determining the costs of data preservation and access in different disciplines. Such research should be conducted in collaboration with researchers and trusted repository partners, and the findings disseminated to inform subsequent data management plans.

COMMENT 7

Researchers need an easy way to create and manage a unique persistent identifier for digital datasets. One possibility is EZID, developed by the California Digital Library.[23] The EZID service enables users to create identifiers for objects on the web, to maintain their current locations so people can click on the identifier and link directly to the object, and to store associated metadata with the identifier. Another alternative is the Digital Object Identifier (DOI) provided by DataCite. DataCite DOIs resolve to a public web page with information about the associated dataset and a link to the dataset itself.[24]

As with the NIH public access policy, federal agencies should require researchers to include the dataset ID in reports, publications, and subsequent grant proposals. However, unlike the NIH PMC ID, a dataset ID might not necessarily signal compliance with an open data policy. For example, the ability to update locations and metadata enables researchers to get a preservation-ready EZID identifier before they gather the data or deposit it in a trusted repository.

To enable federal agencies to measure and verify compliance with open data policies, researchers (or their designates) should be required to report when the dataset has been deposited in a trusted repository, preferably with an easy-to-use interface that generates a record for the federal government's registry of publicly accessible datasets funded with taxpayer dollars. Deposit of the dataset and creation of the registry record should be required by the end of the grant period or when the final report is due, though for some datasets there may be a period of restricted access (i.e., the dataset resides in

[22] See *Digital Research Data Sharing and Management* (December 2011), p. 6.
[23] See http://www.cdlib.org/services/uc3/ezid/index.html.
[24] See http://datacite.org/whatdowedo.

a dark archive) before it becomes publicly accessible. (See comment #2.) Compliance – creation of a registry record for the dataset – should be *quid pro quo* for future funding from federal agencies.

COMMENT 8

Potential users must be aware of the existence and location of federally funded, publicly accessible datasets. Dataset IDs referenced in research publications and discoverable in an Internet search will facilitate discovery and use. In addition, the government should maintain a searchable registry of federally funded open datasets. Registry records would provide public access to descriptive metadata about each dataset, including licensing terms and attribution (researchers, agency, grant ID), the name of the trusted repository where it resides, and a link to the dataset.

COMMENT 9

Ideally, the descriptive metadata bundled with the dataset will convey the licensing terms and include a list of those to be attributed. (See comment #2.) However, the attribution of credit for datasets is a relatively new field of endeavor. Many groups are in the process of determining best practices for data citation in the sciences and humanities. Strict guidelines for data citation cannot yet be provided, but federal agencies requiring data sharing and management can provide ongoing guidance for data citation, keeping close watch on new developments in the field. In addition, federal agencies should fund research into best practices and systems for data citation to accelerate the development of guidelines for researchers in different disciplines.

Two current development activities relevant to unique identifiers for attribution deserve mention. The National Information Standards Organization (NISO) is developing a recommended practice for use of the International Standard Name Identifier (ISNI) to identify institutions.[25] The ORCID (Open Researcher and Contributor ID) project is developing unique identifiers for individual researchers to resolve name ambiguity problems in scholarly communication.[26] Federal agencies should monitor these develops closely, and disseminate and encourage use of best practices and standards as they develop.

Standards for Interoperability, Reuse, and Repurposing

COMMENT 10

The data must be in an open, not proprietary, format. Trusted repositories – required to support relevant standards and best practices for interoperability and preservation – must be prohibited from converting data deposited in an open format into a proprietary format upon retrieval or download. Standard licenses tailored for open data must be applied.

COMMENT 11

Standards development is essentially a three-step process undertaken by a community of interest: develop, implement, promote. Experts familiar with a problem and stakeholders that will be affected by

[25] See http://www.niso.org/publications/isq/2011/v23no3/gatenby.
[26] See http://www.orcid.org/.

the proposed solution convene to draft a standard that meets their needs. The standard is released as a draft for trial use. During the trial period, implementers test the standard and the draft is open for public review and comment. At the end of the trial, the standard is balloted, revised, or withdrawn. If issues reported by implementers and stakeholders require significant revision of the draft, standards developers reconvene and produce a subsequent draft, released for another trial period. The process iterates until members of the relevant consensus body overseeing the process vote to approve or withdraw the proposed standard. Approved standards are promoted by relevant standards organizations. (See comment #12.)

The key elements of successful efforts are broad stakeholder involvement and commitment, the period of testing and feedback, and the development of consensus. Best practices are developed in a similar way, but can be accomplished much faster than standards because they serve as guidelines, allowing for experimentation, while standards require strict compliance, making consensus more difficult to achieve. Federal agencies should not underestimate the value of best practices, which are often forerunners to the development of standards.

COMMENT 12

Federal agencies, in so far as they represent the interests of their constituent communities, are in a strategic position to encourage the development of international standards for digital data. They can promote effective coordination of standards by working with their communities and repository developers to identify problems that standards will solve and by participating in the standards development process. Furthermore, they should monitor significant initiatives in digital preservation and disseminate relevant information to their constituencies. For example, the project *Planets* has built services and tools to help ensure long-term access to digital assets.[27] *DataCite* supports data archiving that permits verification and repurposing of the data and works to establish easier access to data.[28] The Science and Technology section of the Association of College and Research Libraries is convening a panel discussion on January 22, 2012 to provide standards development organizations with input from publishers, vendors and librarians about metadata and technical descriptors needed to enhance access to scientific datasets.

Federal agencies that are not already members of the American National Standards Institute (ANSI) should join ANSI as Government Members.[29] They should join the National Information Standards Organization (NISO) and serve on relevant working groups formed under the auspices of NISO. To achieve global reach, agency representatives should participate in the International Organization for Standardization (ISO).

The American National Standards Institute (ANSI) works to enhance the global competitiveness of U.S. businesses by promoting and facilitating standards and ensuring their integrity.[30] ANSI does not

[27] Preservation and Long-term Access through Networked Services (Planets) was a four-year project funded by the European Union. See http:www.planets-project.eu. The Planets project ended in May 2010, but the documents and deliverables are being maintained and developed by the Open Planets Foundation (OPF). Government bodies may join the OPF. See http://www.openplanetsfoundation.org/.

[28] See http://datacite.org/whatisdatacite.

[29] The list of ANSI Government Members is available at https://eseries.ansi.org/Source/directory/Search.cfm.

[30] See http://www.ansi.org/standards_activities/overview/overview.aspx.

develop standards, but rather accredits the developers that build consensus among qualified groups.[31] ANSI provides a forum for accredited developers to work together to develop American National Standards (ANS), and promotes the use of ANS internationally. ANSI also encourages the adoption of international standards as national standards when these meet community needs. ANSI is the only U.S. member of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). As such, ANSI plays an important role in creating international standards. "[T]he success of these efforts often is dependent upon the willingness of U.S. industry and government to commit the resources required to ensure strong U.S. technical participation in the international standards process.[32]

The International Organization for Standardization (ISO) initiates development of new standards in response to established need for them. Stakeholders submit a request for a standard. The relevant ISO technical committee reviews the request. If the committee verifies international need for the requested standard and most committee members support the work, a standard will be developed.[33]

The National Information Standards Organization (NISO) is accredited by ANSI to identify, develop, maintain, and publish technical standards to manage information, including storage, retrieval, re-use, metadata, interchange, and preservation. NISO standards serve those who publish information or provide tools to access, use, or preserve information. NISO represents (on behalf of ANSI) U.S. interests as the Technical Advisory Group to ISO's Technical Committee on Information and Documentation, and serves as the Secretariat for the Subcommittee on Identification and Description. NISO offers programs on standards issues and workshops on emerging topics. Committees are often formed after these events to develop new standards. Alternatively, best practices are developed and released as guidelines.[34] In addition, white papers are often written prior to standardization activity to explore key questions or to identify opportunities and possible approaches for standards development.[35]

Federal agencies should help fund and participate in NISO workshops and programs. Agency representatives should serve on NISO working groups and committees. (See comment #13.)

COMMENT 13

Federal public access policies for digital datasets must require researchers to include the dataset ID(s) in publications reporting findings based on the dataset. The dataset ID should be discoverable in an Internet search and link directly to the dataset.

A project to develop a best practice for supplemental journal article materials is underway under the leadership of the National Information Standards Organization (NISO) and the National Federation of Advanced Information Services (NFAIS).[36] The project focuses on publishers, specifically a best practice for publishers to include, handle, display, and preserve supplemental journal article materials.

---

[31] To maintain accreditation, standards developers must consistently meet the Institute's requirements for openness, balance, consensus, and due process. The requirements ensure that ANS are responsive to the needs of all stakeholders.

[32] See http://www.ansi.org/standards_activities/overview/overview.aspx.

[33] See http://www.iso.org/iso/about/how_iso_develops_standards.htm.

[34] See http://www.niso.org/publications/rp/.

[35] See http://www.niso.org/publications/white_papers/.

[36] See http://www.niso.org/workrooms/supplemental.

We have argued here that publishers that meet the criteria for trusted repositories may provide data management services. Other trusted repository partners, e.g., university libraries, could find the forthcoming best practice useful. The three groups established to develop the best practice – the Stakeholders Interest Group, the Business Working Group, and the Technical Working Group – are addressing a wide range of concerns, from semantic and policy issues (including metadata and persistent identifiers), the responsibilities of various stakeholders (e.g., authors, editors, publishers, and peer reviewers), interoperability, accessibility, and preservation (including migration). We encourage federal agencies to monitor the progress of this project and provide feedback on document drafts by joining the Stakeholders Interest Group at www.niso.org/lists/suppinfo.

In closing, we at Carnegie Mellon believe strongly that prompt, free access and use rights to federally funded datasets will eliminate unnecessary redundancies, accelerate science, and provide opportunities for innovation and commercialization unrealized to date because the data are unavailable for re-use. Public access to federally funded datasets – data freely available on the Internet where anyone may download, copy, analyze, process, pass them to software or use them for another purpose without financial, legal, or technical barriers – is the desired state. However, many unanswered questions and unresolved issues clutter the path to this desired state. Federal agencies are in an ideal position to move us forward on the path by

- Facilitating development of needed standards and best practices, including timelines for when data must be made open
- Establishing criteria for trusted repositories and maintaining a list of trusted repositories
- Implementing a registry of federally funded, publicly accessible datasets to facilitate discovery and assess compliance
- Funding research designed to remove obstacles in the path to open data

Thank you for the opportunity to provide comments on this important initiative.

Sincerely,

Gloriana St. Clair, Dean, Carnegie Mellon University Libraries
gstclair@andrew.cmu.edu

Denise Troll Covey, Principal Librarian for Special Projects
troll@andrew.cmu.edu

Rationale for comments

Mandating prompt public access and use rights to federally funded research datasets, working to ensure the development and dissemination of standards and best practices for data management, monitoring compliance with public access policy, and facilitating discovery (via a searchable registry) will grow existing and new markets by not only encouraging re-use, but by enabling use by more users and different kinds of users. The diversity of users and uses will yield innovations and

commercializations that stimulate investments and create jobs. Small businesses in particular will benefit from free access to federally funded datasets.

Prompt public access and use rights to digital datasets will increase the productivity of science by eliminating redundant efforts at data gathering, and enabling researchers to reproduce, verify, and validate previous work, thereby accelerating confirmations or rejections of research findings. Open data will also enable new uses and applications of the data, leading to new findings that advance science. Furthermore, open data will increase exposure, discouraging research misconduct.[37] Open data will bolster the productivity and integrity of science, and in so doing, bolster the public trust.

To achieve these goals and provide the maximum benefit to all stakeholders, data must be open. For data to be open, it must meet the following conditions[38]:

- The dataset must be available without charge in its entirety and in a convenient and modifiable form.[39]
- There may be no licensing restriction against or fees levied for redistribution or re-use.
- There may be no technological restrictions that obstruct free redistribution or re-use. The data format must be open, not proprietary.
- If a license requires attribution, the metadata for the dataset must provide a list of those requiring attribution.
- The license must not discriminate against persons, groups, or fields of endeavor.

The Knowledge Investment Curve graphically conceptualizes the advance of science as a function of conducting research and of sharing the results. While the actual shape of the curve is unknown, if 0% or 100% of funding is invested in sharing, the pace of scientific discovery will be zero. The optimum amount to be invested in sharing will vary by discipline.

> We can ask then what the federal investment should be in Web-based science sharing. Conceptually, points on the Knowledge Investment Curve to the left of the optimum imply that the pace of science discovery would be accelerated by increasing the percentage of funding for sharing results. One thing we know is that the investment in sharing is highly uneven across the various sciences. The fraction of health science research funding dedicated to sharing knowledge is greater than for physical and energy sciences. The latter is unlikely to be near the optimum.[40]

Federal policies on open data, incentives (appropriate licenses), and monitoring of compliance will, over time, provide much needed information about the costs of data sharing and preservation and the optimum amount of funding to be allocated to conducting research and sharing the results.

---

[37] The blog *Retraction Watch* routinely reports journal articles retracted for plagiarism and other types of research misconduct. See http://retractionwatch.wordpress.com/.

[38] See http://opendefinition.org/okd/ for further details.

[39] All data cannot be shared over the network because of bandwidth issues. Reasonable fees may be levied to cover the cost of media to transport such open datasets.

[40] Warnick and Wojick, 2009.

Thu 1/12/2012 9:30 AM
Response RFI: On Public Access To Digital Data Resulting From Federally Funded Scientific Research FR Doc. 2011-28621

Effective reuse of data resulting from scientific research is a multi-faceted challenge that goes beyond simply archiving and distributing data files. To effective build upon the prior work of their peers, researchers must be able to both find and interpret relevant data:
voluminous data archives will be of little value without specific clear and informative metadata and tools that leverage that metadata to help identify data of interest.

The generation of this metadata presents significant challenges. Although the success and evolution of metadata formats like the MIAME model cited in the RFI provides examples of best practices, these models suffer from several deficiencies that limit their impact. So-called "minimal information" models like MIAME are, by definition, limited in their expressiveness. Effective data sharing might require metadata that goes significantly beyond the baseline "minimal" description. However, the generation of more fully-descriptive metadata is a time-intensive task that is often not well-supported by existing data management tools. This difficulty is compounded by a lack of incentives: data annotation is most often the responsibility of the data generator, who may see this task as a cumbersome overhead requirement with little direct value-added.

A combination of data models, annotation tools, and search tools that leverage those annotations is needed to address these shortcomings. To be successfully and widely adopted, these tools must be designed to be well-integrated with existing tools and work practices.

Responses to Specific questions:

1. What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Scientific agencies should take several steps to encourage public access and preservation:

* Provide specific requirements for preservation and publishing of public access data
* Identify specific data models and tools to be used for various data types
* Promote the development of more extensive and usable tools for annotating and finding research data
* Support the development of tools that promote best practices for archiving and managing data
* Require specific data sharing plans and dedicated resources in appropriate funding rewards

2. What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Continuing embargo policies that provide researchers with the opportunity to use data for publication and to apply for patents should be promoted and adopted to the need of specific communities, particularly with respect to delays relative to publication, patent, or other trigger time points.

3. How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Engagement with appropriate research organizations, discipline-specific workshops, and additional RFIs can be used to understand the needs of specific communities and to plan accordingly.

6. How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Preservation and archiving costs could be considered as "overhead" costs that would go "above the line" and therefore be included above and beyond current funding limits.

9. What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Standard guidance for reporting data reuse, and community recognition might help with attribution and credit. For examples, effective reuse of data -either in reusing data from others or having one's own data reused by others - might be considered as a positive factors during grant reviews.

11. What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Scientific ontologies such as those curated by the OBO Foundry and the NCBO provide well-defined semantic models that encourage interoperability.
The use of these ontologies to publish scientific data as linked open data should be encouraged.

13. What policies, practices, and standards are needed to support linking between publications and associated data?

Unique identifiers for both publications and data sets, along with tools for using those identifiers, might be used in combination with linked open data on both publications and datasets, to support this linkage.

Harry Hochheiser
University of Pittsburgh
Department of Biomedical Informatics

January 12, 2012

Office of Science and Technology Policy (OSTP)
Executive Office of the President
725 17th Street Room 5228
Washington, DC 20502

RE: Request for Information: Public Access to Digital Data Resulting From Federally Funded Research 76 Fed. Reg. 70176, November 10, 2011.

To the Office of Science and Technology Policy:

AVAC welcomes this opportunity to comment on the recent RFI: *Public Access to Digital Data Resulting From Federally Funded Research*. AVAC is a non-profit organization that uses education, policy analysis, advocacy and a network of global collaborations to accelerate the ethical research and development and global delivery of vaccines, male circumcision, microbicides, pre-exposure prophylaxis (PrEP) and other emerging HIV prevention options as part of a comprehensive response to the AIDS pandemic. AVAC believes that the road to safe, efficacious, accessible and affordable HIV prevention options for all who need them will necessitate IP and data sharing for the public good and also effective management of data and materials produced from disparate sources but generated for a potentially common purpose IP collaboration is a critical part of AVAC's mission,[1] and we commend the OSTP for addressing digital data sharing under the America COMPETES Reauthorization Act of 2010 (ACRA).

The public's need to know the complete results of research it pays for is clear and immediate. Immediate access to public or consortia managed access to data that has undergone quality assurance is justified by the benefits which accrue from increasing comparability of results, delivery of information supporting cross disciplinary approaches and promoting cost savings in complex research. Digital data sharing and management are a long over-due means to enhance collaboration in research and speed the translation of scientific advances into quality, affordable health care. Since tax dollars underwrite this work, all Americans should benefit from the broader utilization of digital data and its use in scientific research. Policies also need to be developed to ensure that digital data are collected and stored in formats that are accessible and interpretable by others.

Our responses note the several questions to which they relate.

[1] See *Intellectual Property at the Crossroads* in AIDS Vaccines at the Crossroads, AVAC Report 2005, *http://avac.org/pdf/reports/2005_Chapter4.pdf* and *Data and Materials: A "to-do" list for the future, AVAC Report, 2010,* http://www.avac.org/ht/a/GetDocumentAction/i/28317.

Advocacy to accelerate ethical research and global delivery of AIDS vaccines and other HIV prevention options

423 West 127th Street
4th Floor
New York, NY 10027
USA

T +1 212 796 6423
F +1 646 365 3452
E avac@avac.org
W www.avac.org

1. *What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

AVAC recognizes the potential benefits in productivity to HIV prevention research arising from greater access to digital data generated through Federal funded grants and contracts. Unequal or restricted access to information by researchers creates inefficiencies in scientific research. As a recent article found "the observed differences in access levels between institutions suggest an un-level playing field, in which some researchers have to spend more efforts than others to obtain the same information."[2] Eliminating this un-level playing field will facilitate increased productivity. Even minor increases in research efficiency can have significant effect on the United States economy. With the United State's investment in gross expenditure on research and development at $312.5 billion in 2007 and assuming social returns to R&D of 50%, a 5% increase in access and efficiency would have been worth $16 billion. [3]

Significant obstacles to digital data sharing exist even if greater access to data is required under ACRA. As noted in the recent National Science Foundation report:

> Successful digital research data sharing and management plans depend, in part, on adequate consideration of funding, resources, and structural issues that may either facilitate or impede acceptance and implementation. These plans are especially important for small research institutions and research grants that may not have the resources available to share and manage long-lived data. Thus, just as a single data sharing and management policy will not apply to all research communities, a one-size-fits-all business model will not apply to all institutions and awards.[4]

AVAC recommends that OSTP solicit suggestions on digital data sharing from a variety of stakeholders. We would like to share insights gained from the HIV vaccine research field. The Global HIV Vaccine Enterprise is an international alliance of more than 30 independent research, funding, advocacy and stakeholder organizations and governments, engaged in unprecedented collaboration to speed the development of a safe and effective HIV vaccine.

---

[2] Voronin Y, Myrzahmetov A, Bernstein A, *Access to Scientific Publications: The Scientist's Perspective.* PLoS One. 2011;6 (11):e27868. Epub 2011 Nov 17.

[3] Houghton J.W. and Sheehan, P.J. (2006) *The Economic Impact of Enhanced Access to Scientific Publications,* Centre for Strategic Economic Studies, Working Paper, No 23, Victoria University, Melbourne. (http://eprints.vu.edu.au/archive/00000472/).

[4] National Science Foundation, *Digital Research Data Sharing and Management* December 2011 (Task Force on Data Policies Committee on Strategy and Budget National Science Board)

The Enterprise convened meetings on management of the large data sets including deep sequencing of HIV virus populations, as well as B-cell and T-cell repertoires.

The Enterprise meetings have confirmed first that digital data store requires significant investment of time and resources, in some cases estimated to include storage requirements in the range of petabytes. Vaccine researchers have also identified gaps in the subject matter content of available databases that are critical to the field.[5] Federal policies that support development of web-based systems and assemble field expert panels to identify priority digital data needs can increase the efficiencies and use of such data. NIH should accommodate expert outside panel recommendations when transitioning previous activities from the now (recently) defunct National Center for Research Resources.

Federal policies must address the significant non-IP related challenges. Obstacles identified by the Enterprise working group include: 1) training to develop human expertise to input, catalogue and design data bases; 2) establishing procedures for depositing and curating data bases; and 3) addressing privacy issues related to patient information.

We also recommend invigorating existing data and resource sharing plan policies to yield substantive results. Currently, NIH grant policy requires submittal of data and resource sharing plans within applications.[6] In practice, however, application content in many grants is minimal, lacking in detail or extremely modest in promise, if any. NIH may invigorate this element by elevating the importance of data and resource sharing plans within applications during award scoring, requiring content focused on appreciation for field wide benefit of use, and mandating substantive content directed towards identifying those posting and forum opportunities where a grantee's results can have best translational benefit.

2. *What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?*

AVAC does not assume that greater access would, in fact, harm or diminish the IP of those that initially develop digital data. IP discussions, in our view, focus too exclusively on mere ownership without sufficiently discussing responsible IP management or use for public purposes. Useful models of IP management and use for public good are available, for

---

[5] An open-ended plea for the development of a global database of HIV vaccine responses Wilkinson, Peter; Filali-Mouhim, Abdelali; Li, Shuzhao; Ahlers, Jeffrey; Schatzle, John; Pulendran, Bali; Sekaly, Rafick-Pierre; Cameron, Mark J. *Current Opinion in HIV & AIDS*, POST AUTHOR CORRECTIONS, 21 November 2011 doi: 10.1097/COH.0b013e32834e390a

[6] NIH Data Sharing Policy http://grants.nih.gov/grants/policy/data_sharing

example, such as the regulatory implementation schemes of the California Stem Cell Bond Act program. [7]

3. ***What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?***

There are strong currents in the present research and product development models which favor unilateral research and which inhibit data sharing. To overcome these barriers, in the HIV field, for example, the AIDS Clinical Trial Group has developed a system for equitable allocation of secondary uses of data in papers.[8].

4. How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

AVAC supports continued participation and follow-up to the Board of Research Data and Information, National Academy of Sciences, efforts to develop an internationally supported Microbial Research Commons.[9] Future efforts should also include participation by public advocacy stakeholders to help design content, public use content and allocations of rights.

Finally, we note that the RFI's description of stakeholders - "e.g., research communities, universities, research institutions, libraries, scientific publishers" – does not include patients, advocacy knowledge users and organizations devoted to promoting research initiatives. AVAC hopes the RFI implementation will place these interests firmly within the outcomes results of this effort.

Thank you again for this opportunity to comment. If you have questions about this letter, please contact me at mitchell@avac.org.

Sincerely yours,

Mitchell Warren
Executive Director

---

[7] California Institute for Regenerative Medicine http://www.cirm.ca.gov/Regulations
[8] ACTG *Publication and Disclosure of Study Results SOP 111* available at https://actgnetwork.org/node/430
[9] National Research Council. *Designing the Microbial Research Commons: Proceedings of an International Workshop*. Washington, DC: The National Academies Press, 2011.

Thu 1/12/2012 10:24 AM
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Name/Email
Ahmet Erdemir
Affiliation/Organization
Computational Biomodeling (CoBi) Core
Department of Biomedical Engineering
Lerner Research Institute
Cleveland Clinic

City, State
Cleveland, OH

Comment 1

Relevant to

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Data from federally funded research should be publicly accessible for use and for further dissemination. Ideally if one goes to a federal agency's research reporting tools, e.g. NIH RePORTER, http://projectreporter.nih.gov/reporter.cfm, a search should not only provide project details but also links to data/models generated by the project. For example, one can easily access list of publications generated by the project and if uploaded in PubMed Central, the publication itself. Why not have a link to data that is used in that publication? Or, links to other data that is not necessarily associated with a publication but still a result of that project? There are many federal or institutional resources to upload data, where longevity is more warranted than an investigator initiated website. For example, in our computational biomedical research we utilize SimTk (https://simtk.org), which is provided by the National Center for Biomedical Computing at Stanford.

Comment 2

Relevant to

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

While many federal agencies has initiated various data and resource sharing requirements, these does not seem to be enforced adequately. The way and the timing of the data dissemination is still at the discretion of the funded investigator. As a tax payer and a fellow investigator, I feel that when I read about the results of a federally funded publication, I should be able to contact the authors to access the

data as well. If authors cannot provide me the data, they should provide me their data/resource sharing plan (which is part of NIH submissions) that explicity dictates why they cannot share data.

Comment 3

Relevant to

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

It is certainly possible that some investigators may not want to share data due to potential benefits from intellectual property. Nonetheless, during the submission of a federal grant, this intention should explicity be provided and a cost sharing plan to accommodate lack of public sharing to data should be outlined. When full public access of data is planned and executed, cost sharing should not be necessary.

Comment 4

Relevant to

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Investigators should be rewarded for their efforts to provide their data in a usable form. These efforts take time and resources.

Investigators are hard pressed to spend their time to write grants and to write peer-reviewed articles. The former is dictated by the desire to keep a sustaining research program and the latter is the traditional way for knowledge dissemination and the way investigators prove themselves to their institutes that they are productive. Sharing data should have the same priority as these. A while back, to accommodate this cultural thinking, we discussed on the possibility to establish a Journal for Dissemination, (see, http://www.imagwiki.nibib.nih.gov/mediawiki/index.php?title=Journal_for_Dissemination). We have also addressed how the dissemination can be rated. In addition, we provided a tabularized comparison of Data/Model/Software dissemination methods.

Comment 5

Relevant to

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Time to dissemination and extent of data sharing is very likely to be different between disciplines; nonetheless, the principal goal should remain the same; data collected through federal funds, unless justified rigorously and supported by cost sharing, should be publicly accessible.

A relevant discipline related problem has been summarized for the patenting system and prospective reform of it (see Schacht, 2007, Patent Reform: Issues in the Biomedical and Software Industries, Congressional Research Service Report RL33367). For example, computational sciences evolve very rapidly. A long lasting ambargo to access data/models/software will likely hinder innovations by upcoming generations to innovate.

Comment 6

Relevant to

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Another concept to consider in federally funded research is open development. Commonly accepted norm of doing science in the US is that, you write a grant, you do the work in your lab, you publish, and if you are generous you disseminate your data/models/software afterwards. Why not investigators, when writing a grant, are encouraged to write a plan for open science and have plans for including other stakeholders to help decision making during the research process. Open development practice by definition will dictate transparency and require early and frequent dissemination as well.

Comment 7

Relevant to

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?Show citation box

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

The National Science and Technology Council's Interagency Working Group on Digital Data are encouraged to browse through multiscale modeling community's discussions on model and data sharing at http://www.imagwiki.nibib.nih.gov/mediawiki/index.php?title=Model_Sharing_Working_Group, http://www.imagwiki.nibib.nih.gov/mediawiki/index.php?title=Data_Sharing_Working_Group, http://www.imagwiki.nibib.nih.gov/mediawiki/index.php?title=Working_Group_10

Comment 8

Relevant to

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Standards are necessary for seamless exchange of data. Nonetheless, the priorities to accomplish public data access should be: 1) Share the data (so that it can be accessed), 2) Document the data (so that it can be used), 3) Conform to data standards (so that it can be exchanged easily). From an investigators perspective, as one goes from item 1 to item 3, the workload increases heavily. Along those lines, and to address item 2 for our specific discipline, we have provided reporting considerations for complicated biomechanical models (see http://www.imagwiki.nibib.nih.gov/mediawiki/index.php?title=Reporting_in_FEA/JB_Edition)

Comment 9

Relevant to

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Publication system is archaic and slow and is not always public access. The way to communicate scientific results need to adapt the speed of our evolution of social communication networks. This will require the review system and publication companies to adapt new approaches. It may be interesting to explore why a system like Wikipedia (http://www.wikipedia.org) should not work for scientific communication.

Comment 10

Relevant to

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Search for data should be easy. If Google or any other search engines can find the data, so does any interested parties. Innovations for filtering search results from vast variety of data sources are necessities.

Comment 11

Relevant to

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There seems to be a wide range of innovation and business opportunities to establish federal and institutional repositories, to assemble relational databases among them, and to provide tools that will facilitate the investigators to annotate and publish their data (not just the knowledge acquired from the data).

Comment 12

Relevant to

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

A scholarly publication only provides a knowledgebase. Reproducibility, reusability, and therefore accountability are not warranted with the lack of underlying data. Some disciplines has established requirements for uploading of data with a publication. This good practice can be generalized quickly if federal agency policies prescribe provision of data with the publication.

==========

Cleveland Clinic is ranked one of the top hospitals in America by U.S.News & World Report (2010). Visit us online at http://www.clevelandclinic.org for a complete listing of our services, staff and locations.

COGR
an organization of research universities

COUNCIL ON GOVERNMENTAL RELATIONS
1200 New York Avenue, N.W., Suite 750, Washington, D.C. 20005
(202) 289-6655/(202) 289-6698 (FAX)

January 12, 2012

Interagency Working Group on Digital Data
National Science and Technology Council
Office of Science and Technology Policy

SUBJECT: Request for Information: Public Access to Digital Data Resulting from Federally Funded Research

The Council on Governmental Relations (COGR) is an association of 188 research universities and their affiliated academic medical centers and research institutes. COGR concerns itself with the influence of federal regulations, policies and practices on the performance of research conducted by its member institutions. Our goal is to ensure that federal policy goals can be met in an effective and efficient manner without creating administrative structures that may hinder compliance.

COGR offered written and testimonial comment to the National Institutes of Health (NIH) as it developed its policy to enable public access to NIH-funded research. As we noted in our January 2010 response to the Office of Science and Technology Policy's (OSTP) request for information concerning Federal government-wide public access policies, we support the goal of providing timely and less costly access to data that result from federally funded research and observed that the public does have such access through various depositories.

This new request for information raises a series of specific questions but we want to begin by introducing a question that may burden any efforts on the part of OSTP to coordinate Federal agencies' policies concerning the long-term stewardship of research data. The meaning of "data" as used in Federal regulations and policies has become increasingly ambiguous. There does not exist a common definition among Federal agencies, and the definitions used by agencies do not always reflect the meaning applied within the research community, which itself does not have a uniform definition. The rights and responsibilities surrounding ownership, access to and retention of data will be affected by the variety of meanings ascribed to "data." Frequently, the term "research data" is confused with what are, by definition, research materials.

Generally, we would argue that research data consists of information that provides a quantitative and/or qualitative description or characterization. This definition is consistent with the Office of Management and Budget's (OMB) Circular A-110, *Uniform Administrative*

Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations, definition "as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." And yet, various agencies provide a broader definition.

For example, NIH defines "data" as "recorded information, regardless of the form or medium on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data." Some would argue this definition embraces "research materials."

For much of the research community research materials are those materials from which data can be extracted. Materials are tangible or physical objects, e.g., writings like a database, cells, molecules, designs, plans, forms, flow charts, planets, plants, and/or animals. Thus, in making the distinction between research *data* and research *materials,* it's important to distinguish between the entities *containing* the data and the data *themselves*. For example, a lab notebook, a recording, or an insect are not data but contain data or represent entities about which data (description or characterization) can be created.

All Federal agency policies and regulations do not employ a similar distinction between data and materials. As noted above, NIH's definition of research data includes materials such as data files, which are recorded but in most cases will not provide a quantitative or qualitative description or characterization in and of itself. The Federal Acquisition Regulations (FARs) that provide general terms and conditions for Federal contracts includes computer software and software documentation in its definition of "data;" the Defense contract regulations (DFARS) reference "technical data" which includes computer software documentation but not the software programs or source codes. The Environmental Protection Agency (EPA) goes further by carving out "raw data" to include "laboratory worksheets, memoranda, notes or exact copies thereof, that are the result(s) of original observations and activities of a study and are necessary for the reconstruction and evaluation of the report of that study" (40 CFR Part 742).

For the research community under OMB's Circular A-110, the OMB definition which applies across Federal agencies may provide the most useful general framework for discussing the access to and retention of research data. In this definition, preliminary or "raw" data or research materials without analysis should not be included for the purposes of access by the general public.

However, if the OSTP truly intends to harness digital data for public access, the challenges are even greater. If the intent of the OSTP's proposal is to provide greater public access to "digital data" then the consequences for research institutions is a significant financial and administrative burden. In its 2009 report, "Harnessing the Power of Digital Data" the Working Group on Digital Data broadly defined digital data to encompass:

> . . . born digital and digitized data produced by, in the custody of, or controlled by federal agencies, or as a result of research funded by those agencies . . . . [including] the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses.

As with the distinction between "data" and "materials," it is important to distinguish between the different types of "digital data" generated by different disciplines into those that will be of use to other scientists, e.g. genomic data, long-term climate data, versus digital data, or perhaps more accurately "digital materials" that really is of no use except to the originating scientist or investigator because it requires access to notebooks and other information to analyze, laboratory-level digital data. In some disciplines given the nature of the equipment and processes used in experimentation, "digital data" encompasses essentially all the original data because virtually all primary data is collected in a digital format. The latter can be massive and is retained by the investigator as support for publications, etc., versus the former such as genomic or climate data sets that will be needed for many years by a variety of scientists and investigators.

If required by Federal agencies to retain laboratory-level digital data or materials, institutions would need to develop extensive management systems that could store the huge diversity of digital data that our researchers develop, remembering that most laboratory-level "digital data" would include all the original data we collect every day. The growing number of cooperative data repositories for digital data with a larger, more global interest to other scientists helps institutions meet their responsibilities for data access and data sharing with the scientific public that can make good use of the data. Yet, these repositories come with challenges as well including maintaining the integrity and security of the data housed in the repositories.

The RFI addresses "digital data" albeit reading through questions many of them use the more generic term "data" or "research data." It is not clear whether the OSTP is making a distinction between digital data and other types of data or not. Thus, we return to our original concern over the meaning of "data" as used in Federal regulations and policies and, as noted, in this RFI. The rights and responsibilities and, perhaps more challenging, the costs and burdens associated with providing public access surrounding ownership, access to and retention of data will be affected by the variety of meanings ascribed to "data."

With these overriding concerns, we offer the following responses to some of the questions.

Preservation, Discoverability, and Access

Comment 1: Growing markets related to access and analysis and using those markets to grow the economy and improve productivity of the scientific enterprise.

> Data resulting from research are the foundation for the continuing dialogue among scientists that advances our scientific enterprise and productivity. Research data, whatever the format, serve as the source for inventions, publications, and can with sufficient support, serve as the origin for expending existing and creating new businesses.
>
> The individual most capable and most interested in using the data is the person who created it. It would be a mistake to jeopardize the ability of that individual scientist or investigator to exploit the potential of the data.
>
> Data derived from Federally supported basic and applied research are data that can be intriguing to potential investors. But without further investigations and directed proofs of the concepts suggested by the basic research, the ability of the general public to use the data is limited. Like access to research publications, access to research data will not provide a direct, uninterrupted link to a new business or activity without significant investment.

Comment 2: Protection of Intellectual Property

Unlike patentable inventions, where the Bayh-Dole Act and implementing regulations provide a uniform Federal framework, there is no single source of authority on ownership and protection of data resulting from Federally funded research. Generally the research agencies that provide Federal financial assistance allow recipients to copyright and own data developed under the award, subject to the right of the agency to use the work for Federal purposes and, as appropriate, subject to specific requirements like the NIH and National Science Foundation's data sharing policies. Under Federal contracts, the government may allow copyright, but normally the government retains the ability to exercise all the rights of the owner, e.g., distributing copies to the public. The FAR provides that universities and colleges may claim copyright in data developed under a contract for basic and applied research that they perform solely. This provision may be a serious constraint in the current funding climate that otherwise encourages teaming arrangements and public-private partnerships. The DFARS makes no special provision for educational institutions.

Institutions have been aggressive in identifying research activities with the potential to stimulate economic development and work to patent and license the intellectual property to the benefit of the business partners and, ultimately, the public. The ability to protect the intellectual property is what attracts businesses to make the investment in time and resources to license the technology and bring products to the market. Public access may not serve these purposes, nor may the approach followed in the various Federal acquisition regulations.

OSTP notes the examples of NIH and the National Science Foundation policies concerning data management and data sharing. These policies work because the focus of access and sharing is the scientific community. The repositories established by various agencies, notably NIH, are built on carefully crafted policies and procedures that protect the intellectual property rights associated with the data. Any agency contemplating policies or regulations must ensure similar protections.

Comment 3: Differences in Disciplines and Data

It is precisely these differences in disciplines and data that pose a significant challenge to data management and preservation and, we would add, the establishment of standards for interoperability, reuse and repurposing. Rather than building separate, prescriptive policies similar but different across agencies and disciplines, OSTP should advocate a simple policy requiring federal grantees and contractors to enable the transfer of research data in a manner that meets the specific disciplines and data. In this way, those who understand the data can establish a framework for its transfer to others with appropriate contractual protections including privacy, confidentiality, etc. Institutions can transfer information on how the data were collected, what unique tools are needed for analysis, etc. This type of approach avoids the inevitable pitfalls in attempting to create a single set of standards applicable across disciplines, data and agency.

Comment 6: Real Costs of Preservation and Access

For grantee institution, the costs of preserving and sharing are significant. As a general rule, most institutions rely on the investigator to help meet its obligations to preserve and share research data. The creation of national repositories in some disciplines has helped in the

management of access to research data. Charging reasonable fees for the duplication of research data can not cover the full and significant costs associated with long-term preservation.

Standards for Interoperability, Reuse and Repurposing

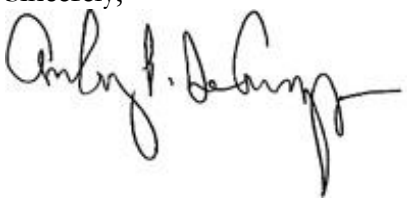
Comment 10: Digital Standards

As noted above, we would caution OSTP and the Federal agencies to avoid establishing standards for data preservation and transfer that penalizes the investigator in terms of the loss of first use and requires a significant investment in re-formatting of data to meet a national standard. Establishing a general principle that requires transfer or sharing of data within a reasonable time frame will ensure that the data is available to other interested scientists and investigators while preserving appropriate protections.

Thank you for this opportunity to comment on OSTP's continuing consideration of the value of public access to peer-reviewed publications. We would note that the America COMPETES Reauthorization Act of 2010 (PL 111-358) Sec. 103 does not assume that a single, government-wide policy is appropriate and charges the Interagency Public Access Committee with coordinating agency activities concerning access to publications and data.

COGR has long supported harmonization and coordination among the Federal agencies in order to streamline the compliance with Federal mandates and regulations. In the case of access to data, we would suggest that setting a principle of long-term preservation and access without prescribing the institutional approach is the wisest course. As OSTP observes, disciplines have begun defining standards that meet the needs of the research community. Relying on those standards and efforts of institutions to transfer new, innovative technologies to the marketplace has led to extraordinary economic developments to date.

Sincerely,



Anthony P. DeCrappeo
President

American Statistical Association

Promoting the Practice and Profession of Statistics

732 North Washington Street, Alexandria, Virginia 22314 USA
(703) 684-1221 ● Fax: (703) 683-2307 ● Email: asainfo@amstat.org
Web site: http://www.amstat.org/

January 12, 2012

Interagency Working Group on Digital Data
National Science and Technology Council
Office of Science and Technology Policy
Executive Office of the President
Washington, DC 20502

Dear Working Group Members,

The American Statistical Association and its Committee on Privacy and Confidentiality appreciate the opportunity to offer the attached comments on the request for information, "Public Access to Digital Data Resulting from Federally Funded Scientific Research."

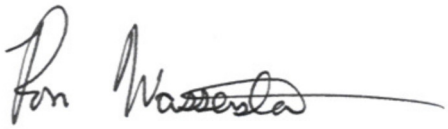
As background, the American Statistical Association (ASA) is the world's largest statistical society, with over 18,000 members in some 90 countries (though most are in the United States). One of its core missions is to advise government on matters related to data-centric research and policy-making. The Committee on Privacy and Confidentiality is an appointed group of ASA members with expertise in the technical methods and policy issues related to data access and confidentiality. The Committee members who endorse the comments in this letter includes:

Jacob Bournazian, MA, JD	Energy Information Administration
Julia Lane, PhD	Committee Chair
Krishnamurty Muralidhar, PhD	University of Kentucky
Aleksandra Slavkovic, PhD	Pennsylvania State University
Lance Waller, PhD	Emory University
Simon Woodcock, PhD	Simon Fraser University

The Committee, and the ASA more broadly, would be delighted to share our expertise with the Interagency Working Group on Digital Data.

Sincerely,

Julia Lane
Chair, ASA Committee on Privacy and Confidentiality



Ron Wasserstein
Executive Director, ASA

We applaud the OSTP initiative to enhance the availability and preservation of digital data. Such access is consistent with the principle of reproducible research, which we believe is vitally important to the scientific and policy-making process. Such access also opens opportunities for new insights based on existing data.

Preservation of data is equally important, since it is impossible to know what currently collected data will useful for future generations of scientific inquiry.

We know that the OSTP recognizes the importance of preserving data subjects' confidentiality in making digital data access policy. The American Statistical Association's Privacy and Confidentiality subcommittee wishes to offer the following general suggestions on confidentiality issues related to the OSTP initiative.

1. Research has shown that many data subjects are skeptical of government agencies' abilities to protect confidentiality. A policy that greatly increases possibilities for confidentiality breaches could severely weaken the ability of both researchers and of government agencies to collect data. Researchers are subject matter experts in their own area of research, not in data dissemination and confidentiality protection hence the development of sound policy requires input from experts in these areas. The statistical community can contribute to ensuring both that data confidentiality is protected and that research subjects understand what is being done to protect the confidentiality of the data.

2. Researchers in confidentiality protection methods often distinguish two general classes of methods. Restricted access limits who can use the data, for example via secure data enclaves, remote access, or licensing.

   Restricted data limits what data are made available, for example via data suppression, aggregation, swapping, or simulation. Each approach has a purpose. Restricted access is arguably the best solution for purposes of reproducible research and data preservation. It is also the best approach for complex data (e.g., relational or high dimensional data), which are difficult to protect adequately without degrading the usefulness of the data.

3. For restricted access, recent cyberinfrastructure advances have led to the development of data repositories that are managed by national data producers -or contracted to non-government parties--which investigators can use for data dissemination and preservation. These repositories can be staffed by data dissemination professionals and advised from committees of experts on data confidentiality practice.

   Such remote access systems have been developed at the NORC/University of Chicago data enclave in the US, the Secure Data Service in the UK, the Microdata Online Access (MONA) system in Sweden, and the remote online system at Statistics Netherlands. These enable users to analyze the data at their own computers; however, users cannot save or print data locally.

   When coupled with vetting and education of data users, as well as penalties for misuse, such systems can provide access while minimizing risks.

4. For restricted data, the ASA subcommittee has commented on recent HIPAA revisions to recommend against the adoption of safe harbor type standards. Put simply, each dataset has unique disclosure risks that cannot be captured with a list of common prohibited identifiers.

The subcommittee also wants to provide specific feedback on the following questions.

Question 1) What specific federal policies would encourage public access to and the preservation of broadly valuable digital data...

Users need a centralized location for searching available data bases. Websites such as Data.gov are possible choices where federal agencies can place research data that is publicly available from their agency website. The metadata supporting the files would be standardized and there could be a single path through the "Raw Data" for viewing what digital data may be accessed. Also, federal agencies need to reach consensus on standardized coding to use in the front section of the URLs that link to digital data so that researchers can easily identify digital research data files and can direct their search queries appropriately.

Federal agencies should consider reaching consensus on some criteria for setting time limits on the sensitivity of certain categories of information when products or procedures are no longer produced or applied, or structural changes occur within an industry due to changes in regulations or the application of new technology. Not all information that is protected at the time of collection needs to remain protected for decades into the future.

Question 2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal Agencies, and other stakeholders, with respect to any existing or proposed policies...

Authors of published research should provide a proprietary notice on all work that is publicly released. Federal agencies should provide researchers with the capability to claim copyright protection in their work that is federal funded. Many times, federal agencies receive FOIA requests from the public for this information as a method for gaining a copy of the research and the releasing agency cannot impose or enforce any protections for the author if the authors do not claim any proprietary protection.

UNIVERSITY LIBRARIES

January 12, 2012

Office of Science and Technology Policy
The White House

Re: Request for Information: Public Access to Peer-Reviewed Scholarly Publications Resulting from Federally Funded Research

The University of Maryland Libraries write in response to the Request for Information, published in the *Federal Register* on November 4, 2011, by the Office of Scientific and Technology Policy regarding public access to peer-reviewed scholarly publications resulting from federally funded research. We appreciate this opportunity to comment.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

UMD Answer: NIH Public Access policy (NOT-OD-05-022) and the NSF Data Management Plan requirement.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

UMD Answer: A variety of steps can be taken to protect the intellectual property interests of data creators. Some examples might include allowing for the ability to "embargo" works, much as we do with theses and dissertations, for a set amount of time. We can design preservation architecture that allow for permissions management. In addition, researchers should have some leeway in how they package the data that they wish to have publicly disseminated. For example, they may choose to make a subset of the data fully available, but withhold other elements of the data that while not crucial to the final results, could require future users to come up with their own methods of display and analysis. In some cases, however, having a deadline by which data must be made publicly available can actually serve as an incentive to researchers to publish their most critical findings first and in a more timely manner.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

UMD Answer: Federal agencies need to be flexible and design preservation systems that can manage diverse types of data. They also need to be highly aware of the costs and size of the data. Some scientific research can generate hundreds of terabytes of data in a short amount of time and on a regular basis, while others, only a few gigabytes over long periods of time. First, researchers should not necessarily feel that they must restrict their research or computations based solely on size of data, but second, accounting for larger datasets is something that should be taken into account when asking for funding.

Since presentation of and access to data from various disciplines vary widely and since customizing access for various purposes requires expertise about the data depending on the discipline from which use cases develop, it might be best for Federal agencies to limit their services to presentation and access of whole datasets in their raw form. Instead, Federal agencies can enable, through crowdsourcing tools and other means, the user communities to develop special and customized presentation and access services.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

UMD Answer: It is tricky to talk about the costs of long-term stewardship and dissemination without somehow placing restrictive quotas or penalizing those disciplines that naturally create large datasets of digital data. It is most important that agencies first recognize that there are differences in costs of different types of data, and then second perhaps conduct some focus groups and studies about the usefulness of different types of data to future researchers. Much of this discussion has to come from within the disciplines and with the data creators themselves. The big question is an appraisal one – what of the data needs to be retained? Of 100 terabytes of astronomical data, perhaps only 10% is useful or worth retaining. But without the other 90%, it may take a future researcher an incredible amount of time to use the data or reconstruct an experiment. Packaging data for reuse and preservation may be an extremely time-intensive process and something that should be accounted for in the costs when assigning grants, as well as when writing data management plans.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

UMD Answer: Stakeholders can best contribute to implementation of data management plans by working together to provide guidelines, templates and services to better prepare data for long-term preservation. By working with the data creators from the beginning of their research, and by collecting relevant information to enable future access and use of the data, libraries, research institutions, etc. can ensure that they have the correct tools in place to fulfill the data management plans. In addition, stakeholders can provide staffing, in the form of data scientists, archivists, etc. who can assist in the process of preparing the data for final deposit/dissemination.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

UMD Answer: Allow or require that the data management plans included in grants contain a line-item in the budget to account for start-up and maintenance costs of digital data.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

UMD Answer: Registries would be one way to measure and verify compliance with Federal data stewardship and access policies for scientific research. The caveat is that registries are often dependent on self-selection and user submission. Adding a level of prestige or legitimacy to inclusion in a registry would encourage more active participation. Peer review is another mechanism to improve compliance by having panels of peers review and monitor, by audits and tests, compliance with policies. In addition, more direct involvement in the compliance process from agencies (e.g. National Archives and Records Administration, Library of Congress) for whom data stewardship is an important part of their mission and existing skill set, would help greatly in legitimizing the verification process. In general, use of public data over time by peers will serve as a natural audit mechanism, as well. Data that is important to a group of researchers will be used heavily and with scrutiny. Any failure in preserving the data will be flagged by users. It is important that Federal and other stakeholders monitor and act on the flags, to help recover lost data and in cases where data is not recoverable, learn from lessons over time and reduce risk of data loss.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

UMD Answer: In order to stimulate innovative use of publicly accessible research data in new and existing markets, agencies could devote more time and funding to communicating availability of the data and provide, as part of the communication, an effective description of the available data. Stimulus grants, such as the National Science Foundation's Digging into Data Program, would help to jump start the process.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

UMD Answer: There are currently no standard and reliable mechanisms to ensure this happens with non-digital data. The best mechanism is to instruct researchers on how to properly cite and document their work. Use technology (like watermarks for images) to ensure that data can always be traced to its source. Most technologies do not lend themselves well to the entire array of scientific data. The various disciplines likely have some way of ensuring this, and the most useful thing would be for the Federal policies to document all of this, so that it becomes very clear how secondary use of data is cited.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

UMD Answer: This is an area where librarians can step in and help facilitate a process. Issues to address would be determining common metadata, and minimum metadata required. In cases where each discipline has solutions for these problems, librarians can develop processes and mechanisms to ensure that access and preservation systems can interpret various types of data standards, rather than requiring all to fit within a narrow scheme.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

UMD Answer: The library and archives professions have created a host of long-lasting and effective standards over the years: MARC, Dublin Core, Resource Description and Access (RDA), OAI-PMH, Encoded Archival Description (EAD), to name a few. What makes these standards successful is widespread adoption by their communities, extensive documentation on usage, their ability to set "minimums" to enable compliance at various levels, their interoperability, their appropriateness to the material that they are describing, and metadata creation tools that allow for consistency and uniformity.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

UMD Answer: Some existing programs that can help with this are the National Digital Information Infrastructure Program (NDIIP), and international bodies such as Digital Preservation Europe, who are already working with international communities within each discipline. Another example is the program to design a Microbial Research Commons as

described in a recent BRDI (Board on Research Data and Information) and BLS (Board on Life Sciences) symposium.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

UMD Answer: Embrace and encourage open access. Standards for minimal set(s) of metadata for the datasets that have clear semantics and are normalized, at least within a discipline, bibliographic and keyword or subject heading standards for the publications that are compatible with the dataset metadata standards.

The Association for Research in Vision and Ophthalmology
1801 Rockville Pike, Suite 400 ▪ Rockville, Maryland 20852-5622
arvo@arvo.org ▪ +1.240.221.2900 (Tel) ▪ +1.240.221.0370 (Fax)
www.arvo.org

January 12, 2012

Response to OSTP **Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research**

On behalf of The Association for Research in Vision and Ophthalmology (ARVO), I submit the following comments in response to the RFI issued on November 3, 2011. ARVO is the largest and most respected eye and vision research organization in the world. Our members include more than 12,600 researchers from over 80 countries. ARVO encourages and assists research, training, publication and knowledge-sharing in vision and ophthalmology. ARVO publishes two medical/scientific research journals which are published online only and are hosted at HighWire Press which is considered by libraries internationally as a trusted site and archive. In mid-2012 ARVO will launch a new online-only journal on the topic of translational ophthalmic science & technology, which will also be hosted at a trusted site. In addition, ARVO voluntarily deposits complete articles of all NIH-funded research published in its journals in PubMed Central on behalf of authors and at no charge to the authors.

ARVO supports the principle of providing the public with access to the federally funded scientific research. However, we believe that releasing the peer-reviewed research articles in direct competition with scholarly publishers undermines the ability of associations and societies to maintain the high quality standards of selection, review, production, publication and protection of the scientific record.

Scholarly publishers provide essential services that ensure the quality and integrity of journal content. Through peer review publishers and the scientific community identify scientific shortcomings and inadequacies which continues through the revision and re-review of articles. Over 50% as for some journals as much as 75% of submitted articles are ultimately rejected because of these inadequacies. The continuous feedback to authors through review and editing immeasurably improves the final published product. Publishers also serve as guardians of scientific ethics and standards to ensure accuracy, reliability, ethical treatment of patients and humane treatment of animal subjects.

In addition, in our opinion, the current NIH policy confuses the community and the public regarding the completeness of the "public" record and who the actual publisher of the scientific material is. NIH has established itself in direct competition with private publishers while using public taxpayers' funds to complete their redundant work. These activities jeopardize the financial viability of journals, particularly those published by learned societies and associations

that are dependent on subscription revenue and author charges to sustain their journals and educational activities.

ARVO appreciates your consideration of our responses below to the specific questions posed in the RFI regarding ***Preservation, Discoverability, and Access*** **and** ***Standards for Interoperability, Re-Use and Re-Purposing.***

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

RESPONSE: Since most association and non-profit publishers in all scientific and medical areas already offer free and open access to ALL content, whether or not it is federally funded, anywhere from three (3) months to twelve (12) months after publication and usually provide open access to the abstracts immediately upon publication, ARVO does not believe that additional policies would improve the access or dissemination of information. In addition, by requiring federally funded research to be available earlier only presents 30% to 50% of the published research, thus presenting an incomplete resource on any given topic. Increasing funding for scientific research would better serve the public, the economy, and researchers. Like most association and non-profit publishers, especially those hosted at trusted sites, ARVO also participates in the LOCKSS (Lots Of Copies Keep Stuff Safe) which permits any interested libraries to deep archive all journal content and update those files for future use if deemed appropriate. It would better serve the public to cease requiring deposits of published articles in a taxpayer funded site such as PubMed Central (PMC), which is redundant and by all reports from the National Library of Medicine costs over $2,500 per article to create and maintain, and allows indiscriminant distribution of all content.

It should be noted that private sector publishers represent over 30,000 workers in the U.S.; spend millions of dollars to provide peer-review, editorial support, and production and distribution of over 45% of the scientific peer-reviewed articles published each year for researchers around the world.

(2) What specific steps can be taken to protect the intellectual property interest of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

RESPONSE: As mentioned in Response to Question (1), publishers should be permitted to control the access and distribution of their content and are currently providing wide access within one year of publication, which often includes all back issues, as is the case with ARVO. We provide free and open access to all content published from Volume 1, Issue 1, page 1 of our journals through all information published up to 6 months ago. We are investigating the possibility of watermarking all content to ensure that the copyrighted content cannot be indiscriminately copied, reused, and redistributed without providing appropriate credit and attribution to the authors and publisher. This would apply to all supplementary data that is published with an article. Our intent is to protect the integrity

of the research and ensure proper accreditation of the research. In addition, agencies should consider supporting America's Research Act, H.R. 3699.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

RESPONSE: Regarding online only publications in the medical area, PubMed Central has recently advised published that ALL content, regardless of whether it is federally funded, must be deposited and made freely available when ready through PMC. This, in our opinion is counter to the Digital Millennium Copyright Act and US Copyright Act, Title 17. It seems reasonable that Federal or federally trusted, perpetually maintained data repositories could be established to store data tables or genetic/genome information databases in which all authors could deposit such content. Medical and scientific associations and non-profit publishers are already providing trusted sites and archives for all of their content and these facts should be considered before requiring additional federal resources be used to establish a redundant system of managing data. In these days of world-wide discoverability through the use of robust search engines, associations and non-profit publishers are already meeting data management and discoverability needs.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

RESPONSE: ARVO suggests that as a first step, agencies should evaluate **existing** trusted source resources that currently exist and examine the practices for access, dissemination, and preservation and that are currently being paid for by publishers. These sites would include HighWire Press, Allen Press online, and such organizations as the American Geophysical Union. Establishing any federal repository that would include data already stored in such a site would be redundant and an unnecessary use of taxpayer funds that could be used for supporting additional research.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

RESPONSE: If stakeholders participated in already existing plans, such as LOCKSS, CLOCKS, and other deep archives, as well as agree on requirements for qualification of trusted hosting sites then all could begin public discussions of data management standards.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

RESPONSE: First, and foremost, an evaluation of the costs incurred in maintaining already established archives and deep archives should be undertaken. ARVO suggests that organizations such as HighWire Press, The Stanford Libraries, and OCLC's activities, as well as others such as Portico, should be reviewed and considered. These organizations have already committed to digital preservation and perpetual access. Their costs could be used as benchmarks for current and projected costs to best address funding levels and mechanisms.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

RESPONSE: If a centralized list of trusted hosting and perpetual archiving resources were developed and associated with recognized scientific journals and publications (books, meetings abstracts, etc.) then authors could report where their research is published and be in compliance with any and all funding institutions. This would minimize cost of compliance and verification. Since publications lists are an integral part of grant applications this would further consistent reporting of compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

RESPONSE: Encouraging the continued development of affordable semantic search and discovery tools and supporting the publishers in the use of these tools would provide stimulus to a variety of markets and industries. It is important to support innovative U.S. companies in these any other technology areas.

(9) What mechanisms could be developed to assure that those who produce the data are given appropriate attribution and credit when secondary results are reported?

RESPONSE: One method would be to support the use of Digital Object Identifiers (DOIs) for all elements and parts of all published content and embedding of DOIs in figures, tables, and content and parts of content. Publishers already use DOIs and deposit the information with groups such as CrossRef. Whenever a DOI is searched or used it is associated with the original citation. If DOIs were used universally by indexers and abstracters and checked verified by publishers during the publishing process and use a reverse lookup at regular intervals after publication, original publication information of results would be readily apparent and available. Note that over 1,200 publishers, libraries, and organizations world-wide are members of CrossRef and already support the use of DOIs. The organizations also agree on standards for use and presentation of DOIs for all facets of scholarly publishing. CrossRef also participates in establishing NISO standards for use.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29,371) is an example of a community-driven data standards effort.

RESPONSE: ARVO has no definite response to this question at this time other than allowing each scientific community to develop its own standards.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

RESPONSE: ARVO has no definite response to this question at this time but agrees with the need for standards but recognizes the need for flexibility in the standards because of ever-changing technologies and discoveries.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

RESPONSE: One method of promoting effective coordination internationally is to work with and through international specialty organizations such as ARVO. Many associations and scientific societies, including ARVO, have members worldwide who represent government agencies and institutions in their home countries. This would be a very effective means of developing good lines of cooperation and communication. Our organizations continually promote and support international collaboration in research.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

RESPONSE: Please refer to earlier discussion of DOIs in Question #9.

Thank you for the opportunity to respond to this RFI issued by OSTP.

Submitted by: Karen Schools Colson
ARVO, Director, Publishing Projects
On behalf of The Association for Research in Vision and Ophthalmology
1801 Rockville Pike, Suite 400
Rockville, MD 20852

ACS Submission to the Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Research

FR Doc. 2011-28621

Submitted January 12, 2012
by:
John P. Ochs
j_ochs@acs.org
Vice President, Strategic Planning and Analysis
American Chemical Society
Publications Division

The American Chemical Society (ACS) is the world's largest scientific society with more than 164,000 members. ACS advances knowledge and research through scholarly publishing, scientific conferences, information resources for education and business, and professional development efforts. The ACS also plays a leadership role in educating and communicating with public audiences—citizens, students, public leaders, and others—about the important role that chemistry plays in identifying new solutions, improving public health, protecting the environment, and contributing to the economy.

ACS Publications is a division of the American Chemical Society. The Publications Division strives to provide its members and the worldwide scientific community with a comprehensive collection, in any medium, of high-quality information products and services that advance the practice of the chemical and related sciences. Currently, over 40 peer-reviewed journals and magazines are published or co-published by the Publications Division. Over 290,000 pages of research material are published annually, representing over 37,000 research papers. With the introduction of the ACS Journal Archives in 2002 and the C&EN Archives in 2011, we provide searchable access to over one million original chemistry articles dating back to 1879.

ACS Publications offers both sponsored and author-enabled open access to research articles through our ACS Author Choice and ACS Articles on Request programs. In addition, digital data that supports the findings of articles and bibliographic information, including abstracts of research articles, are freely available on our website. Since the beginning of the transition to electronic publishing in the mid- to late-1990s, we have developed, and are continuing to develop, innovative and accessible business models, policies, and practices to support the scholarly communication process and broaden information access.

As a socially responsible organization deeply rooted in the scholarly community, we share the interest of the Federal government in maximizing the dissemination and discoverability of knowledge. ACS believes that success in this area will hinge on these efforts being sustainable for publishers over the long-term. We welcome for the opportunity to respond to the invitation to contribute to the Request for Information (RFI) on Public Access to Digital Data Resulting from Federally Funded Scientific Research published by Office of Science and Technology Policy (OSTP) in the *Federal Register* on November 4, 2011.

Our response is in two parts: first a summary of our overall comments and recommendations, and second, answers to the specific questions posed in the RFI.

I. Summary

ACS supports the view that Federal agencies should work with researchers and other stakeholders to create appropriate policies to make digital data resulting from federally funded scientific research freely available to the public. ACS sees an appropriate role for governmental and other funding agencies to identify standards and best practices for the management of primary scientific data that are generated via taxpayer or other research grant funding that supports independent investigators. This governmental role could also include standards for the interoperability of data repositories with the published research literature. As part of this process, agencies should investigate and establish contacts where appropriate with a number of initiatives already underway or recently concluded that are examining data stewardship issues.

Within the context of standards and best practices that have been identified, the Federal government can develop effective, evidence-based policies to enhance public access to and preservation of digital data. We recommend that these policies be established in collaboration with researchers and other key stakeholders.

Grants should earmark specific funds to support researcher data management and deposit activities. The amount should be determined in collaboration with representative bodies of key stakeholders who are involved in the data preservation and deposit process. It may need to vary with discipline. In parallel with this activity, the government should ask the General Accounting Office to undertake a study of existing federal data archives to determine the full costs required for start-up and ongoing access, preservation, and migration of data depositories.

Federal policies should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. They should provide for the establishment of security protocols that protect stored data from unauthorized modification, damage or deletion and liability arrangements if data is lost or affected. Key policy terms should be defined and policies should take into account that there are differences between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets').

Federal intellectual property policies should recognize that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost which, in certain circumstances, the host should be entitled to recover. Databases themselves – i.e. collections of data specifically organized and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – merit intellectual property protection, under copyright or database protection principles.

To reduce legal uncertainty for data users and producers, federal policy should give clear direction as to what data may be shared publicly – e.g. no personal data related to volunteer subjects. Penalties for the misuse or abuse of data should be established, such as grant bans for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity.

Policies should not require researchers to fund the establishment or maintenance of data archives nor should they be required to pay submission fees for deposit. Federal policy should encourage, but not require researchers to supply their data when submitting manuscripts to scientific journals. This is because certain forms of publication, e.g. letters and other short communications, act as early alerts to results of potential interest and the requirement to supply data can add a burden that slows scholarly communication to the detriment of all.

Policies could create an incentives hierarchy for scientists to share their data, with the greatest reward for those who publish data with articles and short communications but also recognition for those who publish data only.

II Response to RFI Questions

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Before specific policies are adopted, ACS sees an appropriate role for governmental and other funding agencies to identify standards and best practices for the management of primary scientific data that are generated via taxpayer or other research grant funding that supports independent investigators – e.g. recommendations for best practices in the PARSE.Insight report *Insight into Digital Preservation of Research Output in Europe*. This governmental role could also include standards for the interoperability of data repositories with the published research literature.

Once standards and best practices have been identified, the Federal government will be in a stronger position to adopt effective, evidence-based policies to enhance public access to and preservation of digital data. We recommend that these policies be established in collaboration with researchers and other key stakeholders.

Policies should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. They should provide for the establishment of security protocols that protect stored data from unauthorized modification, damage or deletion and liability arrangements if data is lost or affected. Key policy terms such as data and data integrity should be defined since, for example, there are differences between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets').

To reduce legal uncertainty for data users and producers, clear direction should be given as to what data may be shared publicly – e.g. no personal data related to volunteer subjects. Penalties for the misuse or abuse of data should be established, such as grant bans for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity.

Policies should not require researchers to fund the establishment or maintenance of data archives nor should they be required to pay submission fees for deposit. Federal policy should encourage, but not require researchers to supply their data when submitting manuscripts to scientific journals. This is because certain forms of publication, e.g. letters and other rapid communication formats, act as early alerts to results of potential interest and the requirement to supply data can add a burden that slows scholarly communication to the detriment of all.

Policies could create an incentives hierarchy for scientists to share their data, with the greatest reward for those who publish data with articles and short communications but also recognition for those who publish data only – i.e. with no discussion, analysis, or interpretation of such material.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Policies adopted by the federal government should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. Penalties should be established for the misuse or abuse of data, e.g. bans on grant eligibility for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity. Key policy terms such as data and data integrity should be clearly defined to differentiate between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets'). To reduce legal uncertainty for data users and producers, clear direction should be given as to what data may be shared publicly – e.g. no personal data related to volunteer subjects.

The ACS endorses the view that researcher-validated primary data should be made freely available but federal intellectual property policies should recognize that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost which, in certain circumstances, the host should be entitled to recover. Databases themselves – i.e. collections of data specifically organized and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – merit intellectual property protection, under copyright or database protection principles. Such databases are often characterized by the sophistication of their data field structuring, searchability tools, and contain valuable and useful information for scholarly research. The value of researcher validated data sets and individual data points is different from specific databases that have been organized and compiled to serve particular research needs.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The ACS supports the position that researcher-validated primary data should be made freely available and that Federal agencies should work with the scientific community and other stakeholders to create appropriate policies that reflect different standards currently in use or commonly accepted. If no consensus emerges from such efforts, ACS believes that the government has an appropriate role in working with key stakeholders such as researchers and publishers to develop best practices that will advance scholarly communication and the public good.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Agencies should investigate and establish contacts where appropriate with a number of initiatives already underway, or recently concluded, which are examining data stewardship issues. These include:

- Opportunities for Data Exchange (ODE, www.ode-project.eu), whose aim is to gather and promote best practices around the way scientific data are treated. Its *Report on*

Integration of Data and Publications is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/

- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), a project of the Alliance for Permanent Access which includes over thirty research institutes, national libraries, IT providers and research funders working together to create a Network-of-Excellence on digital preservation
- PARSE.insight (http://www.parse-insight.eu/), who developed a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The *Insight into Digital Preservation of Research Output* report is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf and the *Science Data Infrastructure Roadmap* is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf
- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science (ICSU) working on an initiative for a World Data System
- DataCite (http://datacite.org/), convening members of the datasets community to collaboratively address the challenges of making research data visible and accessible, and
- NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental), looking at policy and technical issues surrounding the definition, publication and linking of journal articles and supplemental materials, including data, as well as archiving, preservation and migration of different file formats.

Interaction with these and other initiatives should give Federal agencies a good base from which to estimate the relative costs and benefits of long-term stewardship and dissemination of different types of data. Agencies may also find the data sections of the *Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (available at http://brtf.sdsc.edu/) to be relevant in evaluating the relative costs and benefits of long-term data preservation and migration.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Keys to successfully implementing data management plans from the Federal government, and other funders, include the following:

- Requirements for data management plans should clear, complete and unambiguous. They should specifically address liability issues
- Data management policies established in collaboration with researchers and other stakeholders such as publishers
- They should take into account the practices of different research communities and be developed in collaboration with representative bodies of all stakeholders who will likely be affected – e.g. researchers, funders, publishers, universities, data repositories, etc.
- FAQs, training courses, and e-learning modules should be available for researchers to gain a more complete understanding of data management plan requirements as well as the data deposit process

- Grant funds should be earmarked to support data management and deposit activities
- Incentives to deposit, such as the possibility for receiving research credit for data deposit, should be provided as well as penalties, like grant bans, for noncompliance after a clearly-defined and collaboratively-set time frame
- Data deposit, integrity, provenance, and access at repositories should be fast, efficient and clear.
- Data repositories should be certified and audited to foster trust. Researchers should not be required to maintain the accuracy or integrity of the data once it has been deposited but depositing researchers should have the right to modify or correct data they have deposited. Liability
- The administrative burden on researchers should be kept to the barest minimum possible

Stakeholders should work collaboratively on these issues since more than one stakeholder can contribute to each. There is no one stakeholder that has, or should have, a monopoly on any of these activities.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Grants should earmark specific funds to support researcher data management and deposit activities. The amount should be determined in collaboration with representative bodies of key stakeholders who are involved in the data preservation and deposit process and may vary with discipline. In parallel with this activity, the government should ask the General Accounting Office to undertake a study of existing federal data archives to determine the full costs required for start-up and ongoing access, preservation, and migration of data depositories. Agencies could also investigate the Open Archive Information System (OAIS) Reference Model (ISO standard 14721:2003, available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683), used by many as a model for building a sustainable digital archive. Last, agencies note the assessment of funding models from the *Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (available at http://brtf.sdsc.edu/):

"*There is no single "best" funding model for digital preservation. Selection of an appropriate model requires an in-depth knowledge of the circumstances surrounding the effort, preservation goals, the stakeholder community, and so on.*" (p. 44)

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

As in other areas related to preservation where significant activity is already underway, the federal government could establish relationships with groups like the ISO *Repository Audit and Certification Working Group* (see http://wiki.digitalrepositoryauditandcertification.org/bin/view) to learn about standards and best practices already in development. Once standards and best practices have been identified, the Federal government will be in a stronger position to adopt effective, evidence-based measures related to the assessment of compliance with its data stewardship policies.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

In addition to the measures already discussed in previous questions, agencies could set aside funds to promote to use of the data depositories or develop special sections of their websites promoting the availability and characteristics of the data they hold.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Agencies should seek collaborations with DataCite (see http://datacite.org/), a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and re-purposed for future study. DataCite is currently active in supporting researchers by helping them to find, identify, and cite research datasets with confidence; supporting data centers by providing persistent identifiers for datasets, workflows and standards for data publication; and support journal publishers by enabling research articles to be linked to the underlying data. They are currently working primarily with organizations that host data, such as data centers and libraries.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The PARSE.insight *Science Data Infrastructure Roadmap* (available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf) notes the following initiatives that may prove to be use examples of digital stewardship: CASPAR (http://www.casparpreserves.eu/), Planets(http://www.planetsproject.eu/), DCC (http://www.dcc.ac.uk/), OAIS (http://public.ccsds.org/publications/archive/650x0b1.pdf), SHAMAN (http://www.shaman-ip.eu/), and nestor (http://www.langzeitarchivierung.de/)

Also the Technical Working Group of the NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental) is preparing an initial draft of its recommendations. The narrative form is expected to contain a table outlining the minimum metadata elements recommended to describe supplemental materials and establish their relationship to the main article, as well as a more detailed discussion of optional elements to more comprehensively characterize the materials for future applications. A non-normative DTD is also expected in draft form. This DTD, once finalized, will not be an official standard. Rather it will be a model to more precisely define a hierarchy for the recommended metadata, and could be used as a starting point for organizations seeking to adhere to the NISO/NFAIS recommendations.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Most standards development in the field of digital data stewardship is either ongoing or prospective. However, the initiatives cited in the answer to question 4 above are good examples of active projects in this area and are reproduced below for ease of reference:

- Opportunities for Data Exchange (ODE, www.ode-project.eu), whose aim is to gather and promote best practices around the way scientific data are treated. Its *Report on Integration of Data and Publications* is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/
- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), a project of the Alliance for Permanent Access which includes over thirty research institutes, national libraries, IT providers and research funders working together to create a Network-of-Excellence on digital preservation
- PARSE.insight (http://www.parse-insight.eu/), who developed a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The *Insight into Digital Preservation of Research Output* report is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf and the *Science Data Infrastructure Roadmap* is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf
- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science (ICSU) working on an initiative for a World Data System
- DataCite (http://datacite.org/), convening members of the datasets community to collaboratively address the challenges of making research data visible and accessible, and
- NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental), looking at policy and technical issues surrounding the definition, publication and linking of journal articles and supplemental materials, including data, as well as archiving, preservation and migration of different file formats.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies should join the international community of organizations already actively involved in the development of digital preservation standards, best practices and policies that have been cited in answers to questions 4 and 11.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Agencies should become involved with three initiatives already well underway in this area:

- Opportunities for Data Exchange (ODE, www.ode-project.eu) – whose aim is to gather and promote best practices around the way scientific data are treated. See its *Report on Integration of Data and Publications* available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/
- The NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental) which is preparing an initial draft of its recommendations. The narrative form is expected to contain a table outlining the minimum metadata elements recommended to describe supplemental materials and establish their relationship to the main article, as well as a more detailed discussion of optional elements to more comprehensively characterize the materials for future applications. A non-normative DTD is also expected in draft form. This DTD, once finalized, will not be an official standard. Rather it will be a model to more precisely define a hierarchy for the recommended metadata, and could be used as a starting point for organizations seeking to adhere to the NISO/NFAIS recommendations
- DataCite (http://datacite.org/), a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and re-purposed for future study. DataCite is currently active in supporting researchers by helping them to find, identify, and cite research datasets with confidence; supporting data centers by providing persistent identifiers for datasets, workflows and standards for data publication; and support journal publishers by enabling research articles to be linked to the underlying data. They are currently working primarily with organizations that host data, such as data centers and libraries.

Brian M. Bot

Sage Bionetworks - 1100 Fairview Ave N • Seattle, WA 98109
E-Mail: brian.bot@sagebase.org Web: www.sagebase.org

Date: January 12, 2012

To: Office of Science and Technology (OSTP), on behalf of
National Science and Technology Council (NSTC)
digitaldata@ostp.gov

Re: Request for Information:
Public Access to Digital Data Resulting From Federally Funded Scientific Research

Open access to biomedical research is essential. It will provide an increased return on the federal government's investment and ultimately will result in improved patient care.

From the onset of my career as a biostatistician, I was sheltered from the struggles plaguing biomedical research. Fortunate to have my first job in the innovative and intellectually stimulating atmosphere of the Mayo Clinic, my perspective on the public access conversation was limited to the tiresome albeit valid argument as to whom "owns" data resulting from federally funded research. This debate merely scratches the surface of the issue of public access to research data.

Current infrastructure is not equipped to handle the data deluge that is omnipresent in the research setting. Even mighty research institutions are being forced to rethink the way data is managed. Startups, innovators, and changemakers don't stand a chance.

Now biomedical research has reached its tipping point.

The research community has benefited from open access to digital data with the Human Genome Project (HGP) providing a forward thinking model of public access to valuable research data. The guiding principles set out in the Bermuda Accord undoubtedly avoided much duplication of efforts and thus years of wasted time, effort, and resources.

Other examples such as the Gene Expression Omnibus (GEO), the Database of Genotypes and Phenotypes (dbGaP), and The Cancer Genome Atlas (TCGA) provide value as well, but are flawed by usability and accessibility issues.

Arguably, half-baked public data resources are worse than not making the data available at all. Usable data in standardized formats should be required, not just the act of making data available. Governance inhibiting access of "public" research data presents unnecessary obstacles for researchers trying to reuse data already made available.

Data alone is never enough. Metadata describing an experiment from beginning to end is as important as the raw data itself. Data without links to clinical characteristics make it impossible to build predictive models of disease. Data without experimental confounders can bias model building and confuse downstream interpretation. I would like to see a push for not only public access to digital data, but also public access to full experimental information. The distinction is subtle, but I feel necessary for the true value of open access to be realized.

The logistics of making research data, especially genomic data, widely consumable is no small task. Instituting policies that provide infrastructure, or incentivize the building of infrastructure, can help ease the burden. By building a genomic information superhighway, great opportunity become available for a burgeoning sector of the economy dealing with digital data stewardship.

It is evident that policies seen as optional or mandatory (but loosely enforced) will not be prioritized. Motives of the Individual do not align with the greater good. Individual researchers have many more reasons to conceal their research data than to make it broadly available and useful.

Sadly, scientists are a stubborn bunch. The scientific culture is stale, and simple policy shifts could help to transform it in the blink of an eye. The very scientists living on the bleeding edge of innovation in their respective niche fields ironically lag in the technological advances that allow other sectors to flourish. Transparency, crowdsourcing, and meta-analytics have proven useful but have not been adopted by the masses. Open access policies will force researchers down these advantageous paths.

It is my hope that the spirit of the HGP and Bermuda Accord will continue to live on through policy statements within the National Institute of Health (NIH), but an appropriate incentive structure to encourage those ideological policies is severely lacking. Current incentive structures for career scientists, especially in academia, do not reward altruistic endeavors. There is no place on a CV for having provided readily accessible, usable, and valuable data for the community at large. Measures such as publication rate, grant funding levels, and ultimately tenure are no longer good surrogates for a researcher's impact.

Great opportunities are being missed. The longer public access to digital data is delayed, the less opportunity our creative free market system is able to take advantage of the wealth of taxpayer-funded resources that are currently locked up. Lowering the barriers of entry to usable research data is a proven catalyst for innovation and entrepreneurship. If we can further encourage this type of environment through policy changes, our next generation of researchers and the scientific community as a whole can reap the benefits. Imagine the collaborative atmosphere and exponential return on investment for federal research funding that would be possible if we succeed.

It is my hope that publicly informed and expert driven policy choices can shape this promising revolution. The opportunity is present to create a synergistic movement that points all researchers, public and private, towards common goals that will benefit the greater good and ultimately more rapid progress toward important scientific advances.

Regards,

Brian M. Bot

The Ornithological Council

PROVIDING SCIENTIFIC INFORMATION ABOUT BIRDS

American Ornithologists' Union

Association of Field Ornithologists

CIPAMEX (Sociedad para el Estudio y Conservación de las Aves en México)

Cooper Ornithological Society

Neotropical Ornithological Society

Pacific Seabird Group

Raptor Research Foundation

Society for the Conservation and Study of Caribbean Birds

Society of Canadian Ornithologists/ Société de Ornithologistes du Canada

The Waterbird Society

Wilson Ornithological Society

Ellen Paul
Executive Director
5107 Sentinel Drive
Bethesda, MD 20816
Phone (301) 986-8568
Email: ellen.paul@verizon.net

12 January 2012

Ted Wackler
Deputy Chief of Staff
Office of Science and Technology Policy
Attn: Open Government
725 17th Street, NW.
Washington, DC 20502

Submitted via e-mail to digitaldata@ostp.gov

Dear Mr. Wackler,

The Ornithological Council, a consortium of twelve scientific ornithological societies in the Western Hemisphere, submits these comments in response to the request by the Office of Science and Technology Policy (OSTP) for input on the Administration's interest in enhancing public access to digital data generated in federally funded research.

Ornithology is rich in data that are underutilized because they are not accessible. Decades of data are disappearing rapidly and irretrievably because the scientists who collected the data had no opportunity to archive it in a physical or electronic form. Whether on paper or in some kind of electronic medium, datasets collected over the past century could contribute greatly to our knowledge of avian biology.

Our organization strongly supports the concept of archiving and sharing these data. We have investigated and discussed the possibility of developing an archive for the types of data generated in ornithological research, but found that the cost is prohibitive and that it might not be realistic to expect that scientists will voluntarily undertake the somewhat burdensome effort of learning metadata standards and routinely labeling their data for deposit into an archive.

As a preliminary and key issue, we stress the need to allow researchers to have exclusive access to and use of their data for reasonable time after the grant period has ended, so as to allow them to complete their publications. The "reward system" for scientists in both academia and in federal agencies stresses publications. The number and quality of publications is a large factor in determining promotion and tenure, and also strongly affects the researcher's success in obtaining grant funding. We assume that OSTP is fully aware of the fact that the misappropriation of a researcher's data could have substantial negative impacts on the researcher's career and will take care to assure that any public access policy includes ample protections for the researcher.

As a second key issue, we would like to address something that seems to be outside the scope of the OSTP request and existing agency data management requirements, probably because it would be impossible to impose these requirements retroactively. We would like to stress that if resources are available, the government should commit those resources to help "stabilize" those data, convert them to a digital format, and submit them to appropriate data repositories. The data collected a decade ago or a century ago are, in our field, at least as valuable as the data collected today, if not more so, as these baselines are necessary to assess change. The attics full of paper, note cards, field notes; the offices full of punch cards, floppy disks, and magnetic tape – all need proper storage to guard against physical loss and all should be digitized and contributed to publicly accessible repositories. We cite the example of the North American Bird Phenology Program created by the Patuxent Wildlife Research Center of the U.S. Geological Survey. Using volunteers and a high-speed scanner, this remarkable program preserved six million hand-written note cards recording bird migration observations, dating back to 1881. The scanned records were then uploaded to the internet to make it possible for volunteers to enter the data into a database. The USGS and the other partners of the National Phenology Network provide analytical tools, guidance documents, and other resources. More recently, the U.S. Bird Banding Lab was able to stabilize decades of hand-written records by scanning, and it is hoped that funds will be made available to make these critical data available to researchers by digitizing the data and making them available on a public-access website. To date, researchers and others have been able to access these data only by making a request to Banding Lab staff, who would then retrieve the physical records for copying and mailing. The records were at extreme risk of physical deterioration or loss, having been stored in a variety of facilities that were subject to rodent infestation, fire, dampness, and flooding.

Therefore, we strongly encourage OSTP to work with the Office of Management and Budget to provide funding and direction to the agencies to stabilize existing physical data records, to digitize those records, and make them available on publicly accessible databases. These processes should not be limited to agency-held data but should be opened to private researchers as well.

We would also like to address certain of the questions asked by OSTP, as follows:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Response: The key issue here is funding. Developing and maintaining these systems is costly. The intricacy involved in creating any one metadata standard is substantial. Interoperability is a daunting challenge. In our discipline, for instance, DataOne <www.dataone.org> is intended to "ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE will transcend domain boundaries and make biological data available from the genome to the ecosystem; make environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; provide secure and long-term preservation and access; and engage scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualizations." The five-year NSF grant alone amounts to $15,257,190 from the Office of Cyber Infrastructure and it is supplemented by support from the NSF Computer and Information Science and Engineering Directorate (CISE) Pathways Computational Sustainability, the NSF INTEROP Programs, NASA, the Leon Levy Foundation, the Moore Foundation and (until its recent demise), the National Biological Information Infrastructure of the U.S. Geological Survey.

There is already ample evidence that federal funding does result in the development of successful data repositories. Federal funding was largely responsible for the development of a suite of taxonomic databases – ORNIS for birds, MANIS for mammals, HERPNET for herps, and FISHNET for fishes – each a distributed database and all interoperable, mappable, and publicly available.

The complexity of these systems requires that they be done right; if not, the end result is a system that hampers, rather than facilitates public access. The federal government must be willing to commit the resources to enable excellence or the undertaking is not worthwhile. We would have an expensive warehouse where nothing can be found, much less retrieved.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Response: Assure that researchers have a reasonable time after the completion of data collection or the end of the grant period, whichever is later, to publish the results before making the data publicly accessible. There may be some situations that merit a longer period of exclusive access. In some fields, research may extend over decades. For instance, studies of long-lived organisms will typically continue over the full life-cycle of the organism and possibly over several generations. A researcher will likely publish papers throughout this period, but later papers will often make use of data collected at a much earlier stage of the study. A set of criteria to determine when such extensions are appropriate is needed.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Response: Consult with the professional societies. We can provide the data and insight as to the policies and practices that will make it possible for our members to archive and share data without jeopardizing their intellectual property interests. We can also provide information about the ability of our discipline to create and maintain these repositories and the appropriate metadata standards. We can identify gaps in opportunities for data management. In ornithology, the existing repositories, though stellar, simply cannot accommodate many kinds of data collected by ornithologists. We have, as a result of the NSF data management plan, been collecting information about all potential data repositories that may be suitable for this kind of data, and we are still finding significant gaps. At the moment, NSF's data management website simply directs those who are unable to find an appropriate public repository to "Contact the cognizant NSF Program Officer for assistance in this situation." We suspect that if NSF were to attempt to compile a comprehensive list of relevant data repositories, these gaps would be quite evident. Further, while it may be that among all the existing repositories, a researcher could find suitable repositories for some parts of the data in a given dataset, it is not reasonable to expect a researcher to have to submit data to two or more different datasets, particularly as it is possible that the two datasets may not use the same metadata standard.

We can also compile and provide data about the range and median grant size in our discipline. This information should be taken into account before imposing another time-consuming grant requirement

on researchers. The OSTP notice mentions that the NIH requirement applies only to grants with direct costs exceeding $500,000 in a single year. In our discipline, that threshold would exclude most grants. For instance, the average grant size made by the NSF BIO program in 2011 was $149,238. In 2010, it was $140,064 <http://dellweb.bfa.nsf.gov/awdfr3/default.asp>. Most NSF grants in our discipline come from the Division of Environmental Biology (DEB) or the Division of Integrative and Organismal Systems (IOS). In DEB, the average grant in 2010 was $95,649 and in 2011, it had declined to $85,919. In IOS, the average grant size was $150,000 in 2010 and $151,181 in 2011. Smaller grants simply do not allow the researcher to hire administrative staffers or other technicians to handle this additional work.

If no additional funding is provided, the data management requirements could constitute an unfunded mandate such as would trigger the provisions of 2 U.S.C. §1501. We recognize that the Administrative Procedure Act exempts matters ""relating to agency management or personnel or to public property, loans, grants, benefits or contracts" and that therefore, a formal rulemaking as would trigger the Unfunded Mandates Reform Act (UMRA) would likely not occur. Nonetheless, the agencies have made it a practice to use notice-and-comment procedures outside the Federal Register process for this and other policy matters. These quasi-rulemakings should be regarded, for the purpose of the required UMRA analyses, as the equivalent of a rulemaking. Therefore, any agency that wishes to mandate data management should be required to conduct an "UMRA-like" analysis to assure that the requirements are the least costly, least burdensome, or most cost-effective option that achieves the objectives of the rule, or explain why the agency did not make such a choice (2 U.S.C. §1535).

The scientific community should also be consulted with regard to the release of certain types of data. For instance, we have long been concerned about the potential online, public access release of location information associated with bird banding. Most of the birds banded are legally protected at the federal or state level. Information about the location of banding could facilitate activity that is prohibited under the Endangered Species Act. Other species, protected only under the Migratory Bird Treaty Act, are very vulnerable to disturbance during the breeding period. If the public could use the location data associated with bird banding to determine breeding locations, the disturbance resulting from human presence could lead to failed breeding attempts. The same concerns would pertain to location data of other animals or plants protected under the Endangered Species Act, and to other animals that are vulnerable to disturbance, should location data be made available. Even species that are not endangered (whether or not legally protected but that are in commercial demand could be over-exploited and small populations could be driven to extinction by over-collecting. In cases such as these, the researcher should be permitted to omit, obscure, or coarsen the location data.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Response: It is not clear that this can or should be done. Suppose that the number of queries or data retrievals were reported by each repository? That information would not tell us if or how the data were used, and of course, the determination of the value (benefit) of that use is subjective and not comparable across disciplines.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Engage the scientific community full in every stage of the development of data management plans. The NSF data management plan requirements, flexible though they may be at this time, seem to have been developed by the National Science Board without any, or any significant, input from the scientific community. There were two workshops – one in 2003 and another in 2004. At the first, only two hours were allotted to discussion; at the second, only 45 minutes. A relatively few public comments – most from other federal agencies, data management firms and professionals, and only a few from researchers, research institutions, or scientific societies – were received in response to a 2005 request for comments. Between the task force recommendation in the 2005 report and the actual development of the NSF data management plan requirement that went into effect in January 2011, there seems to have been no opportunity for input from the scientific community.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Response: As noted above, grants in our field typically do not permit researchers to hire staff to undertake the work associated with effective metadata labeling and deposit of data. There is no point in warehousing data if it is not done in such a way as to make the data easily retrievable and to assure that subsequent users are able to identify the characteristics of those data so they can determine if they are appropriate for the later use. Without additional funding, data repositories are not likely to be of adequate quality and any resources devoted to them will have been wasted.

This is not a hypothetical concern. The U.S. Geological Survey devoted more than a decade of effort to develop the National Biological Information Infrastructure. It is now being dismantled; it never began to approach the original goal of providing access to distributed data, but for the support afforded to efforts such as VertNet.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Training is probably the key to improving compliance. The existing biological data profile and the numerous metadata entry tools that exist are not, in many aspects, intuitive. It is likely that scientists who have not had training will struggle to use these systems and will either give up entirely or will not enter complete information. Training is likely to reduce the barrier to use of the metadata entry tools.

Verification could be a step in grant close-out.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

For a number of years, we have discussed this very question with regard to the potential release of bird banding data. It has been the practice of the Banding Lab to interact with those who request data and to remind them of the professional standards for attribution and credit. This interaction is possible only because data requests are made by individual contact to a staffer who then transmits the data to the requester. In fact, the Banding Lab website makes no mention of these professional standards. The

U.S. Bird Banding Lab Advisory committee could not devise a more robust solution, saying that a web-based public access site should be developed "...in consultation with banders and users of banding data, review and revise the current policy for use of banding data, and require all data users to agree to this policy. The BBL should also encourage the adoption of this policy by ornithological societies and scientific journals as part of their scientific code of ethics."

The reality is that there is no effective mechanism to force users to give appropriate attribution and credit. It may be evident, given the age of the data or the geographical or temporal range of the data that the author did not collect all the data used in the paper. In those cases, editors will likely insist that the author provide attributions. However, there will be many cases where this is no evidence that the data used were collected by other than the author, and in those cases, there is really no adequate solution. However, the use of the data in a subsequent analysis is the purpose and benefit of public access; the lack of attribution is not a sufficient reason to curtail access. The real value of the data to the original researcher is that researcher's own publications; the unattributed use of the data in a subsequent analysis does not diminish the value of the original publication or of the use of the data for that original publication.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

In our discipline (the taxonomic sciences), extensive effort has gone into the development of a metadata standard known as the Darwin Core. Numerous extensions have been developed that will support the addition of "ancillary" data such as ecological conditions, and weather data. We hope that there will someday be extensions for the behavioral data that is commonly collected in ornithological research.

The use of this common metadata standard and extensions would permit interoperability with any other system that uses the same standards. For instance, the Darwin Core has led to the development of ORNIS, HerpNet, MANIS, and FishNET (birds, herps, mammals, and fishes) and these are integrated with GEOLocate, AmphibiaWeb, Map of Life, Specify, Arctos, DataONE, Encyclopedia of Life, and Animal Diversity Web.

These repositories and the metadata standards were initiated by the community and achieved with federal funding. Other organizations (most also federally funded) then built user tools and applications, such as the Avian Knowledge Network at the Cornell Lab of Ornithology. This project also received significant federal funding.

However, no amount of scientific zeal and energy can achieve this kind of result without significant federal funding. Unless the federal government is willing to continue to devote appreciable sums, the government and the public cannot expect to achieve the goal of providing public assess to data derived from federally funded research. The termination of NBII may also result in the termination of funding for the single coordinator position and a single programmer position for VertNet. The participating

institutions are all suffering from the economic downturn and cannot readily replace the funding for these two positions.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Response: Science knows no geopolitical boundaries. Scientists have long been working on an international basis to develop metadata standards. The Global Biodiversity Information Facility, established in 2001, already holds 8,594 datasets to which access is free and unrestricted. However, the sole U.S. representative to GBIF is a single employee of the now-terminated National Biological Information Infrastructure. The NBII termination page states with regard to GBIF that "While USGS does anticipate continued collaboration with some of these activities, we have yet to determine at what level this will occur." We are informed that it is likely that USGS will continue to participate at the minimal level (i.e., one FTE) that was the case prior to the termination of the NBII.

The federal agencies must commit to increased participation in these international bodies, and commit the necessary resources for that participation.

If the federal government is unable or unwilling to continue funding this activity at an adequate level, then it should hold in abeyance any mandate that scientists submit data to any repository. If there is no assurance that the repositories will persist and will be properly managed, and that there will be a continued development of science-driven metadata standards, then the burden imposed on scientists to label their data and submit to data repositories is not warranted.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Response: The Global Biodiversity Information Facility has developed a protocol for the use of universal identifiers that can be used to refer to a range of digital objects in data sets, documents, and repositories. The single identifier would allow the user to retrieve all digital objects (such as datasets) associated with a publication or, conversely, all publications associated with a dataset. The use of a universal identifier also facilitates the tracking of digital object retrieval and, should that item be used in a subsequent publication, could also help determine the extent to which that information was actually used.

We thank the OSTP for considering our concerns and views, and hope that this response will prove helpful in shaping federal policy on public access to digital data.

Sincerely,

Ellen Paul
Executive Director

January 12, 2012

RFI: OSTP, Public Access to Digital Data Resulting From Federally Funded Scientific Research
Contact: Francis P. McManamon fpmcmanamon@asu.edu
Center for Digital Antiquity [Arizona State University]
Tempe, Arizona

Questions and Comments:

Preservation, Discoverability, and Access:

1) What specific Federal policies would encourage public access to and the preservation of valuable digital data resulting from federally funded scientific research, to grow the US economy and improve the productivity of the American scientific enterprise?

We suggest several specific Federal policies that would encourage preservation and access to data resulting from federally funded scientific research. These include:

(a) Require that data generated from federally funded research be archived in an appropriate trusted digital repository or archive dedicated to the preservation of and access to data and supporting documentation. Part of this requirement should include the creation of appropriate and sufficient metadata for discovery, so that data are not simply preserved, but also readily accessed and interpreted for future research. Regarding the kinds of digital repositories most appropriate for data archiving, we suggest that discipline-specific repositories provide a rich context of similar materials so that users, as they search, are provided with search results tailored to their expectations and needs. Metadata in disciplinary repositories contains phrases, keywords, and categories that match subject matter domains, making search results much more targeted to information resources that are especially useful. In addition, specialized digital repositories, as opposed to institutions or organizations with generalized missions, can increase efficiency and productivity of tasks related to digital archiving, and minimize the costs of data access and preservation to researchers and traditional archives maintained by libraries, museums, and universities.

(b) Encourage data repositories to include in their archive any documentation relevant to the original data set. For example, repositories should include reports related to the data. Repositories also should ensure appropriate linkage to metadata, which would ease searches among related data and make background research more efficient.

DIGITAL ANTIQUITY
School of Human Evolution and Social Change
an academic unit of the College of Liberal Arts and Sciences
Francis P. McManamon, Executive Director
PO Box 872402
Tempe, AZ 85287-2402
Direct: (480) 965-6510 Digital Antiquity: (480) 965-1369 Fax: (480) 965-7671
http://digitalantiquity.org http://tdar.org fpm@digitalantiquity.org

(c) Encourage the re-use of existing successful software tools across disciplines where available (to both decrease costs of implementing archival systems and to foster interdisciplinary and integrative research).

2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Public access and use might be enhanced by expanding the "fair use" concept applied to sharing documents (even when copyrighted) and data, if the use is for non-commercial educational, scholarly, or scientific purposes.

In increasing public access, it is important to develop and employ administrative procedures, data organization, and/or publishing formats that control appropriate access to "confidential data," for example, very specific locations of archaeological sites that might be subject to looting if their locations were generally known. It also would be desirable to develop procedures to provide appropriate warnings concerning sensitive information that may be available in widely shared data, documents, or images.

In developing procedures for wider access to scientific data, agencies must balance the requirement of placement in digital archives with a short embargo within the archive to facilitate the completion of ongoing research while ensuring future public access via the archive. Creative Commons and other open access licensing formats have worked well for document and other text data, but are not completely appropriate for research data. There are several current efforts (e.g., the Open Data Commons Project) to develop appropriate licenses for research data. On balance, though, every effort should be made to encourage public and professional access to data through appropriate channels. Related to this wider access to data, policies and procedures are needed to ensure and encourage the appropriate citation, crediting and attribution for individual researchers and organizations that carry out research and produce the data being shared.

3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Federal agencies should work with other agencies or organizations (like the National Science Foundation) that have well-established disciplinary directorates to develop policies relevant to their area of expertise. A number of disciplines are already addressing issues of data preservation and access specific to their research; it would be cost-effective and reduce duplication of effort to work with the professional societies in developing policies. For example, within the discipline of archaeology, the Center for Digital Antiquity (http://digitalantiquity.org), the Open Context repository (http://opencontext.org/), and the Archaeology Data Service in the UK have cooperated on various topics related to the digital archiving and providing for access and use of archaeological data.

4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Considering the cost of lost data versus the costs of long term preservation, lost or inaccessible data essentially means that all Federal monies spent on a research project were expended for no result. In such instances, there is no potential for current or future benefit to the public or the American scientific enterprise. Alternately, preservation and access may require marginally more Federal funding to ensure that the research results are accessible and preserved, but those costs are amortized over a very long time period and will have a broad range of economic, scientific, and educational benefits.

5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Stakeholders are typically generalists (such as universities) or specialists in specific disciplines or topics other than digital archiving. Their efforts would be best steered toward implementation of policies that require submission of data to appropriate disciplinary digital archives. Stakeholders should participate in the development of guides to good practice for researchers contributing data, both broadly (libraries, universities) and in specific disciplinary research communities. Stakeholders should require authors and publishers to provide data related to their publications in ways that facilitate archiving and in standard file formats that are susceptible to archiving and ultimate migration as new formats develop. Researchers should be required to organize their data in ways that facilitate archiving and in standard digital formats that are amenable to archiving and conversion as new formats are developed. Likewise, authors should provide appropriate documentation for all archived materials to ensure that the information is useful to various stakeholders.

6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Computational science models predict that the cost of storage will continue to decrease over time. However, the cost for long-term archiving and curation are more complex because files have to be converted as software becomes obsolete and hardware and software advances are made. A "pay once, store forever" model (the so called Princeton model, developed by Goldstein and Ratliff, 2010 [http://arks.princeton.edu/ark:/88435/dsp01w6634361k]) is currently the most reasonable option. Grant funded projects should include an appropriate direct cost line item for long-term curation and preservation of digital research data.

Life-cycle cost analysis is a useful method of determining or estimating the real costs of preserving and making digital data accessible. Requiring funding mechanisms to incorporate ways to explicitly address the life-cycle costs associated with the maintenance of digital data will be an improvement over current approaches. Relying on discipline-specific and specialist repositories will both minimize costs associated with digital data maintenance (economies of scale) and those expert facilities will be well situated to realistically determine the costs associated specifically with the long-term preservation of and access to digital resources.

7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access polices for scientific research? How can the burden of compliance and verification be minimized?

Agencies can require that researchers submit as part of their invoicing for research grants or in their reports on grant expenditures, official documents from the digital archive in which their data and documentation has been placed. These documents need not be complicated, but should verify that the research data and documentation have been deposited in the digital repository along with appropriate metadata and that they are now accessible. The archive also should affirm that the deposit of the data and documentation means that these will be preserved for future use in the archive.

In the planning for a research project and as part of the data management plan submitted with the research proposal, the steps should be described to track digital data from its creation through placement in an appropriate archive.

8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Agencies should provide incentives through grants specifically targeted for integrative research drawing on digitally archived data. These could have very high returns at small cost: new data would not need to be generated - but bigger-picture questions beyond any single project could be addressed.

Tools like the Sunlight labs have shown the value of data when available, by taking the step of requiring it to be captured digitally and available this would facilitate significant progress for entrepreneurship. Agencies also could offer grants to "rescue" important data in out-of-date software programs or media. This would be particularly important for lost or threatened data from past federally funded research. A systematic approach to digitally archiving legacy data would be an ambitious and extraordinarily important project for many disciplines.

9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

As a condition of grants, agencies could require thorough citation details as part of the metadata for archived resources. Appropriate attribution (whatever the source) is a fundamental part of the responsible conduct of scientific research and should be explicit in the training of students, scientists, and other professionals. Disclaimers and terms-of-use could be required by the digital repository laying out expectations of proper attribution if the data or supporting documentation is used in any fashion. Repositories also should be encouraged to offer services that provide, for example, a standard format for citing the research data set (similar to the formats used for citing published works). Included in this citation should be permanent identifiers that are associated with a particular data set (e.g., Digital Object Identifiers [DOIs]).

Standards for Interoperability, Re-Use, and Re-Purposing:

10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment:

see Brazma et al., 2001, Nature Genetics 29:371) is an example of a community-driving data standards effort.

The effort should start with better practices for the long term preservation of digital data. (an example would be the *Guides to Good Practice* (http://guides.archaeologydataservice.ac.uk/) developed for archaeological archiving and data files by the Archaeology Data Service in the UK and the Center for Digital Antiquity in the US.

Libraries and archives already have a number of standards to support basic interoperability, but discipline specific metadata requirements would be a significant first start.

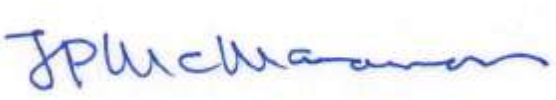
12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Agencies could provide funds for cooperative activities between disciplinary archives in different countries, e.g., for archaeology, the Archaeology Data Service in the UK and the Center for Digital Archaeology in the US. Agencies also could fund archival projects that develop links among repositories, e.g., the TransAtlantic Gateway project (funded by the US National Endowment for the Humanities and the JISC in the UK).

13) What policies, practices, and standards are needed to support linking between publications and associated data?

Policies that encourage publishers, specifically those who publish top tier journals for specific disciplines, to work with established archives to link data with articles should be developed. This has worked well in the field of ecology, where the data archive, DRYAD, holds research data related to publications in various ecological journals. When an article is accepted for publications, the journal requires that the relevant data sets be deposited in DRYAD. Similar approaches should be developed and encouraged in other disciplines by agencies through funded initiatives.

Sincerely,



Francis P. McManamon, Ph.D., RPA
Executive Director and Research Professor

January 12, 2012

Catherine Casserly; cathy@creativecommons.org
Timothy Vollmer; tvol@creativecommons.org
Creative Commons
Mountain View, CA

Re: OSTP Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research [http://federalregister.gov/a/2011-28621]

Creative Commons (CC) is pleased to submit comments to the Office of Science and Technology Policy's Request for Information (RFI) on the topic of Public Access to Digital Data Resulting from Federally Funded Scientific Research. Creative Commons (http://creativecommons.org) is a 501(c)(3) U.S.-based nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright. CC develops legal and technical tools used by individuals, cultural, educational, and research institutions, governments, and companies worldwide to overcome barriers to sharing and innovation. Creative Commons operates globally. The international CC affiliate network consists of 100+ affiliates working in over 70 jurisdictions, and there are over 500 million CC-licensed works available on the web.

Thousands of academic researchers release print and digital datasets, journal articles, and educational materials under Creative Commons copyright licenses and waivers, allowing those materials to be easily found, accessed, reused and re-purposed around the world. CC licenses offer a flexible set of permissions so that authors and publishers can release their data while ensuring—if desired—that they receive attribution for their work, or explicitly place their research into the public domain in cases where copyright does not apply (e.g. for collections of factual data).

We answer specific questions laid out in the RFI below. Note that we have not answered all the questions, as we believe there are other organizations and stakeholders with greater expertise in those areas.

Question 1
What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Comment 1
The federal government should establish policies that insure the public has cost-free, unimpeded access to the digital data resulting from federally funded scientific research. Data should be made accessible in a manner that explicitly communicates the rights available to data re-users. The public should have the right to reuse publicly funded data free of any legal restrictions, or with the only restriction being that the source of the data is credited.

Access to this data should be made available as soon as possible, with due consideration to confidentiality and privacy issues, as well as the researchers' need to receive credit and benefit from the work (for instance, a short embargo period for data is not antithetical to the recommendation of public access, as original researchers invest significant effort into the creation and analysis of data and would like to be able to capitalize on their efforts).

If the federal government wants to maximize the impact of digital data resulting from federally funded scientific research, it should provide explicit, easy-to-understand information about the rights available to the public. The simple process of posting federally funded scientific data in a publicly accessible repository on the web so that they can be viewed and downloaded will not realize the full reuse potential of the data. Even where it is the intention that digital data created with public dollars be widely shared, as long as those materials are not clearly marked with information describing the rights and permissions under copyright law, the use of those data will be diminished and impact of public investment lessened.

With a renewed attention to clarifying the rights available to data re-users in advance, the federal government can increase the speed of scientific discoveries, promote innovation, and support new economic opportunities. And as scientific researchers, creative startups, and traditional commercial businesses are granted the reuse rights they deserve, these and other groups can best leverage federally funded digital data to advance the scientific enterprise. Unimpeded access to the digital data resulting from federally funded scientific research can increase the speed and variety of scientific discoveries and boost U.S. competitiveness by encouraging the development of new products and services.

The federal government can grant these permissions to the public by 1) dedicating the data to the public domain or 2) adopting a liberal licensing policy where at most downstream data users must give credit to the source of the data.[1] Creative Commons offers public domain tools and licenses to help accomplish these goals. CC0 (read "CC Zero") is a tool that allows data publishers to dedicate data to the public domain by waiving all rights to the work worldwide under copyright law.[2] Waivers such as CC0 are the gold standard with regard to global interoperability and innovation potential because it removes all barriers to reuse from data. In certain domains, such as science and public sector data, there are important reasons to consider using waivers like CC0. Waiving copyright and related rights in a domain like science eliminates all uncertainty for potential re-users, encouraging maximal reuse. Echoing the Panton Principles,

> "[I]n science it is STRONGLY recommended that data, especially where publicly funded, be explicitly placed in the public domain via the use of the Public Domain Dedication and Licence or Creative Commons Zero Waiver. This is in keeping with the public funding of much scientific research and the general ethos of sharing and re-use within the scientific community."[3]

[1] By public domain, we mean the legal concept whereby content is not protected by copyright.

[2] http://creativecommons.org/publicdomain/zero/1.0/

[3] http://pantonprinciples.org/

Use of explicit waivers of rights is necessary to avoid uncertainly around what uses may be made of data, e.g., to integrate with other data to create new datasets or reformat it to support its long-term preservation. Even in the frequent case of data that is automatically in the public domain, e.g., because it is strictly factual in nature and so not subject to copyrights under U.S. law, explicit declarations of the rights-free status of the data is necessary since researchers are not expert in the nuances of copyright law and often make incorrect assumptions about their rights to reuse data. In other cases datasets of mixed factual and creative content, making their legal status murky even to legal experts. This uncertainly can result in researchers avoiding reuse of existing data for fear of inadvertently violating a non-existent copyright or to avoid involving legal counsel to determine possible reusability.

The situation is even more complex for international research collaborations where different countries and legal jurisdictions have different or conflicting laws and policies regarding protection of data, such as the sui generis data rights that apply in the European Union but not in the U.S. For data produced by such international collaborations (an increasingly common phenomenon) a rights waiver such as CC0 is the only existing method of making the data unambiguously legally available to U.S. researchers for reuse.

Where attribution is desired, the federal government might consider requiring the Creative Commons Attribution 3.0 (CC BY) license for digital data resulting from federally funded scientific research. CC BY is a copyright license that grants permission to the public to reproduce, distribute, perform, display or adapt the licensed materials for any purpose so long as the user gives attribution to the author or as otherwise directed by the copyright holder.[4]

Question 2
What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2
The current climate around intellectual property (IP) is increasingly one of ratcheting up enforcement and maximization of copyright and other IP rights. However, it's important to understand the underlying priorities of scientific researchers to determine whether the default copyright regime reflects their needs and supports their academic endeavors. The primary motivation for scientific researchers is "to inform others about their work," thus contributing to the scholarly canon and promoting the advancement of science.[5] Scholarly authors wish to have their research read, consumed, and cited. To do their work, scientific researchers must acquire and use data from a wide variety of sources, and need to know how they may reuse data—either by itself—or more likely, in combination with other datasets. So, it makes sense to be able to provide the widest access with the fewest encumbrances to digital data resulting from federally

[4] http://creativecommons.org/licenses/by/3.0/

[5] Hansen, et al., *Intellectual Property Experiences in the United States Scientific Community*, 2007, p. 8. Available at http://sippi.aaas.org/Pubs/SIPPI_US_IP_Survey.pdf.

funded scientific research. For this access to be meaningful, scientists need clear and unambiguous information (metadata) about the rights they are granted in using the data for their research.

Rather than focusing on Intellectual Property as the means for capturing the value of research performed, mechanisms are needed to provide academic and public credit to researchers for their work, to the agencies that funded the work, to publishers who promoted the results, and to other stakeholders such as data archives that made the data available. Such mechanisms will require broadly available and standard infrastructure such as persistent and globally unique identifiers for the data, researchers, institutions, funding agencies, and other stakeholders that allow data use and reuse to be tracked over time and made part of the scholarly record.

Some argue that placing digital data resulting from federally funded scientific research directly into the public domain is suboptimal because it will be impossible to track the attribution, provenance, and credibility of the published data. But these are separate issues: citation has always been a normative practice that scholars enforce through social mechanisms. Whether the cited research is in a peer-reviewed publication or posted on the public Web does not affect its citability. Citation and credibility are made possible via mechanisms like associated metadata and disciplinary norms for determining quality and provenance. Even licenses that require "attribution" (common for copyrighted creative works) does not guarantee appropriate citation, since licenses and attribution are legal tools that are usually enforced via lawsuits. A public domain policy for the sharing of digital data resulting from federally funded scientific research is not misaligned with the need for citation and quality control.

Question 3
How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3
With regard to the management of copyright, the federal government should adopt common, global standards for the communication of rights for digital data resulting from federally funded scientific research. Scientists and researchers around the world currently use CC tools such as CC0 to release data into the public domain and CC BY (where copyright adheres to the data) to allow its reuse with simple attribution. Standardized, public licenses and waivers should be strongly recommended instead of customized solutions for each discipline because it increases compatibility and clarity to the end user in how they may use the data and combine it with other datasets.

More generally, federal agencies should adopt a standard data format for the communication of rights and permissions for digital data resulting from federally funded scientific research. The Creative Commons Rights Expression Language (ccREL) is a specification for how license information is described using RDF and how licensing information is attached to works.[6] ccREL

[6] RDF stands for Resource Description Framework. RDF is a "family of W3C specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats." For more information, see http://en.wikipedia.org/wiki/Resource_Description_Framework

is the standard recommended by Creative Commons for machine-readable expression of copyright licensing terms and related information, so that content and data can be exposed via search engines like Google.[7] ccREL is embedded within CC license metadata and the CC0 public domain dedication tool metadata.

Question 8
What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8
Especially for the scientific community, digital data resulting from federally funded scientific research should be made immediately available in the public domain in order to speed the discovery of cures for diseases and promote the advancement of science.[8] Immediate access speeds up the research and development cycle, thus leading to faster development of value-added research, products and services.

Clear, unambiguous policies around the management and sharing of digital data resulting from federally funded scientific research should be provided to federal grantees, and the federal government should ensure compliance with these policies. As discussed above, federal grantees that create digital data should be required to place that data in the worldwide public domain using CC0 or made available with the minimal attribution requirement via CC BY. Adopting a clear data management policy should help rectify the widespread over-reach of many data provider Terms of Use statements. Many of these Terms of Use statements curtail re-use through overly restrictive licensing agreements. In addition, the federal government can help educate grantees about copyright with regard to data. For instance, strictly factual information does not rise to the level to warrant copyright protection, and should not be claimed as copyrightable material.

Question 10
What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma *et al.,* 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Comment 10
Digital data resulting from federally funded scientific research should be shared in the worldwide public domain using the CC0 public domain dedication tool or made available via the liberal CC BY license. Both CC0 and CC BY metadata include ccREL specification for machine-readable expression of copyright licensing terms and related information. Ensuring that licensing metadata

[7] http://wiki.creativecommons.org/images/d/d6/Ccrel-1.0.pdf

[8] The Tuberculosis Commons (TB Commons) Initiative Open Innovation Team "believes that open models can accelerate knowledge turns resulting in a faster drug development process." Content provided by end-users to the site is released into the public domain using the under the CC0 Public Domain Dedication tool. See http://www.tbcommons.org/initiative/.

and related information is clearly communicated to humans and machines promotes interoperability, reuse, and repurposing of digital scientific data.

Question 13
What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13
For the OSTP RFI on Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research, we recommended that scholarly articles created from federally funded research be released under full open access. Full open access policies will provide to the public immediate, free-of-cost online availability to federally funded research without restriction except that attribution be given to the source.[9] The standard means for granting permission to the public is through a CC BY license. This license is aligned with the principle of full open access because it allows rights (including commercial rights) to be communicated with the only requirement that users give credit to the rightsholder. Full open access policies for publications are necessary to maximize the impact and reach of federal research investments, and promote scientific discovery and economic activity. By adopting a common legal licensing framework for publications and data, the federal government can ensure maximum interoperability and reuse of federally funded research outputs.

We thank OSTP for the opportunity to provide comments to this RFI, and we're happy to answer any other questions you may have.

Sincerely,

Catherine Casserly, CEO, Creative Commons
Timothy Vollmer, Policy Coordinator, Creative Commons

[9] Carroll, Michael. *Why Full Open Access Matters.* PLoS Biology, November 2011. Volume 9, Issue 11. Available at http://www.plosbiology.org/article/info:doi%2F10.1371%2Fjournal.pbio.1001210.

Thu 1/12/2012 6:22 PM
RFI:Public Access to Digital Data Resulting from Federally Funded Scientific Research

[Assigned ID #]
[Assigned Entry date]
Name/Email
Morteza Gharib, Vice Provost / vpr@caltech.edu
Affiliation/Organization
California Institute of Technology
City, State
Pasadena, CA 91125

Caltech is a PhD university employing 922 principal investigators whose research funding comes largely from 6-10 different federal agencies. In addition, Caltech is committed to education and recognizes its profound obligation toward public dissemination of its research results ideally unfettered by the demands of commercial profit so that learning and discovery, two major pillars of the enterprise, will thrive. The global network provides the means to ensure maximum access for uptake of new knowledge via electronic distribution of publicly funded research results. Therefore, Caltech urges and supports action to require prompt public access to results of all government funded research.

In response to the request for information from your office released on November 3, 2011 on the topic of public access to peer-reviewed scholarly publications resulting from federally funded research we offer the following comments.

922 includes professorial faculty, research faculty and postdoctoral scholars.
The Dept. of Defense is counted as one agency.

Comment 1
The National Science Foundation and National Institutes of Health require data management plans for all current grants. This is an important first step in creating incentive to improve access to and reusability of data generated by federally funded research. Broadening the data management requirement across the full range of unclassified, federally funded scientific research will significantly improve the efficiency and productivity of the American scientific enterprise.

Standardized data deposit policies across the spectrum of funding agencies will go a long way toward ensuring ease, and therefore consistency, of deposit. Where possible, discipline-specific data repositories (e.g.; ICSPR, IRIS, CDIAC, PDB) should be utilized to create content-rich destinations for data seekers, both human and machine. These databanks will facilitate discovery, access and preservation activities, moving past the widely varying interpretations of data access policies of individual primary investigators.

Comment 2
Existing publisher copyright policies and attendant business models for peer reviewed publications are ill suited to data. Publishers focus on dissemination and licensing for all of the content to which they hold rights. Preservation and universal access are not core values or business functions, especially in light of the vagaries of business mergers, consolidations, divestitures, and bankruptcies.

Data, as facts, are not subject to copyright. The American Chemical Society, in 2010, ceased to claim exclusive rights to supplementary material published in its journals. Dryad provides data repository services for 100 journals.

Creative Commons CC-BY and CC-0 licenses represent excellent consistent and standard starting point from which to build licenses for data, acknowledging potential copyright issues, while maximizing the prospects for both people and machines to build services that maximize discovery and access.

A rational argument for embargoes can be made to acknowledge the unique efforts of the primary investigator or original scientific team. The efforts of primary investigators should be acknowledged through a term of exclusive access to data (i.e.; throughout the publication process) allowing time for data to be thoroughly processed and analyzed. Primary investigators and their teams should be given the first opportunity to make discoveries and produce publications. It is important to note that a right of first publication does not preclude the deposit of the data into a certified data repository even during an embargo period, particularly to initiate archiving activities.

Comment 3
Different scientific disciplines offer a broad spectrum of requirements for data management practices and policies. There are some basic conditions for archiving and preservation that apply regardless of scientific discipline. Datasets require significant documentation (e.g.; equipment and equipment settings, provenance, data processing) to make comparisons and combinations with other datasets valid. Identifiers, fixity information, Persistent URLs are a few of the critical pieces of data infrastructure supporting archiving and preservation that should be applied to all scientific datasets.

Each scientific community is best qualified to address its specific data management needs. However that perspective tends to be narrowly conceived and minimally applied. Not all domain scientists may be explicitly aware of data management issues and needs. Basic requirements from funding agencies offers the opportunity to gain the necessary attention from the researchers. A major point is that they are held accountable to the public access concept and focus on scientifically defensible criteria for any deviation from full and complete access for humans and machines.

Comment 4
Different types of data and the needs of specific scientific communities will introduce different relative costs and benefits. It may be useful to consider needs as they relate to baseline services that apply across all scientific disciplines (e.g., archiving) and secondary services (e.g., discovery and specialized query capabilities). Agency policies should accommodate the relative emphasis between these two categories of services in different disciplines, as it relates to the distribution of costs and benefits across the full array of stakeholders. Federal agencies might wish to fund libraries and museums to develop the data archiving capacities, yet expect those libraries and museums with their parent organizations to bear the long-term costs given their cultural memory missions over the long haul. Agencies could provide seed funding for preservation while provide ongoing funding to a scientific community to develop secondary services.

Comment 5
Data management practices within scientific communities are currently diverse. This is not necessarily a bad situation. Some scientific disciplines already benefit from ongoing, centralized data repositories (e.g.; ICSPR, IRIS, CDIAC, PDB). It would be negligent not to build upon this already existing infrastructure. Of significant importance is the need for these data repositories to demonstrate their ability to preserve data functionality over time, not just assure the community that preservation is being done. Archives, libraries and museums have an extensive track record with these functions and could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions of preservation. However the parent institutions need to participate in the strategic re-ordering of resources to meet the new needs. With a clear set of requirements, it will become possible

to identify how various stakeholders can implement data management plans, noting that these roles will vary by scientific discipline or community.

Comment 6
The real costs of preserving and making digital data accessible are legitimate and important costs of the scholarly infrastructure necessary to support research. Grant proposals will need to include funding for data curation for preservation similar to the usual practice of providing funds for publishing. That being said, funding of cyberinfrastructure to create community-based data repositories, for scientific disciplines where the cost benefit ratio supports the notion, should not be ignored. Some funding in research grants is necessary for the preparation for data sharing, but the bulk of the cost of preservation, discovery, and dissemination services may reside in the operation and maintenance of discipline-specific data repositories.

Comment 7
Workflows of library-based or community-based data archives are implementable and effective platforms to ensure compliance. Plus such community based services are more likely to offer the researchers a consistent and reliable set of rules over time. Persistent identifiers (ORCID, doi, Persistent URL) and appropriate licenses (CC-By, CC-0) represent critical mechanisms through which compliance and verification can be automated thereby reducing costs. NIH currently requires PMCID reporting for all articles published under a grant in progress reports and final grant reports. A similar reporting mechanism for data deposition would flesh out the NIH data management requirement and could be generalized to other agencies, enabling verification mechanisms for data management requirements.

There are two activities that require the researcher's interaction with a third party: proposal submission and publication submission. Proposal submission is the place to insert data management requirements, as evidenced by the NIH and NSF data management plan requirements. Publication submission is a key point at which investigators have validated the datasets which are relevant and trustworthy for sharing via publication and deposition in data repositories. By embedding appropriate data management planning with proposals, data deposition with publication, the use of appropriate licenses (CC-By, CC-0) and cyberinfrastructure requirements into these workflows, the prospects for efficient compliance and verification, through standard grant reporting mechanisms, are heightened considerably. Researchers will resist any additional burden. However their institutions or their communities must develop capacity to support and implement data management plans, so those "burdens" can be shifted to support infrastructure (libraries, museums, disciplinary data repositories) that view such activity as part of their core mission.

Comment 8
Federal agencies could stimulate the development of discipline-based and institutional data archives that support discovery, download, and preservation. A uniform mandate across agencies to make data freely available through such archives, under appropriate licenses (CC-By, CC-0), would encourage the growth of such archives and, by extension, the development of APIs to allow individuals and machines to develop new capabilities and services. This type of open system would facilitate new business opportunities even by smaller businesses. The licensing arrangements (CC-By, CC-0) would be critical to ensure that no single entity or group has an exclusive right to generate such new business opportunities.

Comment 9
One of the most important components is author and institutional identifiers (e.g., ORCID) that would support developing attribution and credit processes. Machine-based access argues for CC-0 licensing. The seismology community already acknowledges use of datasets deposited in IRIS. While the original

primary investigators are not necessarily credited, the community benefits from the ability to compare and compile datasets from a wide variety of seismic arrays.

Comment 10
Barcode of Life Data Systems (BOLD) is a community-based data repository supporting evolving tools and data standards that aid in the collection, management, analysis, and use of DNA barcodes. While there are many community-driven data standards for scientific data, most of them deal with interoperability or sharing rather than archiving or preservation. There are too many discipline specific efforts to list. Basic requirements for more x-disciplinary sharing need to be developed.

Comment 11
There are examples of successful data standards development efforts within various domains (e.g.; Structural Biology Knowledgebase (SBKB) of the Protein Structure Initiative (PSI), FITS (Flexible Image Transport System) in astronomy, FGDC (Federal Geographic Data Committee) geospatial metadata standards). In each of these cases, there are undoubtedly several, perhaps common, characteristics or reasons for the success of the effort (or alternatively reasons why such efforts did not succeed in other cases).

Comment 12
While there exist groups that work in this area (e.g.; the National Academies' Board on Research Data and Information (BRDI), CODATA (Committee on Data for Science and Technology)), it would be helpful for Federal agencies to support community-based efforts that connect nodes of data infrastructure development activities, both disciplinary and institution-based repositories. For example, the European-based EUDAT project has already reached out to projects within the U.S. regarding a Data Access and Interoperability Task Force (DAITF) along the lines of the Internet Engineering Task Force. NIST could be helpful in this context supporting the development of a "data grid" that would operate in a similar manner to the power grid.

Comment 13
There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Traditionally, the linkage, if it existed, has been through supplementary material in journals. The Society for Neuroscience, for example, no longer accepts supplementary materials for distribution with articles in *Journal of Neuroscience*. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. Dryad has created partnerships with 100 journals, from a wide variety of publishers, to host, preserve, and provide long-term access to the data underlying formal publications. A key consideration from a policy, practices and standards standpoint is a requirement to use persistent, unique identifiers (ORCID, doi, Persistent URLs/handles) for publications, data, authors, and any entity of interest. These identifiers not only bolster the linking and attribution of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as individual investigators' website URLs, a step which *The Astrophysical Journal* is in the process of undertaking. Agencies may have to specifically require that a national identifier scheme be used. Many researchers do not understand the difference between a url and a supported identifier that resolves to the currently active url.

J. Willard Marriott Library
THE UNIVERSITY OF UTAH
295 South 1500 East Salt Lake City, Utah 84112-0860 (801) 581-8558

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Experience so far indicates that to be effective, federal policies must mandate the data created in the course of federally funded research be deposited into publically accessible repositories. The National Science Foundation requirement of a data management plan is a laudable step toward awareness of the need to manage data, but a mandate will be required to create the critical mass of available data that will support rapid scientific innovation and encourage the commercial reuse of data that can underlie economic growth.

One aspect of the success of the NSF approach has been to set an expectation but not to require a specific method of implementation. Because of the variety of approaches and types of data across different disciplines, flexibility in compliance is called for, even within the context of a mandate. This flexibility can help the scientific community come to view data preservation and sharing as an issue of principle, necessary for good research and scientific accountability, rather than as merely a burdensome compliance issue.

A policy mandating data deposit will need to be accompanied by the development of standards and services that make data sharing economically feasible and data reuse as accessible as possible. Incentives are as important as requirements if the goal is to make usable data available for scientific verification and commercial reuse. By creating systems that are as simple, standardized and open to reuse as possible, the maximum potential of economic growth will be achieved. A useful metric for public access to data is whether someone, or some computer, can discover, access, interpret and use the data without having to contact the original data producer; such access is both economically beneficial and less burdensome to the data producers.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

An important aspect regarding data is that under U.S. law, raw data is not subject to copyright and that rights even to protectable collections of data usually remain with the data producer, rather than being subject to transfer to publishers. So the rights issues are not as complex for data as they may be for publications.

The basic premise of federal policy should be that more openness is better, and restrictions should be applied only when genuinely necessary, for example, when the data makes it possible to identify a particular person involved in the research study. Basically the default, which is currently that data is managed locally (if at all) and idiosyncratically, should be changed to openness and standardization.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

The key to convincing data producers, who are also the holders of whatever IP rights exist, to participate is to provide easy roads to compliance and incentives, usually in the form of norms and expectations within their disciplinary communities, to comply. Other incentives involving credit for their work, easy citation methods, and ability to demonstrate the impact of their work by using metrics such as the number of times their data set has been cited are important incentives and need to be built into any plan for compliance.

The federal government could assist in making data preservation and sharing as seamless as possible by supporting the creation of successful data management systems that resemble currently successful programs such as those at the National Library of Medicine's National Center for Biotechnology Information. There are four different methods for submitting data into GenBank, along with a number of other tools to make this process easy. Working with publishers in this process to deposit and store data may initially seem the obvious path, but publishers are in a business to make money. Publishers do not have the commitment to preservation or open access that other stakeholders, such as libraries, have already demonstrated in their work with books, journals, video, audio, maps, microfiche and other rare materials.

There is a reasonable argument for embargoes, in some cases, based on the unique effort exerted by the data producers or original scientific research team. Although effort alone cannot justify copyright protection (based on the Supreme Court's 1991 decision in *Feist Publications v. Rural Telephone Service Co.*), the need to protect data for some short period of time while the team or lab completes its own analysis could be respected by allowing a fixed-term period of exclusive access. Such an arrangement, however, does not preclude the deposit of the data into a certified repository even during that embargo period, particularly so that archiving activities can be begun.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Federal agencies can take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data by adopting a relatively general mandate for data sharing while requiring more specificity for the practices within each discipline. This is because there is ample evidence that different scientific disciplines present a variety of requirements for the management of data. Data sharing policies should be viewed as flexible requirements that remain open to modification as problems arise or best practices emerge from within specific communities of scientific practice.

Some baseline conditions or requirements, especially related to archiving and preservation, can be applied across the board. This is a vital place to begin, since many scientific disciplines have focused on access or discovery rather than preservation, yet the latter is key to fostering efficiency and innovative reuse.

In some disciplines, a funder requirement will serve as a first step toward creating awareness of the fundamental need for data management, preservation, and access. We have seen this take place among working scientists as awareness of the NSF data management plan requirement has spread, and further mandates will facilitate this awareness.

Funding agencies should be willing to provide funding for data management expertise that is available locally at researchers' institutions, and/or through disciplinary repository services (such as the DRYAD repository at the National Evolutionary Synthesis Center). Such support will assist researchers in applying data management approaches that are appropriate to their specific disciplines.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

In general it would be difficult for agency policies to consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research because of the particularized and ad hoc nature of so many approaches to research data up until now. It may be most useful to think in terms of baseline services that should be supported across all disciplines (i.e. archiving) and more particularized secondary services (such as specialized query capabilities). Agencies might consider the relative emphasis that is appropriate in the area of research that that agency funds, and what areas are appropriate for local institutions to assume responsibility for. Thus an agency might provide seed funding to institutions for preservation, but recognize the need for ongoing funding to a scientific community to develop secondary services.

A potential technique to establish a baseline cost would be to set an allowable cost for data management for funding requests, then analyze, after several rounds, what approaches have been applied and how effective they have been based on metrics such as use statistics, the verifiable integrity of the data over time, and third-party costs to discover, retrieve and use the data. If funding is provided to disciplinary repositories, reports based on these metrics should be required.

The benefits of shared data will also be difficult to measure, but they are nonetheless real. Accountability and the ability to verify scientific results are vital, but hard to quantify. Other benefits, such as the support provided for reuse by different teams of researchers or by commercial enterprises, will be easier to track. The opportunities for innovation and commercial exploitation of shared data will be evidenced by increased growth within a sector of the economy.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?
Stakeholders can best contribute to the implementation of data management plans by having access to or creating easy and efficient methods for themselves and others in their field for the storage, description, and sharing of the data they collect.

It is important to keep researchers focused on research; this is vital if a data sharing requirement is going to support innovation and growth and not hinder it. The stakeholders named thus have the important role of providing the services, standards, best practices and infrastructure that make data sharing simple and efficient. Insofar as agencies can provide funding and other incentives to support those functions, they will contribute to the implementation of data management plans.

The best approach is to build on existing infrastructures and practices, learning from what works well while being sensitive to disciplinary differences and the evolution of scientific disciplines over time.

While successful practices should be the model for policy implementation, it is important that success be demonstrated and not merely asserted. Each agency, as part of its data sharing mandate, should identify metrics that are important within that field by which the success of a plan or services can be measured. Those metrics will evolve over time, but with a clearly articulated set of requirements it will be possible to identify how various stakeholders can contribute to the successful implementation of data management plans.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Funding mechanisms can be improved to better address the real costs of preserving and making digital data accessible by acknowledging and communicating that the real costs of preserving and sharing digital data are indeed legitimate and important costs of the overall research enterprise.

It is important to recognize that not all costs associated with good data management will be directly attributable to specific projects. As data management expectations become more widespread and routine, an increasing proportion of the costs will need to be considered indirect costs. While some disciplines or projects may present exceptional needs, many other research projects will likely rely on baseline services provided by institutions or disciplinary groups that need more general formulas for funding.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

There are several approaches agencies can take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research. For example, reporting metrics should be developed and applied to early efforts at improving data stewardship, and the results shared broadly. As best practices emerge and community norms support good data management, researchers will have an incentive to preserve and share their data.

In addition, compliance should be verified through systematic approaches, which can be much easier and efficient for the agency and less punitive for researchers. Most researchers pay special attention to two milestone events in the research process – the grant proposal and publication. Policies and metrics that are embedded at these points will get the attention of researchers and make compliance more likely.

Finally, agencies should develop guidelines for those who review both grant proposals and final reports to the funding agency that highlight what to look for in a well-developed data management plan within the specific discipline.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There are several additional steps agencies can take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy. First, the use of open data licenses and platforms that facilitate sharing in standardized ways will make it easier for other researchers and industries to reuse data and increase the return on investment for funded research projects.

Second, support for well-documented APIs that allow individuals and machines to develop new capabilities and services is key to fostering innovation. One of the benefits of the broadest possible access and opportunity for reuse is that federal agencies could help build on "citizen science" efforts, which have up until now largely focused on data gathering and classification. Open licensing and usable APIs will ensure that the maximum number of creative imaginations are looking for innovative ways to use research data.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

There are several mechanisms that can be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported. One such mechanism is to support ongoing efforts to develop data citation standards (such as the DataCite project) and author and institutional identifiers (such as those being developed by ORCID).

Another mechanism would be to require agencies to disclose data sources using common data citation and researcher identification standards in order to build community norms that reward good attribution practice, as is the case for research articles.

Nevertheless, it should be recognized that existing attribution standards for published articles will not translate seamlessly into the world of research data, especially given the importance of machine-based access and reuse. As in so many other areas, this is a case where standards will have to develop as reuse and innovation grows, and agency mandates should remain flexible while publicizing and encouraging best practices.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

There are an astounding number of data standards available depending on the data in question and the subject area. The MIAME example is typical of the development of data standards among researchers in a particular area. A quick search in Google Scholar displays several other articles on a similar theme:
Taylor, et al., 2007, The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology 25, 887.
Bustin, et al., 2009, The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry. 55(4) 611.
Novere, et al., 2005, Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology. 23(12) 1509.
Field, et al., 2008, The minimum information about a genome sequence (MIGS) specification, Nature Biotechnology 26, 541

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Another example involves the development of Common Data Elements, which are jointly created by users of the caBIG database, hosted by the National Cancer Institute (NCI). There are 20 different Common Data Elements in use and NCI is accepting others for review:
https://cabig.nci.nih.gov/workspaces/VCDE/Data_Standards/

One possible element that has lead to the success of many of these efforts to create data standards is the creation of the standards by those working in the field working with the data and attempting to share or use data created by others.

Some of the best examples of data management, preservation, access, and use can be seen at the National Library of Medicine's (NLM) National Center for Biotechnology Information (NCBI).

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Most research projects are completed in teams across institutions and a portion of those teams are international. These teams face the need to create standards for sharing data at earlier points in the process and are often poised to make the recommendations for data standards in their field of study. The fact that the articles listed in Question 10 are authored by international teams is a good indication of a process that works well. Even GenBank from NCBI is an international effort of data sharing, preservation, and storage.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

The National Library of Medicine provides an excellent model of linking between publications in PubMed and data located in a whole host of databases and then back to the publications with single click capabilities.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

UNIVERSITY OF
NORTH TEXAS
DISCOVER THE POWER OF IDEAS

UNIVERSITY LIBRARIES

Dr. Martin Halbert
UNT Libraries
1155 Union Circle #305190
Denton, TX 76203-5017

January 12, 2012

To the Office of Science and Technology Policy:

Thank you for the opportunity to offer comments in response to your RFI, Public Access to Digital Data Resulting from Federally Funded Scientific Research. This issue is of major long term significance to the long-term success of scientific research. The University of North Texas is very concerned with the survivability of scientific research results, and has undertaken a research project aimed at understanding and identifying solutions for the complex problems that you have marked. This research, still in early stages, is funded by the Institutes of Museum and Library Services (IMLS). In this brief response, we would like to a) share the early results of this research project, and b) offer our own recommendations to the questions articulated in your RFI. Our recommendations are informed by both our research effort and our perspectives as members of the National Digital Stewardship Alliance (NDSA) and the Scholarly Publishing and Academic Resources Coalition (SPARC). We participate in the organizational discussions concerning this issue within NDSA and SPARC, and endorse the comments which they have separately provided in response to your RFI.

UNT Research on Data Management

Our research project is entitled "DataRes: Research on Emerging Research Data Management Needs" and will be documented at URL http://research.library.unt.edu/datares/. The central questions of our research include: How are universities actually responding in terms of policy to data management requirements from funding agencies? What are the practical needs of researchers to meet the demands of those requirements? And, finally, how can university libraries, and the library and information sciences field at large, address the needs of researchers for data management, retention, and sharing in terms of services and support, training, and infrastructure?

In our initial findings, we have learned that of the top 200 NSF and top 200 NIH awardee schools (approximately 220 institutions), only 22% have published policies supporting the retention and sharing of research data; of the top 50 NSF awardee schools, only 50% have published such policies. In a recent focus group with NSF Program Officers, they clearly articulated the importance of university-level support in terms of policy and infrastructure to facilitate the recent NSF requirement for data management plans in funding applications.

In general, we have found that university libraries have proactively stepped forward to meet the needs of researchers in preparing and implementing data management plans, sometimes in the absence of top-down institutional support. In the coming weeks, we will conduct a wide ranging survey of stakeholders in research data management – researchers, librarians, office of research administrators, provosts, and others – to further determine the needs and perceptions of this diverse community, as well as conducting further stakeholder focus groups at professional meetings and conferences. We will also conduct targeted surveys of administrators at those institutions without published data retention and sharing policies to better understand the decisions behind the absence of such policies.

Based on our preliminary findings, we believe that data management planning requirements on the part of federal funding agencies have the potential to stimulate significant changes in the way research communities think about the retention and sharing of their data. It is, however, critical, that funding agencies recognize that the preservation of research data does not come without costs in terms of staffing, infrastructure, and ongoing maintenance and repository services; as such, some guidance for incorporating these costs into research proposals may be in order. Likewise, a broader statement of support for open access to research data and published outputs from above the agency level would be extremely valuable to help facilitate the socialization of these policies in the broad research community, and in related support industries, including libraries and publishing.

Responses to RFI Questions on Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Federal agencies should require and enforce stronger mandates concerning the long-term stewardship of research data. A very simple step would be to a) require all federally funded research projects which produce research data to report to the granting agency a permanent URL representing a location in which the data produced in the research project may be found, and b) publish these URLs in public registries or existing web locations which list the grant awards made by the agency. This requirement could be easily audited on an ongoing basis by automated URL checking software to ensure that data continues to be publicly available.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Only digital data which is funded by public funds should be required to be maintained as publicly available. Copyright and other appropriate guarantees of intellectual property are distinct from this class of content and should be addressed separately.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The form and standards associated with data from different fields differ, and should be considered on a case by case basis. The recommendation in (1) above only makes the case that a permanent access point for research data should be provided, not the form the data should take.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Disciplinary differences are real, and (as above) must be considered on a case by case basis. Establishing a fundamental requirement for long term access to research results remains the first priority. Strategies and mechanisms for ensuring the long term survivability of research data is a matter for institutional innovations and is a worthy subject for grant-funded competitive experimentation. Federal agencies should reserve some funding for precisely this kind of long term sustainability research in order to increase the likelihood of successful innovations.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Best practices are still emerging, and practical innovations in this area of expertise should be explicitly fostered and catalyzed through federal dollars. The National Digital Infrastructure and Information Preservation Program (NDIIPP) was a fruitful initial program that involved NSF research efforts. Further federally-funded research in this emerging area is needed to make progress in long term sustainability of digital data.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

One of the priorities for research needed in sustaining research data concerns setting reasonable cost standards for such expenditures. Without such guidelines and concomitant best practices, there will likely be unhelpfully confused variation in requests and responses.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

See the response to (1) above. Automated compliance verification is a manageable burden under this strategy.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Standardized access registries as described above would greatly improve access to research data by industry for new and existing markets. Further, innovations in access mechanisms for searching and culling this registry data would be a new growth industry in its own right.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Provenance of research data for attribution and scientific credit purposes will be improved by publicly verifiable registries as described here. Minimal requirements for attribution could be required as part of the URL strategy described here by linking the permanent data accessibility URLs to registry entries.

Responses to RFI Questions on Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

This is again a disciplinary-specific topic. Interoperability standards arise naturally for specific fields as researchers seek to share data. It is an area that could be catalyzed for some fields by making competitive grant awards available for applicants that adopt or offer to collaboratively establish emerging standards.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Obviously there are many examples to cite, here is one: The ISO WARC standard format for web archives emerged from discussions in the web archiving community on simple strategies for making the results of web crawls more manageable. The content of web crawls is stored in standard formats for long-term access and preservation. This community standard is publicly documented, see the URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717 for full information.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

A first step is to convene international meetings to explore this topic. A good example of such an exploratory meeting was the Aligning National Approaches to Digital Preservation (ANADP) conference recently held in Europe (see http://www.educopia.org/events/ANADP). Such meetings should have associated specific outcomes set forth in the initial planning.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Many protocols are in development for this purpose. A promising candidate is the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification, documented at URL http://www.openarchives.org/ore/.

Thank you again for the opportunity to respond to your RFI. This issue is of tremendous importance to American institutions of higher education, from administration, researchers, students and libraries. If you have any questions concerning any of these responses, do not hesitate to contact us.

Sincerely yours,

Martin Halbert, PhD, MLIS

Dean of Libraries and Associate Professor

Thu 1/12/2012 7:04 PM
Response to RFI FR Doc No: 2011-28621, digital data

General comments:

If publications are the currency of science, then data is the collateral behind the currency's value. By mandating the sharing of this collateral, it changes the way science is transacted. Some scientists have embraced this and fantastic discoveries have emerged. Other scientists are not as enthusiastic. As Wilbanks (2012) puts it "The ugly reality is that sharing data represents a net economic loss in the eyes of many researchers: it takes time and effort to make the data useful to third parties (through annotation and metadata) and that is time that could be spent exploiting the data to make new discoveries….There is pervasive fear that other scientists will "scoop" them if their data are available before being fully explored" (para. 5). Before embarking on technological or financial resolutions, there should be recognition that sharing data may violate long-held beliefs. Only clearly articulating policies, incentives, and minimizing undue burdens on researchers and institutions can overcome this cultural barrier. In addition, "We do not have the sociotechnical infrastructure required to answer questions of data stewardship with any authority" (Wilbanks, 2012, para 8). However, because you ask, I will try to answer.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

While the current data sharing mandates are laudable for their intent to increase exchange of information, which in turn should increase innovation and economic prosperity, they need further clarification. First, the goal should be scientific reproducibility and data re-use, not making data available for the sake of availability. Second, there should be continuing review of policy based on examples of effective data re-use. If policies are to be informed by evidence, then evidence must be collected and evaluated by economists, computer scientists, information managers and others who are qualified to determine the required innovations, costs and trade-offs required to meet the goal. Funding should be provided to study the effectiveness of current data sharing practices and the best use of resources for future data sharing.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Current intellectual property tools have a difficult time accommodating all expressions and uses of data (spreadsheets, computer code, database queries, etc.). While freely available data would be the ideal for creating new innovations and stimulating the economy, this is not always desirable. As well, even freely available data needs to be attributed correctly to protect the scientists' efforts. Placing the burden of responsibility on the scientist leads to confusion regarding issues such as the use of derivatives and designations of non-commercial versus non-profit. This confusion may result in unnecessarily conservative copyright and/or licensing. Alternatively, practices that are too liberal may lead to the loss of commercial potential for the institution or scientist, run contra to export control regulations, or endanger vulnerable populations. The federal government should encourage research institutions to craft intellectual property tools and educational programs for their researchers. This would enable scientists to apply appropriate copyright and licensing to their output. Each institution should create a clear and unambiguous policy on when and where data can be freely re-used, specific to the unique

potential of the data and in alignment with the spirit of data sharing. As well, institutions must provide legal consultation services to scientists as standards and mandates change.

(2) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The current situation is that researchers do not necessarily come equipped with expertise in databases and curation activities and information management professionals may lack subject-specific knowledge about science data. Indeed, it is highly improbable for any one person to have expertise in all physical and life science data requirements, the infrastructure necessary to handle the data, and the curation activities that makes the data findable and re-usable. The only solutions are interdisciplinary.

To facilitate interdisciplinary collaboration, the federal government could sponsor interdisciplinary working groups. These working groups may not follow traditional lines of discipline separation (i.e. departmental hierarchies) and may be best identified by asking professional associations. Professional associations tend to have specialty or interdisciplinary subgroups that may represent discrete data practices. As groups of researchers are identified to have common data practices, they may then delineate how those practices meet, or fail to meet, the data sharing initiatives. Each working group should have access to expertise from the broadly associated disciplines. These broadly associated disciplines include metadata specialists, infrastructure experts, and legal counselors.

It may not be possible for certain groups to release data without significant detriment to commercial goals, research programs, or other contrary regulations. These groups may need waivers or accommodations when faced with data sharing mandates. One suggestion is that "Individual disciplines and communities can opt-out of funder-wide approaches if they make a strong public case that the principles and goals are not applicable to their area, or that they plan to achieve the same goals in a different but equally-effective way " (Piwowar, 2012, #3). Should disciplines not be prepared to either share data or defend why, then they need to elucidate current practices and explore future options. On the other hand, disciplines that have currently have 'dark repositories' or that desire specific data sharing services would be discovered. Specifically requesting that major professional associations report on their constituents' data sharing practices would identify discipline specific differences. Alternatively, simply making data management plans publicly available would allow information managers to get at discipline specific practices and to suggest alternatives.

Lastly, once discipline specific practices have been identified, they need to be unified. In other words, their standards, languages, and metadata schemas need to be interoperable. Without intervention, these standards "will not spontaneously emerge …as long as data are in a tower of Babel of formats, incoherent names, and might move about every day, they will be a slippery surface on which to build value and create jobs" (Wilbanks, 2012, para. 13). Funding agencies need to cite examples of verified emerging standards and stimulate new interoperability through challenges, prizes, and expanding grant opportunities. In particular, the ability for new repositories to federate with existing ones may drastically increase their survival.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Policies need to be informed by working groups that are primarily populated by the scientists. They know best what the relative costs and benefits are of long-term stewardship and dissemination of their particular data. Polling their professional societies and involving economic analysis would go a long way to answering this question in a discipline specific manner.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

While data management plans have served to bring data management issues to the forefront, they reflect the high variability of available options and willingness to share. Since the plans themselves are so highly variable, the requirements for implementation are highly variable as well. To date, librarians have provided consultations, examples and tools for data management plan creation (DMPTool, etc.) and continue to explore options such as data repositories. The Association of Research Libraries has provided a structured course for research librarians to explore these topics and provide recommended actions to their institutions. However, no one entity will be able to answer all the challenges.

Research communities must contribute or effective solutions will not evolve. Universities must actively support their faculty with data sharing by providing legal consultation, infrastructure, and information management expertise. Scientific publishers need to provide avenues for data sharing and work with institutions to apply appropriate copyright and licensing. In particular, publishers must clearly state how they are handling data copyright and ensure that it is compatible with institutions' and scientists' needs. As well, scientific publishers should require and provide unique identifiers and citation of data sets. There are several endeavors currently underway that would facilitate this (datacite, doi's, etc.). There is work for everyone in this endeavor.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Funding agencies should promote digital data sharing by clarifying existing funding mechanisms and targeting new ones. Specifically, repositories need a funding mechanism for that period of time between start-up and when they have accumulated a critical mass of information, value-added services, and a strong user base. Across campuses, there are researchers who compile databases in an effort to organize and use their own output more efficiently. At times, their colleagues wish to contribute to this effort and a repository is born. Since funding agencies typically evaluate grant proposals based on novelty and the ability to generate new ideas, there is rarely funding for maintaining and improving existing repositories (Bastow & Leonelli, 2010). As researchers struggle to find funds and expertise, these 'dark repositories' languish in obscurity.

The logical entity to incubate a burgeoning repository is the institution's library. However, library budgets are not increasing and their service expectations are not decreasing. Therefore, an investment in a data repository must be carefully considered as their cost is "an order of

magnitude greater than that suggested for a typical institutional repository focused on e-publications" (Beagrie, 2008, para. 7). There is evidence that this increased cost is largely attributable to staff efforts in documentation, formatting, and ingest - not necessarily to the archival storage (Beagrie, 2008, Chapter 10). The expertise to properly document and ingest documents usually exists in the library, an entity that funding agencies typically consider an indirect cost, and therefore not eligible to charge fees directly to an individual grant award (OMB Circular A-21, F8). This classification and cost structure is a dilemma for libraries. If a data repository is to be available widely, it would be a major function of the institution, and should be covered under facilities and administrative (indirect) costs. If the data repository will only be used by a few disciplines then it should be charged to those projects that require and use the service (a direct cost to specific grant funds). In reality, any data repository will likely start with a few heavy users and then either generalize to accommodate a whole community or specialize to a particular discipline at a national or international level. How then, should the cost of such services be re-captured? Successful data subject repositories typically subsist on several sources of income, including private and public funds, and even subscriptions. The best solution is to provide a separate budget line for all activities surrounding data sharing (including proper documentation and ingestion) and to allow those funds to go to whatever entity, public or private, that provides the required services. This may also discourage the current practice of eliminating all funds for data dissemination when the proposal isn't fully funded. Otherwise, the letter of the mandate may be met, but the ultimate goal, re-use, will be hampered by inadequate documentation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The simplest approach is to add an information field to the existing funding agency reports. In other words, require that data management plans be verified by a digital object identifier (doi) or uniform resource locator (URL) of where the data is shared. These doi's or URL's could even be published with final reports and summary publications. Safe places to deposit data will be preferentially used, creating new data repositories or increasing the use of existing ones. Proper repositories will specify proper citation techniques, and as data sets are cited, they can be tracked. This is analogous to tracking journal article citations. As we know, the Impact Factor from the Thompson Reuters Science Citation Index has been used to determine tenure, promotion, and publication preferences. Perhaps similar measurements of data re-use will evolve and be used as an incentive for data sharing. (This also applies to question (9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?)

As well, funding agencies need to develop guidelines for reviewers to evaluate data management plans. The inability to distinguish a good data management plan from a bad one negates their value. Grading data management plans and data sharing efforts will help define best practices and improve compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There are three major barriers to wide-spread data re-use. The first is ensuring that the proper intellectual property tools are developed and used, so that scientists are informed and protected. The second is federating existing data repositories through interoperable standards. This increases the ability to locate data. The third is development of analysis and visualization tools. For industry, the most important barrier is problematic intellectual property rights.

Thank you for your time,

Amanda K Rinehart

Very Interested American Citizen

References:

Bastow, R, & Leonelli, S. (September 17, 2010). Sustainable digital infrastructure. EMBO reports (2010) **11**, 730 – 734. doi:10.1038/embor.2010.145

Beagrie, N., Chruszcz, J., & Lavoie, B. (May 12, 2008). Keeping research data safe (Phase 1). Retrieved from http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

Piwowar, H. **(**January 11, 2012**)**. A View Of The Rights And Responsibilities Of The NSF Wrt Data. Retreived from http://researchremix.wordpress.com/2012/01/11/nsf-data-vision/

Wilbanks, J. (January 11, 2012). Response to RFI on digital data. Retrieved from http://del-fi.org/post/15710035064/response-to-rfi-on-digital-data

Thu 1/12/2012 10:59 PM
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Note – A previous draft was sent to this e-mail and should be disregarded. This is the correct version.

Name/Email

Leslie D. McIntosh, PhD, MPH

Jon Corson-Rikert
Director, Mann Library IT Services

Affiliation/Organization

Washington University in St. Louis (LMc)

Cornell University (JCR)

City, State

St. Louis, MO

Ithaca, NY

General Comment

Giving researchers credit and recognition for work they do with data in the form of career advancements and awards will do the most to assure public access to and preservation of digital data. While Federal policies can further the requirements for digital data preservation, discoverability, and access, I believe it will be data scientists themselves that will move this effort forward.

Comment 1

Federal policies can best foster reuse of data by encouraging the adoption of open, interoperable standards for data exchange, notably Linked Open Data (http://linkeddata.org). Linked data reduces the barriers to sharing data and encourages the adoption of ontologies to more clearly express what the data represent.

Comment 2

Reporting requirements for grants could be modified in coordination with data management plan guidelines to require reporting a permanent web address for data, the nature of any access restrictions, and all significant contributors to the data.

Comment 3

Federal agency policies on the management of data should encourage the development of standards for documenting the content of digital datasets using the classifications and terminology of each discipline, ideally through non-proprietary ontologies and controlled vocabularies developed by members of the discipline itself. If datasets within any discipline are documented with explicit references to established ontologies and authoritative databases, the effort to accommodate differences among disciplines can be addressed at the discipline rather than the individual dataset level.

Comment 4

Archivists have developed policies and procedures to assess, and periodically re-assess, the significance of the artifacts they preserve, in large part because it may not be possible to predict the uniqueness, utility, or quality of artifacts until some time has passed. Agency policies should factor in re-assessment of the costs and benefits of continuing to preserve and/or migrate forward digital data, using experts representing the original domain and the most likely domains for data reuse.

Comment 5

Stakeholders can best contribute to the implementation of data management plans by promoting their adoption and encouraging the evolution of review standards through experience with both the costs and benefits of different levels of access and different investments in preservation.

Comment 6

The biggest challenge with current funding mechanisms is the time shift between the period of the award and the need for preservation. Data management plans could require an escrow process to set aside grant funds and assure access to them through a rolling funding model that would use current research funds to pay for the continued preservation of datasets still deemed worthy of preservation.

Comment 7

Datasets can be peer-reviewed much like journal articles. Standards for the review process could be developed; however, there are very few guidelines given when reviewing manuscripts, so the

process could be a very similar.

Datasets can be peer-reviewed much like journal articles. Upon submitting a dataset, it could be made available in a repository with information indicating the level at which the data have been reviewed (e.g. none, manually curated, used by independent source to replicate other findings).

Comment 8

Marketplaces for data and related services have been emerging on their own; federal funds could help support basic infrastructure costs for data registries and to establish workable policies and at least minimal subsidies for preservation of digital data when accompanied by viable business plans.

Comment 9

For individuals and groups to be given appropriate attribution and credit for data used, the data must be identifiable and discoverable, and metadata sufficient for ready discovery must be created for datasets and disseminated in a citable fashion. Once the basic mechanisms for discovery of datasets are in place on the Web, then the data owners can be cited in the same fashion that grants are now cited in publications. Journals should require authors to cite data sources and the authors/curators of the data in order for a manuscript to be published.

Comment 10

Use of any non-proprietary controlled vocabulary and explicit references to accepted international standard scientific and other disciplinary-focused databases will go a long way.

The ANDS/VIVO ontology (http://blogs.unimelb.edu.au/vivoands/2011/07/06/the-vivo-ands-ontology/) is a lightweight extension to the VIVO ontology (http://vivoweb.org/ontology/core) used to submit metadata about research datasets in university collections in the format required for the Australian National Data Service (http://www.ands.org.au/resource/rif-cs.html).

Comment 11

The Open Biological and Biomedical Ontologies (http://obofoundry.org) are examples of standards that have been developed through an open, collaborative process across several disciplines in the life sciences. Open processes that maintain quality standards and encourage iterative improvement through consensus have a higher likelihood of adoption and ongoing maintenance.

Comment 12

A promising way to promote effective international coordination of digital data standards would be to fund a tool that allows for open adoption and development for data, similar to what VIVO (http://vivoweb.org) has done for linking researchers.

Additionally, federal government grants should encourage US Citizens to travel abroad to professional meetings using government grants. This allows personal connections to be made, which facilitate future collaborative work.

Comment 13

Library and government repositories can encourage and, when appropriate, require submission of datasets associated with publications through modification of the publication submission and review process.

Leslie D. McIntosh, PhD, MPH
Center for Biomedical Informatics | Washington University School of Medicine

[Assigned ID #]

[Assigned Entry date]

Name/Email

David W. Robinson, Ph.D.
Executive Vice Provost

Affiliation/Organization

Oregon Health & Science University

City, State

Portland, OR

Comment 1:

What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Free, timely, and re-usable access to the digital data resulting from federally funded scientific research will improve science and human health by promoting transparency, efficiency, and reproducibility. To realize this value, federal policy must speak to several key issues. Policy and practice should create a technical infrastructure that addresses discovery, usability, attribution, and long-term preservation. These specifications must include the archiving of data in publically accessible repositories, using standard record formats, and promoting best practices for interoperability and reuse, such as Semantic Web standards and Linked Open Data. Funding agencies should support awardee data management and compliancy efforts by integrating these expenditures into grant structures and through interagency standards that offer transparent and practical workflows. Finally, award and incentive systems must evolve to recognize the value of data management and sharing to the scientific enterprise.

Markets will emerge to support the management and usability of the data; and, the economy will benefit from derivative products and services. The U.S. can look to the Human Genome Project (HGP) as a strong proof of concept. The HPG has led to groundbreaking discoveries and therapies. For example, our pioneering faculty member Dr. Brian Druker's development of the cancer drug Gleevec is intrinsically linked to the research sharing and advances the HGP fostered. Initially, nearly four billion dollars was invested in the HGP. Since its inception, an entire industry has developed to support genomic research and R&D. The ROI is dramatic; in 2010, the industry produced $67 billion in U.S. economic output, $20 billion in personal income for U.S. citizens, and 310 thousand jobs.[1]

Increasingly, governments, funding agencies, and research institutions are recognizing the scientific, societal, and economic benefits of open data. Since 1999, the NIH has required that crystallography data be submitted to the Protein Data Bank (PDB) upon journal manuscript publication. On December 12, 2011, the European Commission launched an open data strategy

for Europe. Announcing the policy, the Vice President of the European Commission stated that openness is the best strategy for gleaning value from data.[2]

Comment 2:

What specific steps can be taken to protect the intellectual property interests of publishers, scientists, federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Recognition of data sets as a citable, authored sets of information will help protect the interests of individual stakeholders. This provides a framework for recognition that is based on the currently accepted model of citation within publications. Additionally, when appropriate, a phased approach to access can be taken to protect the IP interests of stakeholders. Time-limited embargo periods could be utilized to manage both access and re-use rights. For example, rights holders may choose to restrict commercial re-use for a specific time period. However, embargo periods should not impede depositing of data in a repository.

Comment 3:

How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

In aiming to build policy that acknowledges differences among scientific disciplines, it will be important to leverage the knowledge of scientific communities and experts in the organization of knowledge, such as libraries and information science researchers who are accustomed to providing guidance and resources for disparate kinds of data. Existing discipline/data specific repositories should also be consulted to ensure applicability.

While data management needs differ by discipline, there are qualities and practices that underpin all data types and should inform inter-disciplinary requirements. For instance, there exist upper ontologies that represent the types of things that exist. Classification of data elements can be tied to such upper ontologies via reuse of these upper ontologies. One example is the Basic Formal Ontology as the upper level ontology for all Open Biomedical Ontologies (OBO), which enables representation of things as diverse as a mammary gland, a PCR machine, and mitosis. Similarly, while each discipline's data may require specialized formats, queries and applications, if the federal agencies promote open and extensible standards, the different needs will be met.

Additionally, it must be recognized that data is utilized in ways disconnected from the creator's original research focus. Data from disparate disciplines are combined and analyzed to advance scientific inquiry and support markets. Interoperability standards will benefit these new applications.

Comment 4:

How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

There are data management requirements common to all disciplines and data types; and, there are discipline- and data-type-specific needs and expectations. Agencies should build cost and responsibility frameworks that account for both the shared and unique needs. Mutual requirements can be met via interagency collaboration, standardization, and cost sharing. Unique requirements can be met via discipline and agency specific support and innovation. Such a framework has the potential to control costs and maximize benefits by limiting duplicate efforts, distributing responsibility—for both shared and unique needs—and, encouraging public-private partnerships.

EUDAT, a European based data infrastructure initiative, is currently pursing this strategy, and its work should influence U.S. agencies. EUDAT states, "Although research communities from different disciplines have different ambitions, particularly with respect to data organization and content, they also share basic service requirements. This commonality makes it possible to establish generic…services designed to support multiple communities, as part of a Collaborative Data Infrastructure."[3] The figure below illustrates this framework of collaboration and distribution.

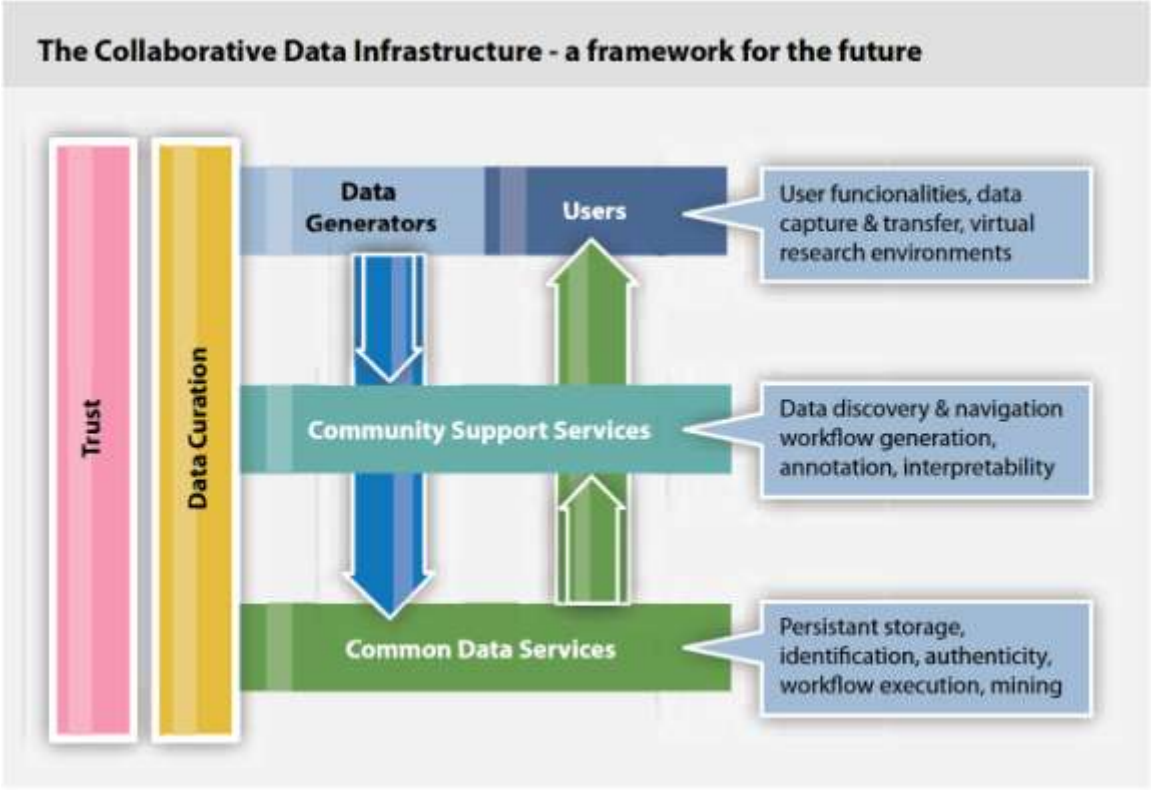


Image courtesy of International Science Grid this Week: http://www.isgtw.org/

Comment 5:

How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

There is room and opportunity for many contributors; however, it is critical that participation is regulated by technical and legal standards (e.g. use rights) that ensure and promote free public access, discovery, re-use, and preservation. The expertise, technologies, and infrastructures of stakeholders should be leveraged both in the development of policy frameworks and their execution. Such collaboration will drive best practices, innovation, market creation, and compliance. The present repositories of research communities, publishers, and institutions can be utilized and developed (e.g. Pangea, TreeBase). Existing partnerships between publishers and repositories, such as Dryad, can be grown. Organizations like DataCite work to improve the discoverability and utility of data. Universities, research institutions, and libraries have been and should continue to be key contributors, building infrastructures to support their researchers' compliancy, as with NIH public access policy, and guiding archival and discovery standards. Libraries are also well positioned to enable these infrastructures to be compliant with the Semantic Web and population of Linked Open Data from these data sources.

Comment 6:

How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Maximum scientific and economic rewards will not be realized if the cost of data management and preservation are an after-thought.[4] Researchers and institutions should be required to document the real cost of data management and publication within their proposals and reports. The cost and effort associated with doing so should be accommodated for in the total budget. Funding mechanisms, requests for proposals, and agency budgets must also address the real costs of long-term preservation, the latter being independent of grant-specific costs. In this regard, leveraging and supporting the services and expertise of institutions and organizations with memory driven missions (e.g. libraries) should be considered. Agencies should consider funding libraries to perform more research "in the field" on making specific data types conform to standards and archived for maximum searchability.

Comment 7:

What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Interagency standards that offer transparent and practical workflows and mandate deposit in publically accessible repositories will improve compliance, facilitate verification, and reduce burden. Ideally, such a systematic approach would require that:

- All data are deposited in publically accessible databases in conjunction with manuscript acceptance.
- Standardizations of record formats and minimum metadata are applied and verified.
- Several submission workflows are supported, including third-party deposit.

- Data and manuscripts are assigned persistent, linked identifiers.
- Compliancy is demonstrated through key events in the research enterprise via persistent, citable data identifiers.

In contrast to data management and access policies organized around the individual researcher or lab, a systemized approach reduces burden by enabling home institutions, libraries, and scientific communities to build effective support services.

Comment 8:

What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

In cases where technical and legal openness are supported, examples of innovation and market growth abound. Open access to US National Weather Service data underpins a large and diverse industry, estimated by the American Metrologic Society to exceed $1.5 billion per year.[5] "Between 1988 and 2010 the human genome sequencing projects, associated research and industry activity—directly and indirectly—generated an economic (output) impact of $796 billion, personal income exceeding $244 billion, and 3.8 million job-years of employment."[1]

To stimulate innovative use and grow the economy, technical infrastructure must support the sophisticated needs of human and machine readers; and, legal infrastructure must support liberal re-use rights, including non-exclusive commercial development. For example, data should be archived according to standards that support multiple formats, using standards metadata and complying with current best practices of data sharing and integration over the Web (e.g. Semantic Web standards and Linked Open data). Licensing frameworks, such as the Create Commons CC-BY, offer a starting foundation for building a license for data that will stimulate investment in new capabilities and applications.

Finally, agencies and institutions should promote their data wealth and encourage its use. The World Bank opened its data in April 2010; in October 2011, it launched the Apps for Development contest, challenging the developer community to create tools and applications using World Bank data. The contest rules ensured that contestants would retain the intellectual property rights of their software. Developers from 36 countries responded, submitting software, mobile apps, games, and widgets aimed at policy makers, educators, health care providers, and the public. Agency promotion of open scientific data would be sure to spur similar participation and innovation, and small funding opportunities or similar contests could be created to do so.

Comment 9:

What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

As with manuscript publication, secondary results should cite primary data. Standardized data identifiers will provide the means to identify and link the resource to other relevant documents,

data, persons, etc. The use of controlled author and institutional identifiers (e.g. ORCID registry) will be critical to support disambiguated and resolvable attribution. Furthermore, use of a common metadata standard to tag various kinds of data with appropriate attribution in a standardized way will ensure proper attribution. It is not always enough to know whom the data came from, but also the version, from where, and how is it related to other documents, data, experiments and grants.

Comment 10:

What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

It would be a burdensome and unproductive to identify, set, and enforce discipline-specific content standards. Rather, what should be pursued are standards that optimize discovery and use by human and machine readers. This includes format standards, minimum metadata requirements, Semantic Web standards, and Linked Open Data. A minimum metadata standard for any kind of content should be created - whether it is a data set, a publication, a patent, a grant, and ontology, a blog, etc. Anything that is reportable as linked to grant funding activity should meet this minimum metadata standard.

Comment 11:

What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful

There are examples of successful standards development in various domains. The W3C standards development process has successfully produced HTML, XML, RDF and other languages. Key to the process is its openness and community participation. Successful standards development relies on the contributions of a diverse population of experts, including scientists, information professionals, and technologists.

Comment 12:

How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

All federal agencies should pursue a technical infrastructure for data that utilizes international standards for interoperability and re-use, such as Semantic Web Standards and Linked Data. Agencies can and should leverage the work of organizations focused on international data sharing and utility, such as CODATA, the Global Biodiversity Information Facility, the Open Archives Initiative, and the Digital Curation Center. It would also be worthwhile for federal agencies to participate in and support international efforts to connect data collections and build

collaborative data infrastructures that aim to deliver cross-disciplinary data services. Similarly, adoption of other international efforts to standardize metadata, for example, coordination between VIVO and CERIF, the European organization for international research information, will facilitate data integration internationally. Promoting such coordination as part of existing granting mechanisms or via new ones to promote international collaboration will be beneficial.

Comment 13:

What policies, practices, and standards are needed to support linking between publications and associated data?

In order to support manageable and meaningful linking, standards and practices must speak to the required use of persistent, unique identifiers for data, publications, authors, and institutions. Unique identifiers will strengthen the visibility of each item and the links between items, as well as enable re-use and the development of new services. This will require a new age of semantic awareness on part of the researcher, the reviewers and the publishers of manuscripts and data. Publishers and granting agencies need to improve their standards regarding unique identification of research entities, and guidelines to authors need to be generated in support of these new standards. Furthermore, enhancing current research training to include these modern information management strategies will be key.

Response to this RFI is voluntary. Responders are free to address any or all the above items, as well as provide additional information that they think is relevant to developing policies consistent with increased preservation and dissemination of broadly useful digital data resulting from federally funded research. Please note that the Government will not pay for response preparation or for the use of any information contained in the response.

1. Battelle Technology Partnership Practice. (2011). Economic Impact of the Human Genome Project. Retrieved December 13, 2011, from http://www.battelle.org/spotlight/5-11-11_genome.aspx
2. European Commission launches Open Data Strategy for Europe | Open Knowledge Foundation Blog. (n.d.). Retrieved December 27, 2011, from http://blog.okfn.org/2011/12/12/european-commission-launches-open-data-strategy-for-europe/
3. Approach | EUDAT. (n.d.). Retrieved December 27, 2011, from http://www.eudat.eu/approach
4. Oecd Follow Up Group. (2003). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal, 3 (29)*
5. Annex 1 – Best Practice and Emerging Evidence - Economic Growth | data.gov.uk. (n.d.). Retrieved December 27, 2011, from http://data.gov.uk/opendataconsultation/annex-1/economic-growth#_ftn9

In accordance with Section 103(b)(6) of the America COMPETES Reauthorization Act of 2010 (ACRA; Pub. L. 111-358), this Request for Information (RFI) offers the opportunity for interested individuals and organizations to provide recommendations on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research. The public input provided through this Notice will inform deliberations of the National Science and Technology Council's Interagency Working Group on Digital Data.

Andrew Sallans andrew.sallans@gmail.com

Bill Corey bill.corey@gmail.com

Sherry Lake sherrylake@comcast.net

Individuals

Charlottesville, VA 22903

Responses

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

- **Require data management plans.** Data management plans are an essential part of the process of capturing critical contextual information about any data package. It is critical that data be managed during the entire lifecycle of a research project. Requiring a management plan in addition to a sharing plan will assure that data is in a state where it can be shared, with meaning and understanding, when accessed. It is essential that data management plans be required for all preserved research data resulting from federally-funded research, and it is essential that the guidelines be consistent at the high level, and relevant at the localized level, in order to be most useful for each data package.
- **Require submission of digital data into a data repository.** First preference should be discipline-specific or subject-based, second preference should be national or regional, third preference should be institution-specific. All digital data should also be submitted to the host institution's repository if one is available. Interlinking and networking of data repositories and metadata (ie. DataONE as a notable effort). The Library of Congress should evolve into being the nation's data repository in the same way that it is the nation's repository for many other types of materials.
- **Develop a national infrastructure for all repositories such that all federally-funded digital data is available openly and without restrictions from a common portal.** Funders should support the infrastructure necessary for the digital data resulting from the research they fund. The NSB (2005 - http://www.nsf.gov/pubs/2005/nsb0540/) report on long-lived digital data: "Participants agreed to a considerable extent on the main policy issues, even though there is one stark difference between NSF and many other agencies: the vast majority of long-lived data collections supported by the NSF are managed by external research organizations, while other agencies, such as the National Aeronautics and Space Administration (NASA) and the National

Oceanographic and Atmospheric Administration (NOAA) focus more heavily on archiving and curating many such data collections themselves." There should be an "if all else fails" option available on a national/regional/state level that will accept digital data from government funded projects whose results don't 'fit' into any of the discipline-, subject-, or institution-specific repositories. Existing data centers, particularly those that receive federal funds, should be strongly encouraged to accept and curate a broader range of digital data in their disciplines, and they should include data from research that is peripheral to their primary focus to encourage inter-disciplinary research.

- **Establish programs for reuse of scientific data.** Stable, secure funding of infrastructure and expertise would allow the growth of a knowledgebase, a foundation for functionally enabling long-term cross- and interdisciplinary data reuse. Metadata is at the heart of digital data reuse, curation and preservation; the capture and standardization of metadata is paramount. It is at the core of many data reuse issues, from discovery to trust, cost, curation, migration, and data quality. Metadata acquisition and/or creation must be an integral part of the digital data repository picture; it should be part of both the technological process of creating the data through research and the social aspects of sharing those research results.
- **Require open access to federally funded research.** Open access enables everyone to stay on top of the current science and research trends, generating new ideas and uses for research results, opening up new windows for educational opportunities, thereby reinforcing the cycle of growth and creativity. Require that all digital data generated by grants provided by federal funders be deposited in a manner that ensures open access to everyone for maximum accessibility and reusability. This will increase citations rates, encourage follow-on research and increase the likelihood of new cross-discipline and inter-discipline research initiatives. [See Piwowar et al: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308]. The resulting research will generate new opportunities for commercial development for the original researchers and also for others who incorporate that data in their own research. It will encourage private investment of research by capitalizing on a public resource, thereby launching new products or services into the marketplace. Commercialization is more likely to happen on data or a database that is fully open and has no restrictions on access or use. [See http://www.battelle.org/publications/humangenomeproject.pdf].

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

- **Clarify the contractual rights, responsibilities, and obligations of federally-funded scientific research in layman's terms.** Scientific researchers working in the higher education academic environment are frequently entangled between the competing interests of federal funding regulations, institutional policy, and commercial opportunities. There is a major opportunity for simplifying the uncertainty in this scenario through clarification of intellectually property rights for those conducting federally funded scientific research.
- **Select a data citation standard for all federally-funded research.** Choose a standard data citation that takes into account reusability, merging of data and versioning so attribution and reuse can work in unison. Additionally, specification of data licensing requirements will simplify full reuse and proper attribution, for both the original work and all subsequent reuses.

- **Specify publisher embargo rights for federally-funded research.** Be detailed and offer clear policy on publisher rights to embargo data for a specific period of time after which it is transferred to an open access repository.
- **Include intellectual property rights and policies outlined above in data standards.** Ensure that approved metadata standards include the approved data licenses, embargo periods, and citation mechanisms.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

- **Establish a vision, clear principles, and a framework at the core, delegate responsibility for localized practice and implementation.** Foster a culture at all levels of the organization which recognizes that policies must be flexible and will need to evolve over time, and create a mechanism to manage those processes, which includes many stakeholders, including scientists, researchers, data managers, funders, publishers, curators, and research institution administrators. In accordance with that plan, require that all funding agencies and their sub-divisions specify the appropriate data management guidelines for their disciplines, as long as they meet the essential base criteria set forth in the overarching principles.
- **Bi-directional communication amongst all stakeholder groups is essential when trying to bridge disparate domains.** [See: "A critical challenge in making policy formation a dynamic, interactive process involving all stakeholders" (Parsons, 2011)].
- **Develop an infrastructure including multiple layers of abstraction, such as seen in a federated web-of-repositories** [Baker and Yarmey: http://www.ijdc.net/index.php/ijdc/article/view/115/118]; a strategy for representing data through both deep domain understandings as well as cross-disciplinary mappings.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

- **Develop a web of domain-based data centers with an ongoing federally-funded mandate to ensure long-term access to data.** There should be funding from federal and private funders for digital data curation. The 'value' of a given digital data set will evolve over time; it may change disciplines, move into a cross-discipline category or even a wholly new discipline, be broken into smaller subsets, or combined with other subsets to create a new data entity. The original digital data set must be held long-term to ensure this level of creativity and evolution.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

- **Develop and provide tools and assistance for the creation of data management plans.** See the DMPTool (https://dmp.cdlib.org/) as a community-driven example of a tool developed for support of data management planning. Likewise, research institutions should collaborate to create a knowledge base of DMPs that have been accepted and approved for funding to provide assistance in policy and tool development, and to assist researchers with the writing of their DMPs. Additionally, stakeholders should collaborate on or develop and provide tools for metadata capture and reuse, management and operation of domain-based infrastructure, and in the development of standards.

- **Clarification of institutional intellectual property policies.** Provide clear and precise information on intellectual property policies at the institutional level. Additionally, provide clear and precise information on open access policies at the institutional and publisher levels, and on contracts at the institutional and publisher levels.
- **Provide information of best practices for data archiving, data sharing and data curation.** Researchers should focus on the research while stakeholders in support capacities advise them on best practices for their DMPs.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

- **Develop guidelines for the curation of data long-term.** Although there is growing awareness of the costs of preserving data long-term, there is significantly less attention given to the processes involved in managing and preparing data for long-term preservation. Specific guidelines and budgetary specifications as they relate to best practices would be valuable in furthering the goals of accessible digital data.
- **Provide funding for the first five years of data management and data curation in the grant-approved repository.** Provide funding in the approved grants for the initial deposit of digital data into an open access or approved domain-, discipline-, subject-, or institution-specific repository. Alternately, provide a mandate and funding for an open access national repository, or a set of open access regional repositories for the retention, sharing, preservation, migration, curation and management of all federally-funded research data.
- **Require a business plan for the research data lifecycle.** Paying for the collection and short-term management of digital data is not enough. Addressing long-term costs, and how to switch from a short-term project data mindset to an indefinite long-term data mindset is challenging. A business plan that addresses the issue of transitioning from federally-funded research to other sources should be required. Business plans could be developed through collaboration with the repositories and institutions directly involved with digital data curation. This would be a non-issue if all digital data went to federally-funded repository infrastructure for long-term preservation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

- **Require that future funding be contingent on full compliance.** If DOIs are required for all federally-funded data sets, then all projects should be required to report the DOIs in final project reports and in future proposals. Additionally, clear and precise information on what 'full compliance' means at all levels.
- **Clearinghouse for all federal funder compliance requirements.** Develop and create a portal that brings together all federal funders compliance requirements, and is available to all researchers, institutions, research communities, libraries, scientific publishers and other stakeholders. (Similar to the SHERPA/RoMEO publishers copyright & self-archiving and SHERPA-JULIET research funders open access policies sites). Also similar to the Office of Management and Budget (OMB) website in some regards.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

- **Establish short-term and smaller grant programs for data reuse projects.** As data reuse requires less start-up expense and time investment, projects could be shorter and less expensive, essentially new analysis initiatives. Give these projects priority and push these programs as an incentive and means of promoting new possibilities.
- **Make data open and available.** Increasing access to digital data can lead to indirect benefits within college and K-12 instructional environments, as well as research environments. This can lead to new software and new small business ventures to add value to the freely-available data packages. Additionally, there might be notable opportunities in the creation of funding opportunities at the college level for undergraduates and graduates to reuse digital data sets to create new ideas and technologies or at the high school level for science, technology, engineering and math students to develop ideas and projects, and to encourage participation at the next stage of their education.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

- **Select appropriate community data citation and attribution standards.** One example would be the CC BY or ODbL type licenses for data which will allow for full reuse and require that credit be given to those who did the work, both the original work and all subsequent reuses.
- **Endorsement of major community data citation initiatives.** One notable example is the DataCite (http://datacite.org) effort.

Standards for Interoperability, Reuse and Repurposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

- **No single standard for everything.** Unfortunately, it's unlikely that a single standard will meet all needs. Just the same, too many standards are also a bad thing. There should be a healthy awareness that standards are emerging and require time to be refined and stabilized through iteration and implementation. Communities should drive development of data standards, but should be guided by independent standards organizations like the ISO.
- **Open, non-proprietary standards.** Most importantly, standards should be developed in an open way and should be non-proprietary, as a means of fostering widespread interoperability, reuse, and repurposing. Standards should evolve independently of software or versions.
- **Crosswalking is the key.** For optimal interoperability, standards should be developed with the expectation that they will in some way need to relate to other standards. Although different types of contents will require different data standards, each will also need to relate in some way in the broader framework of digital data. Clear and accurate documentation of data standards will allow for "crosswalking" of one standard over to another for higher-level aggregation.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

- **The Knowledge Network for Biocomplexity** (KNB - http://knb.ecoinformatics.org/index.jsp) , a national network that helps facilitate ecological and environmental research on biocomplexity, has produced a very successful suite of tools for data and metadata creation, discovery, and analysis. The community is a highly-distributed set of field stations, laboratories, research sites, and individual researchers. It has been successful by developing software products for the community with the community's help and by providing education (training seminars available on the website) and outreach.
- **The Ecological Metadata language (EML)** is the standard metadata specification used by KNB. The EML project was created by the Ecoinformatics.org, a voluntary organization, whose goal was to produce "services that are beneficial to the ecological and environmental sciences".

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

- **Clear identification and agreement on digital data standards.** Before an effective coordination of digital data standards can take place, digital data standards must be identified. A big problem is the discovery of existing standards. When standards are assumed non-existent, communities create their own.
- **Co-development of digital data standards.** Federal agencies should take a leadership role in development of digital data standards across international boundaries, but must involve stakeholder communities in the process in order to increase adoption rates. Development of the standards must occur jointly among those in the stakeholder groups across international boundaries.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

- **DOI's for data and author ("Person") ids** are needed to support linking between publications and associated data. These ids are needed for all data, whether associated with a publication or not. We recommend Federal agencies encourage and support international initiatives which support such principles, such as DataCite (http://datacite.org) and ORCHID (http://orcid.org/). Some publishers, including the Nature Publishing Group and PLoS ONE (open access peer-reviewed journal), are already requiring accession numbers and/or DOIs for supplemental data.
- **Policies like the DRYAD "Joint Data Archiving Policy"** (http://datadryad.org/jdap) should be recognized as a model policy, as it supports further linking between publications and associated data by requiring authors to deposit the data that supports the results of papers published within DRYAD journals.

Submitted By:

Andrew Sallans
Head of Strategic Data Initiatives
Scientific Data Consulting Group

Sherry Lake
Senior Scientific Data Consultant
Scientific Data Consulting Group

Bill Corey
Scientific Data Consultant
Scientific Data Consulting Group

Response to Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

National Snow and Ice Data Center
Boulder, CO
January 2012

The National Snow and Ice Data Center (NSIDC) is a leader in the data science community with decades of experience supporting science through the archiving of ethically open polar and cryospheric data from around the world, including data produced with NASA, NOAA, NSF and other agency funding. As an organization, we are committed to the principles of open data and data stewardship, and we strongly support the IWGDD and NSTC efforts to further develop and harmonize policies for the ethical sharing, access, and preservation of digital data resulting from federally funded scientific research.

Data and metadata issues - including capture, access, discovery, standardization, security, ethics, preservation, interoperability, and synthesis among others – are all extremely complex and dynamic. Therefore, we believe that policies will play a strong though partial role in shifting the trajectory of research science towards a culture of open and ethical data sharing.

The NSIDC response to RFI questions follows:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Based on years of experience in data science, NSIDC recommends a clear, government-wide policy supporting open and ethical data sharing through federated repositories with professional expertise.

Data provide greatest value when they are viewed as a common, networked good rather than as individual or intellectual property. As such, data should be as fully open as possible with restrictions based only on ethical rather than proprietary concerns. Our more than thirty years experience, especially our involvement with the very large, interdisciplinary International Polar Year (IPY), has taught us several lessons on policies around open data.

1. Clear, explicit policy mandating that research data be openly available soon after collection helps make data systems simpler, more robust, and more adaptable.
2. Limited ethical restrictions should be clearly identified. For example, the IPY Data Policy notes legitimate ethical restrictions of data about human subjects, local and traditional knowledge, and where data release may cause harm.
3. We see the greatest success in achieving timely and open data access when data are required to be deposited in an open archive. The requirement should be supported by identified, funded archives and professional data scientists working with the data providers.
4. Licenses, contracts, and other legal agreements restrict the usability and interoperability of data and our ability to address interdisciplinary challenges, because they restrict our ability to use machines to discover, manipulate, and

repurpose data. Wherever possible, legal proprietary rights to data should be waived and data should be exposed to the web in a way that machines can readily interpret as open. Ethical considerations such as fair attribution and accurate documentation of quality should be based on scientific norms rather than legal mandate. This is the concept of an information commons. See, for example, polarcommons.org.

5. While various agencies are recognizing the value of shared data services and are funding development of shared repositories, tools, and services, the funding in many cases has followed the traditional term-limited model. Data preservation and access efforts have been hampered by this short-term structure, and data curated by term-funding in some cases have been put at risk when the repository funding ends. Secure and stable funding of infrastructure and expertise would allow the growth of a knowledgebase, a foundation for functionally enabling long-term interdisciplinary data reuse.

How these policies help grow the U.S. economy and improve the productivity of science will be influenced by a number of factors including the rate of cultural change within the domain communities. Policy, in concert with other efforts such as the including basic data management skills in academic curricula and the supporting ongoing data science research, work in concert to promote change and advance scientific understanding.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

In our view, the critical consideration is the advancement of science and not of scientific institutions or individuals per se. As discussed above, publicly funded data (as opposed to creative works) should be viewed as a common, networked good. Proprietary considerations should be minimal to non-existent. In the current domain, data sharing is uneven across individuals and disciplines. This creates an unequal playing field for scientific researchers. Funding agencies can address this. They can demand that researchers all play by the same rules of openness and they can provide greater recognition and support for data work. This de-emphasis of individual knowledge in favor of an information commons approach provides the greatest overall benefit to the U.S. and indeed to humanity.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The policy-creation process should remain dynamic and flexible, encouraging participation in governance from multiple stakeholder groups while acknowledging that policy alone is not enough to motivate full compliance. Policy must at some point either align with practices on the ground or remain abstracted; engagement with stakeholders and addressing their differing concerns through an iterative process helps align policy with practices. As practices are in a constant state of change, continual bi-directional communication amongst all stakeholder groups is essential when trying to bridge disparate

domains. "A critical challenge [is] in making policy formation a dynamic, interactive process involving all stakeholders" (Parsons, 2011).

Supporting the work of data scientists is important as well; they break down barriers between projects and domains and bring data into an interdisciplinary context. Domain-based data centers for instance, can translate the data and complex local context of that data from individual PIs into standardized form that others from outside domains can access, understand and reuse. Supporting data scientists at local, domain, and cross-domain levels, as might be implemented through a federated web-of-repositories (Baker and Yarmey, 2009), represents a strategy for preserving data and making them accessible for reuse at multiple levels. Leveraging data science expertise maintains the deep domain knowledge of local work through cross-disciplinary mappings into the broad context of interdisciplinary research.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

First, proposal-stage data management planning should include consideration of resulting data products and costs for maintaining data through their entire lifecycle. "End-to-end data management (which includes acquiring, processing, storing, maintaining, updating, and providing access to data) should be planned and budgeted at the outset of any activity that will generate environmental data. This planning should explicitly address data archiving, data stewardship, and data access responsibilities, and sufficient funds should be provided to archive and provide ready and easy access to the resulting data for extended periods of time" (NRC, 2007).

Second, professional data managers should be supported as part of data access and preservation infrastructures (NSB, 2005). This mediating layer of expertise provides a lens through which the costs and benefits of data preservation can be weighed. Data scientists contribute cross-cutting vision and interdisciplinary understanding to multi-stakeholder discussions on maximizing the long-term benefit of valuable data resources.

Third, different levels of service should be created and applied for the preservation and access of each data set depending on importance and community applications (Weaver et al., 2008). These are potential responsibilities for data scientists in collaboration with other stakeholders. Policies should recognize the range of relative contributions and applications of diverse data sets and support strategic delegation of resources by data scientists and curators.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Generally - For the successful implementation of data management plans, many stakeholders with a diversity of roles, responsibilities and expertise must come together under a common goal of ethical data sharing. Recognizing the need for a diversity of perspectives at the table – including those of the scientist and research team, data curator, technologists, metadata experts, digital archivist, user services personnel, and others – is an

initial step towards supporting DMP implementation. Within the context of each support organization, roles and responsibilities must be clearly defined and delegated to the stakeholders with the appropriate expertise. All stakeholders should promote policy compliance along with community-based standards-making.

Universities and research organizations - Clarify intellectual property statements regarding data and promote data publication and sharing as part of tenure considerations. Recognize and support coordinated data management and infrastructure systems and services as valuable institutional facilities for researchers. Offer career paths for data managers and data scientists on campus. Include data management training as part of the science curriculum.

Research communities – Recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement. Proactively contribute to standards-making discussions and the cultural shift towards data sharing and complete metadata capture.

Academic Research Libraries – Consider data as not only part of the scholarly communication cycle, but as a resource to be curated. Apply extensive experience in cataloging and bringing together diverse, interdisciplinary resources to the data preservation and access paradigms.

Publishers – Enable and encourage ethical data sharing and attribution through citations within publications and research documentation. Recommend authors make associated data freely available through open data repositories.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

With the recognition that data are a common scientific good, it becomes imperative that 1) the substantial investment in generating data for research must be protected through preserving and making openly accessible those data for use in additional research, and 2) that the funding responsibility must be borne across all of the sciences. Quality data stewardship starts in the proposal stages, is enacted through the field- or experiment-based management of generated data, and continues through the useful lifespan of data products. Funding should include all aspects of data management; if you pay to collect the data, pay to preserve them. Funding the support of data management could be viewed as "tax" on scientific research used to maintain valuable data resources for all science.

In many ways the real costs of preserving and making data accessible are just beginning to emerge. Recognizing that data preservation and stewardship are not solely technical problems is one step towards uncovering the true costs involved. The cost of data preservation for reuse includes not only the infrastructure and long-term system maintenance fees, but also the expenditure of defining, capturing and structuring necessary metadata, ongoing standardization work, prioritization of data resource allocation, and user support.

Especially for valuable observational data from the 'small' sciences, the largest cost involved potentially comes when not only making data accessible over the long-term, but making data useful well into the future. The amount and quality of metadata required for reuse potentially dwarf the metadata required for access and preservation. One of the biggest obstacles to data reuse is the capture and standardization of metadata. Metadata is at the core of many data reuse issues, including discovery, trust, cost and data quality. To realize the

goals of growing the economy and improving productivity of the scientific enterprise through the preservation of and access to research outputs, we must have a comprehensive strategy for metadata capture and preservation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Current trends in the academic sciences add to the burden of compliance and verification on PIs. First, changes in research funding, for instance a shift from large, single-domain labs to interdisciplinary collaborations of smaller labs, mean a gap in infrastructure support. The traditional lab-based infrastructure including local expertise and hardware has become more difficult to support. Recognizing this, shared resources are under development by many domain groups, universities and other communities of different sizes and foci. However, many of these efforts are operating in temporary funding environments and data access and preservation services remain piecemeal (ex. NBII).

Second, the continuing changes in technology, while opening up potential avenues of research, mean a constant demand on researchers to keep up with changing practices, communities, and policies. Funding for shared and coordinated resources at different levels can reduce researcher burdens by shifting some of the responsibility for metadata creation and structuring, translating requirements and keeping up with tools and services. For instance, the Advanced Cooperative Arctic Data and Information Service (ACADIS) project supported by NSF offers researchers a central portal for information, infrastructure, tools, and expertise to support their efforts to meet NSF Data Management Plan and data sharing requirements.

Along with community-based data management support, repositories like ACADIS are also a step towards quantifiable verification of policy requirements. The presence of a researcher's data in the ACADIS repository clearly indicates compliance with NSF program requirements to program managers. Generally, policies should require deposit into an openly accessible repository as a condition of continued funding. A federated network of repositories and a clear government-wide policy supporting sharing of ethically open data through use of these repositories enable generalizable measurement, verification, and compliance checking.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

NSIDC recommends a clear, government-wide policy supporting sharing of ethically open data through both human- and machine-accessible systems. Through basic APIs, open data are re-usable at scale and accessible to tools in addition to individuals. This allows people to build new and creative applications and services on top of the data that can provide new markets. Consider, for example, the growth of weather apps, visualizations, etc. that resulted from NOAA making their meteorological data more open and machine accessible.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data citation standards (Ball and Duke, 2011; ESIP, 2011), in concert with administrative metadata standards and publisher acceptance and support of expanding traditional attribution mechanisms are a first step towards proper attribution of data. Research into application of data citation standards is ongoing.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Data and metadata standards are emerging and require time to be refined and stabilized through implementation and iterative design. For interoperability, reuse and repurposing of digital scientific data, standards are needed at many different levels. Community-driven efforts to refine and formalize language and representation are necessary and important within a domain reuse and repurposing context. Broader, more general standards are necessary to promote interoperability across disciplines as well as nationally and internationally. A package of human- and machine-readable standards along with crosswalks, tools, and open and accessible data and metadata enable reuse and repurposing.

To encourage these combinations, spread support across international standards development and ontology research while recognizing the important role of community-driven standards creation and implementation. In the Earth sciences, the high-level ISO19115 metadata format and the NetCDF data format are emerging as useful standards. At a community level, the QARTOD->OGC effort has been specifically focused on quality specifications, data dictionaries and sensor-based metadata description for the oceanographic sciences. These examples are not only compatible but complimentary. For top-down, high-level standards to effectively intersect with bottom-up, detail-oriented standards and best practices there will need to be a decentralized model of support, communication, and governance driven by the clear goal of formal integration.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Application tools can often define a standard. For example, KML rapidly emerged as a community standard that was then adopted by OGC because of the popularity of Google Earth and other virtual globes. In another example, UNIDATA worked with the Atmospheric sciences community to develop tools along with netCDF to demonstrate how useful that standard could be.

Standards development needs to be an open, accessible, iterative, and well-supported process with emphasis on communication, community engagement and formal maintenance mechanisms. The Global Change Science Keywords (GCMD) provide an interesting example. They have been very popular and were central to the search interfaces and organization of many Earth science data systems around the world. The keywords were all

defined and openly accessible. The GCMD managed and controlled the list, but they were open to modifications from certain communities, if a need could be demonstrated. The keywords also had a certain amount of internal flexibility with a free form "detailed variable" that could be added anywhere at the end of the hierarchy. Now, however, the GCMD science keywords are losing relevance; they have not evolved well to keep up with modern data systems. The community has repeatedly asked the GCMD to make the keywords and their definitions available as a web service, but the GCMD has yet to provide one. Further, the GCMD made a major revision to the keywords without sufficient community consultation and the revision was not well received. As a result, the GCMD is using the new version internally, but it is not broadly used outside of the organization. External groups have begun to develop and use independent web services based on the older version of the keywords. This is bound to lead to some level of divergence in the "standard". Our point here is not to criticize the GCMD, which remains a valuable resource, but rather to highlight the need for community engagement, formal standard maintenance, and ongoing technical evolution.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

To promote effective coordination on digital data standards internationally, we recommend a multi-level strategy with the following three elements:

1. Work with international organizations such as ICSU and intergovernmental organizations such as WMO and UNEP to promote harmonization of data policy around ethical openness. The Belmont Forum may be an especially appropriate venue (http://www.igfagcr.org/index.php/belmont-forum).
2. Support national- or discipline-based data coordinators to work with counterparts in other nations around common international initiatives. There is growing international interest in standardizing Arctic observing data for example.
3. Seek collaborative science funding opportunities with individual nations or groups such as the EU around common interests like Arctic observing and then develop a common data policy as part of the collaborative effort.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Data citation standards and a clear policy mandating ethically open data accessible through both human- and machine-readable system are the first steps towards linking publications and associated data. With community acceptance of data citation practices and a persistent, machine-readable link to associated data in an open archive, tools to take advantage of the linked environment will emerge in response to specific needs.

References

Baker, KS, and L Yarmey. 2009. Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*. 4(2). Available online: http://www.ijdc.net/index.php/ijdc/article/view/115

Ball, A, and M Duke. 2011. 'Data Citation and Linking'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/

ESIP (Federation of Earth Science Information Partners). 2011. Interagency Data Stewardship/Citations/provider guidelines. Retrieved January 7, 2012 from: http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

NRC (National Research Council). 2007. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington, DC: National Academies Press. 116 pp. Available online: http://books.nap.edu/catalog.php?record_id=12017

NSB (National Science Board). 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Washington, DC: National Science Foundation. 87 pp.

Parsons, M. 2011. Expert Report on Data Policy and Open Access. GRDI2020. Available online: http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=e31a1aab-b01e-4e7e-9b10-0fd93d4b710f

Weaver, RLS, Meier, WM, and RM Duerr. 2008. Maintaining Data Records: Practical Decisions Required For Data Set Prioritization, Preservation, and Access. *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International 7-11 July.* 3:III-617 - III-619. Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4779423&isnumber=4779256

Response to Request for Information:
Public Access to Digital Data Resulting From Federally-Funded Scientific Research

From:
Sage Bionetworks
1100 Fairview Ave. N.
Seattle WA 98109

Summary
It is a critical time to implement and enforce effective rules for sharing federally-sponsored research results. Taxpayers have paid for the work and the entire community should have access.

The United States federal government is the largest funder of medical research in the world. As such its policies have a profound global impact on transforming research into improved healthcare. This is particularly true as we dive into a new world where mega-data genomic technology will require cooperation between data creators, data analysts and medical researchers to support the development of innovative therapies and personalized medicine.

Specific Comments related to the RFI:

(1) *Usable* copies of *all* federally-funded digital data should be placed in a publicly-accessible repository within a short period (30 days) of *creation* as an absolute condition of funding.

(2) Patentable Intellectual property rarely exists in primary digital data. Dissemination of federally funded research data does not normally compromise stakeholder intellectual property interests.

(3) Existing grant review panels have domain experts who can apply discipline-specific criteria.

(4) Long term stewardship is part of the full cost of research and should be handled as part of indirect cost negotiations

(5) Universities can foster the required cultural change intrinsic to rapid sharing of data by requiring and monitoring compliance in the same way that human subjects research is monitored and by using hiring and promotion criteria that include community contributions from data sharing. The former will only occur if required by funding agencies

(6) The real cost of data management should be included in indirect cost agreements.

(7) Applicants should be required to provide public repository ID numbers for every dataset referenced in grant progress reports, renewals and proposals as is currently required for publications.

(8) Funding agencies should support a range of pilot consortia focused on creating community-based platforms for effective storage and use of digital data outside of the traditional investigator-initiated, project-based funding programs.

(9) Data ID systems are a logical starting point for referencing and attributing credit.

(10, 11, 12) Absolute data standards are difficult as technologies evolve rapidly and in unexpected directions. Data repositories must be flexible and promote interoperable formats that allow workflow provenance.

(13) Publications can use data IDs as is already standards in many fields

Medical Research Context: Despite the sequencing of the human genome and tremendous advances in medical technology, genomic research has thus far not made significant contributions to healthcare. A critical factor is the inadequate sharing of genetic data, tools and results required for complex, data-intensive human biology research.

Medical research largely still occurs in isolated labs across the biomedical landscape using data that is not broadly accessible. Results are shared by publication in scientific journals after considerable delay and often lack sufficient detail for reproducibility. Without transparency, the quality of published models cannot be established and the refinement of modeling techniques cannot be pursued. There is also a prevalent attitude that federally-funded scientists who create data somehow own that data. This creates hoarding behaviors that are reinforced by research institutes and universities where credit shared may be promotions lost.

The research community needs to adopt a new set of behavioral norms to fully exploit genomic data where multiple investigators participate in, and are appropriately acknowledged and rewarded for, contributing to common, pre-competitive projects. This evolution needs clear demonstrations that breakthrough research findings require widespread and open community involvement as an incentive to change scientist behaviors. *Revised funding rules will change behaviors*.

Ultimately, progress is dependent on both changes in reward systems and the creation of forums for collaborative modeling with open repositories of curated and readily available and usable genomic data and disease models and tools for data analysis.
The interpretation of genomic data requires innovative computational methodologies. Integrative analyses that combine genotypic and molecular trait data to predict phenotypic outcomes provide a powerful means to improve the mechanistic understanding of disease and to develop novel approaches to treatment. Using computational modeling it is now possible to provide frameworks that describe complex physiological systems and predict the causal relationships between molecular states and disease. Early applications of these approaches have provided a boost in the understanding of disease pathologies. Despite the tremendous potential of this field, there are significant barriers to success related to the public availability of “usable” large-scale genomic data, the rapid development, evaluation and deployment of good methods to analyze the data, and the availability of disease models to drive downstream experimental validation and therapeutic development.

Integrative genomic analysis requires a move towards an open source, community-based modeling environment. It is a big challenge and no one organization has the resources to succeed in isolation. Rapid sharing of digital data and results is a prerequisite to make effective use of the large investments being made in genomic research. The success of many open-source software projects has demonstrated that distributed, decentralized and appropriately incentivized teams can effectively collaborate on complex and large-scale projects in an appropriate framework. The software industry has many examples of how such approaches has been transformative. Many of the most widely-used software projects are available for free and are open source (e.g., Android OS, Apache server, Firefox). Such infrastructure is increasingly hosted and managed by third party providers (e.g., SourceForge, GitHub, Google Code). This has allowed development teams to focus on truly novel areas thereby spawning entire new businesses (Facebook, Twitter, etc.). It has also transformed

the way in which software engineers receive recognition for their work and advancement in their career. For example a GitHub profile is beginning to complement or even replace a traditional resume in the software industry as it links directly to an individual's contributions on hosted projects.

Genomic science and public healthcare need ready access to digital data as well as open forums to compare, refine and deploy new methods for the analysis of high dimensional population-level genomic data.

Submission prepared by;
Jonathan Izant PhD
Vice President, Sage Bionetworks

About Sage Bionetworks: Sage Bionetworks is a 501(c)(3) nonprofit biomedical research organization created to change how researchers approach the complexity of human biological information and the treatment of disease.

Sage Bionetworks' mission has five interdependent themes:

- Research on computational network models of disease
- Pilot projects trialing disruptive models of research cooperation
- Rules and rewards that promote data sharing and collective research
- Building the computational platform for a digital Commons
- Activating public engagement and access

We are driving a cultural change around the elimination of disease by activating patients, shifting scientists to share the data and models needed to build better models of disease. To do this, we are building an open Commons called 'Synapse' where data can be shared and a compute space where predictive disease models can be co-evolved so that industry and academia can jointly benefit from understanding biology.

http://www.sagebase.org info@sagebase.org

Thu 1/12/2012 9:08 PM
RFI response

Public interests in data from federally funded research

A response to the Request for Information on Public Access to Digital Data Resulting from Federally Funded Research.

John Hawks, Ph.D.
Department of Anthropology
University of Wisconsin-Madison

Introduction

The United States provides grant funding to scientists through many federal programs. This funding advances work of public interest that might not happen without federal assistance.

The creation of scientific knowledge may serve the public interest directly by enabling useful inventions or supplying actionable information on issues of public importance. A funded project may also serve the public interest indirectly, by (1) finding negative results that prevent wasted effort or public harm; (2) building the scientific infrastructure that enables future discoveries and advances; (3) training new and established scientists in effective research techniques; (4) enhancing international cooperation and public/private partnerships.

Congress and the Executive Branch have recognized that access to the published results of scientific research is not sufficient to advance the direct and indirect public interests served by federally funded projects. Facilitating the indirect benefits of research is a major aim of federal agencies' "Broader Impacts" and data access rules. These policies have been a qualified success since their implementation, limited mainly by the exceptions carved out by programs and agencies to avoid requiring certain kinds of data to be reported along with research reports.

I argue that open public access to digital data should be a requirement for all federally funded scientific research. Digital data can be maintained by federal agencies as a part of the reporting requirement of federal grant funding. Doing so will advance the interest of the public and ensure that today's science generates a continuing heritage of research excellence.

Data access and transparency

Transparency is essential to public trust. Scientific conclusions are formed by observation and replication, and for this process to be transparent, all data must be available for independent inspection. The possibility of such inspection should not be limited to qualified researchers, because the very existence of special access requirements blocks transparency of the scientific process.

Changing technology has shifted the public's expectations about transparency. Digital technology enables most research data to be shared rapidly and at low cost. If data are produced in digital form, and digital data can be shared at low cost, researchers and agencies cannot credibly claim that the difficulty of reproducing and disseminating data is a sufficient reason to restrict access. Where no competing interest argues for restricted access (such as human subjects protections), a lack of access to digital data itself can now be a compelling reason for public distrust.

Therefore, federally funded researchers should release digital data to the public by default. Federal agencies should facilitate this public reporting by requiring digital data to be supplied as part of final project reporting.

Data access has a well-established record of success

The recent history of human genetics demonstrates that open access to data has unforeseen benefits that can spawn innovation, support more effective education, and catalyze new discovery. In genetics, both federal and journal policies require release of data; raw data from federally funded projects are often available as they are generated, long before publication.

My own laboratory has no federal research funding to date, but is actively engaged in research using data from federally funded projects. Today my laboratory trains undergraduate students in genetics with new data from ongoing federally funded genetic projects such as the 1000 Genomes Project. We use open access data from archaic human genomes to investigate the variation of ancient people and their relationships to living humans. This kind of work would be impractical without clearly established open data access policy.

The open access to data from the Human Genome Project facilitated the rapid development of microarrays that are now used on a broad scale in human genetics to investigate the genetic correlates of human health and disease. Access to data from these studies has enabled other scientists to independently replicate many genetic associations. More important, meta-analysis of such data has shown that many associations cannot be replicated, while also showing some cases in which nonsignificant results across different samples give rise to a significant finding when pooling those samples. Access to negative results and raw data is necessary, in other words, to establish the facts in subsequent research. This goes beyond access to published research results and requires open access to unpublished digital data.

Intellectual property protections and data access

Research data are somewhat distinct from the intellectual property issues relating to research publications. Some kinds of data do not meet the standard of originality necessary for copyright protection, such as sequence data, CT or MRI data, or data from measurement instruments. For raw data from instruments, there is no intellectual property reason why federal agency should not maintain an open archive for the public.

Much research data is unquestionably subject to copyright protection, such as lab notebooks, written descriptions, photographs, and original reconstructions. Yet there is still a substantial

public and scientific interest in inspecting such data. For example, photographic documentation of archaeological sites and specimens are of particular scientific value and are today routinely produced by digital technologies and stored in digital form. Some primary digital records are unique products that cannot be recreated at another time and place: for example, in situ photographs of specimens, photographs and records of sites before excavation, and digital reconstructions. The scientific record would be incomplete without such contributions, and maintaining an archive of such data over the long term is a difficult task for a single investigator, beyond the scope of a grant term.

In cases where it is impracticable to obtain Creative Commons or other open licenses to such content, a funding agency should at a minimum require that a copy of all such archival information be deposited along with the final project report and a limited-use non-commercial license permitting electronic dissemination of these materials to the public as part of the report.

Metadata and data access

Many have noted that raw data may be useless in the absence of additional information about how the data were obtained. Such information is known as "metadata". Researchers generate instrumental data using particular instrument settings and recording standards. They gather observational data under particular research protocols. These standards are may change quickly as instrumentation, technology, and scientific results themselves demand new practices.

Some scientists note the problem of incompatible metadata, using it as an argument against to delay the establishment of open public access to data. In their view, the public are likely to misunderstand or misuse scientific data where metadata are not clearly indicated. Meta-analyses combining data from multiple research projects are an important secondary use of digital data, and such meta-analyses are impossible when data cannot be reconciled into common observational or instrumental frameworks. Performing original work with data collected in heterogeneous contexts is a research speciality of its own, and is itself sometimes targeted by federal grants.

However, meta-analysis is only one purpose of data access. Transparency, replicability, and education are central public interests that do not require the reconciliation of data collection methods from multiple studies. They require only clear description of the methods under which data were obtained. At a minimum, final research reports on federally funded projects must describe the standards of data collection with sufficient detail to allow independent replication, including all unpublished results and data.

Data access in paleoanthropology

I am an anthropologist, and am most familiar with the scientific data relating to human evolution. These data include genetic observations on living and skeletal samples of humans. They also include fossil and archaeological evidence such as photographs, CT scans, isotopic records, anatomical measurements and descriptions.

Successes of data access in paleoanthropology

For many years, nearly all genetic data resulting from federally funded research have been made available for public download. Much genetic data generated by non-federally funded research programs, including foreign and domestic institutes, has also been free for public download. These data have resulted in a massive acceleration of research on recent human evolution and human origins. They have also led to unexpected discoveries and a burgeoning contribution of other disciplines to understanding our evolution.

Data from radiocarbon dating and other isotopic sampling has also been made available to the public. Human occupation sites are among the best sources of evidence about past climates. The investment of federal resources in human evolution research has generated a temporal record that is now essential to studying changes in the faunal and plant compositions of past environments. Free access to records has enabled stronger calibration of radiocarbon dates, the development of a more secure chronology, and a more highly replicable scientific record correlating different regions of the world. Our understanding of such events changes is vastly stronger when data are made public.

Institutions and data access in paleoanthropology

By contrast, CT scans and photographs pertaining to human origins are typically not made accessible by the public. The United States funding agencies are not the only parties with an interest in such data. In particular, museums and institutes that curate specimens often permit data collection under agreements that restrict the dissemination of the resulting data. Such agreements may be equated to "non-disclosure agreements" with respect to scientific data.

An institution has a legitimate interest in controlling the public use of images and access to curated materials. Nevertheless, the lack of access to digital data results in reduplication of effort, overapplication of destructive sampling and measurement techniques, and unnecessary handling of precious and fragile specimens. Where it is practical, the United States should facilitate agreements with institutions that allow the release of digital data produced by public funding. Where release is not possible, funding should be granted only for those activities that will result in the release of data under a limited-use non-commercial license. Non-disclosure of data from instruments such as CT scanners, electron microscopes, mass spectrometers is incompatible with scientific replication.

Scientific careers and data access in paleoanthropology

The economy of federal funding for scientific production sometimes leads to perverse incentives for high-ranking researchers that prevent public access to research data. Some scientists believe that their own future research will require exclusive access to data. Others want to impede research achievements by their academic rivals, or to maintain prestige and future funding opportunities.

Scientific data in some areas may constitute "trade secrets" until they are protected by patents. Even in noncommercial research, federally funded scientists sometimes claim exclusive

ownership over data that they plan to use in future research. In my own field of paleoanthropology, data secrecy supports a clandestine "quid pro quo" economy among researchers, in which established researchers and institutions allow furtive looks at unpublished data, to support and consolidate their power and influence.

This is a game that the United States should simply decline to play. When federal research supports scientific results that are not subject to independent replication, it betrays the public interest in science.

Established collaborations and centers of scientific research will always exert a strong influence the future of science irrespective of federal data access policies. But established players should not use federal funding to construct barriers to open inquiry.

Conclusion

Open public access to data is one indication that a research project is following scientific principles. Making digital data available to the public would be good practice for any researcher, irrespective of funding source. Data access mitigates the risk that negative data will be unreported. Data access facilitates broader stewardship of research projects, in particular where collaborations create data that are distributed across many institutions. Data access and reporting standards enable other researchers to fill in for those who cannot complete scientific project due to health or other personal reasons.

Federal grant agencies already have successful repositories for many kinds of digital data. Such data are shared with the public at minimal cost relative to the overall budget for federal research grants. Supporting digital data repositories has itself been an important granting aim for several federal agencies and continues to be an active part of scientific infrastructure. Limiting such repositories for the exclusive use of a small cadre of researchers is enormously wasteful of resources, when they can be opened to an interested public for a small incremental cost.

The public has repeatedly invented surprising uses for digital data that can complement or enhance the scientific record. But much more important, open access to digital data serves the scientific values of transparency and independent replication, essential to maintaining public trust and investment in the research enterprise.

John Hawks
Department of Anthropology
University of Wisconsin-Madison
http://johnhawks.net/weblog

Virginia Tech's University Libraries responds to the Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting From Federally Funded Research (http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28621.htm)

1. *What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

   All digital data resulting from taxpayer-funded research should be freely available to the taxpayer. This will minimize barriers to information that can be used to drive develop and commercialize new products and services by entrepreneurs either working alone, in small businesses, or in large corporations.

2. *What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?*

   An agency's requirements for data curation and open access should not undermine any patent rights under current or enacted laws or treaties for funded investigators or intuitions.

3. *How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?*

   Individual researchers may not have expertise in the establishment of efficient databases, while information management professionals may not have expertise in the use or form of data across multiple disciplines. Therefore, it would be helpful to develop a core set of baseline standards that are applicable to all disciplines, then develop further discipline-specific standards and guidelines where necessary.

4. *How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?*

   Funding agencies should consider these issues when they review research proposals. If the research is likely to result in data that is costly to manage, but offers little value to the research community, then it may be a weak proposal in the first place.

5. *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

6. *How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?*

   Agency research sponsorship should provide adequate funding for research and archival activities (such as that provided by university libraries) to accommodate all data curation and archival requirements of federal funding agencies. Consideration should be given towards the funding requirements for acquiring, processing, and preserving access to research data, including consideration for the cost of bandwidth and disaster plans.

7. *What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?*

   Federal agencies (NSF, NIH, NARA, NOAA, DOE, EPA, LOC, NARA, etc.) and other stakeholders need to identify best practices. Requirements are currently too vague to help researchers comply. Once requirements are clear and measureable, it will be easier to verify compliance. Furthermore, it will be easier to verify compliance through systematic approaches. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

8. *What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?*

   Provide free tools and services to visualize this data in ways that are useful to researchers, consumers, voters, business owners, investors, etc. These tools and services could be developed through grants, or by federal institutions.

9. *What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?*

   At the institutional level, and organizational level, attribution of data sources should be as important as attribution of other sources. Librarians, funders, researchers, publishers, professional organizations and other stakeholders need to identify citation requirements for repurposed data.

10. *What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.*

XML is a good standard for storing the metadata that corresponds with the research data, including instructive information for how to access the data if it has particular platform requirements.

11. *What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?*

12. *How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?*

    Federal agencies should coordinate international digital data standards by working with existing standards bodies such as NISO and ISO.

13. *What policies, practices, and standards are needed to support linking between publications and associated data?*

    One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

Thu 1/12/2012 10:27 PM
Data RFI – input

Professor Victoria Stodden
Department of Statistics
Columbia University
New York, NY

http://stodden.net

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Federal funding agencies must require the digital datasets and computer instructions needed to reproduce published computational results be made openly available. In the pre-computational days, empirical scientists would describe their methods carefully in published papers with the intent that others would then be able the reproduce their work. Now, with the pervasive use of computers in scientific research, the steps taken in generating published findings are immensely complex and impossible to capture in the short methods sections as before such that the findings can be replicated. This is causing an enormous credibility crisis in computational science. It is impossible to reproduce the vast majority of results published in journals or presented at conferences today.

Sharing the information necessary to replicate the computational aspects of the experiment is a necessary response to this crisis. This means revealing the computer instructions, the code and scripts, as well as the datasets these instructions acted upon to produce the published results, at the time of publication. There is an emergent movement to create reproducible computational science, people and groups voicing concerns and creating sharing solutions from fields as diverse and geoscience, signal processing, statistics, bioinformatics and –omics research, MRI processing and neuroscience and many others.
These folks are working against a collective action problem: sharing is extra work for the scientist and not likely to seen as personally beneficial. Given that the incentive structure faced by computational scientists is heavily influenced by funding agency requirements, responsive funding agency policy is a necessary part of the solution.

One size does not fit all research problems across all research communities, and a heavy-handed general release requirement across agencies could result in de jure compliance – release of data and code as per the letter of the law – without the extra effort necessary to create usable data and code facilitating reproducibility (and extensions) of the results. The National Science Foundation now has a database of Data Management Plans and can collate this information to learn what is a reasonable sharing requirement in each field. These data would permit federal funding agencies to craft release requirements that are more sensitive to barriers researchers face and the demands of their particular research problems, and implement strategies for enforcement of these requirements.

This approach also permits researchers to address confidentiality and privacy issues associated with their research. I would hope for the funding agencies to move aggressively, then adjust if problems are encountered. The standard must be replication of the results by a contemporary in the field, without having to contact the original authors.

Data and code sharing requirements are not a foreign concept to scientists. The prestigious journal, Science, now requires authors to relinquish code and data from all articles they publish to any enquirer, for example.

To make the concept of data sharing coherent, it must connect to the reason scientists have the norm of sharing methods in the first place:
Reproducibility.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

The Supreme Court has established that raw facts are not copyrightable. The sharing of datasets would presumably contain raw facts, although the datasets may have some residual copyright if they meet the standard of "original selection and arrangement." Thus the intellectual property status of datasets is not as clear as for written scientific articles. It is exceptionally important to preserve open access and reuse of the datasets, and I feel certain publishers will not do this adequately since they have not done this for published articles and it is important that the data and code that underlie published results do not become the property of the publishing houses. These must be openly available to facilitate reproducibility as well as transfer the knowledge and methods behinds the results beyond the ivory tower. I have previously published an intellectual property framework for scientific research, called the Reproducible Research Standard (cf. V. Stodden "Enabling Reproducible Research: Licensing For Scientific Innovation" at http://www.ijclp.net/issue_13.html) to untangle intellectual property rights associated with research release and clarify requirements.

The Reproducible Research Standard (RRS) realigns the Intellectual Property framework faced by computational researchers with longstanding scientific norms. The RRS suggests a licensing structure for research compendia, including code and data, that permits others to use and re-use code and data without having to obtain prior permission or assume a Fair Use exception to copyright, so long as attribution is given. The RRS utilizes existing open licenses that permit the free use of licensed work, so long as attribution is given, and is satisfied if the following four conditions hold:

1. The full research compendium, including code and data, is available on the Internet,

2. The media components such as text or figures, (including original selection and arrangement of the data), are licensed under the Creative Commons Attribution License 3.0 or released to the public domain under CC0,

3. The code components are licensed under one of Apache 2.0, the MIT License, or the Modified BSD license, or released to the public domain under CC0,

4. The data have been released into the public domain under CC0 or according to the Science Commons Open Data Protocol.

Using the RRS on all components of computational scholarship will encourage reproducible scientific investigation, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.
Moreover, in evaluating compliance, we would also want to encompass the ability to build, run, and verify any source code. This might be accomplished using

* spot checks of the repository
* automated checks akin to unit tests
* tests run by a separate reviewer at the time of inclusion

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The National Science Foundation has collected field-specific information on difference in data sharing through their Data Management Plan. I believe standard must come from the scientific community but they are foundered by a deep collective action problem – sharing data and code is a burden on the scientist and at the moment is not perceived as providing a payoff. The federal agencies should aggressively pursue data and code sharing policies that made the data and code that underlie published results conveniently available, such that the results can be replicated. Different communities can decide how to implement these standards but working toward these standards is imperative and federal policy leadership is a key part of doing so. Researchers with exceptionally different to share datasets or code bases should be apply for temporary waivers from federal sharing requirements, if more time is needed to create repositories or cloud access for example. Permanent waivers can be applied for on the basis of confidentiality or national security.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

More costly sharing solutions, such as for very large datasets or codes, would reasonably be expected to take longer to implement.
Otherwise this question misses the point of sharing: reproducibility of results. Data is not shared because of potential industrial applications or because downstream users may find it beneficial. These are important effects, but they are corollary. The reason to share is to ensure that what we purport to be scientific facts are indeed reproducible. This is why we are facing a credibility crisis in computational science today and why data and code sharing, such that the underlying results can be reproduced, is imperative, and federal agency policy must take a clear leadership position.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

In my experience there is enormous goodwill and desire to move toward greater reproducibility in computational science. The federal agencies can coordinate meetings between these stakeholders to implement data and code sharing plans. If a researcher does not receive further funding if data and code are unshared, this is the best contribution toward data sharing, and toward reproducible science.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Pilot projects and case studies. Fully fund some grants to be fully reproducible: sharing the data and code that generated the published results emerging from the study. Here are some examples of fully reproducible research that was very inexpensively shared:
http://sparselab.stanford.edu and
http://www-stat.stanford.edu/~wavelab . Both are enormous success stories with many downloads and many citations. The solution for many researchers does not need to be expensive or fancy, the policy just has to demand it. For other researchers, primarily those with very large datasets or codebases, they made need additional funding for cloud resources or repository creation for example. Research that was done with no sharing standards, other than the final published paper, is quite difficult to share in a reproducible way, but new research, where researchers are well aware that the data and code will be shared with publication, are much easier since efforts will be made as the research is ongoing. Choose several new grants and fund the applicants to create really reproducible research. This will provide measures of costs and needs beyond those described in data management plans.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Check if the data and code are openly available. This is trivially straightforward, for example following links in published articles, although does not verify the whether the data and code actually reproduce the published results. Leave it to the community to verify the results, but provide an avenue for downstream users to report whether or not they were able to replicate the results. This provides accountability and further evidence of compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

After the data and code are made openly available, then we can take a look and see how best to proceed on increasing access. Perhaps interfaces to the data and code will be the right way to proceed. At this stage it is imperative to just get the data and code open, without additional encumbrances for 3rd party usability. That can come later.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Scientists follow the norm of attribution. It must become standard practice to cite data and code use. Federal agencies can help provide stable URLs where data and code can reside through federal repositories.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

These are important but, as in the example cited in the question, must emerge from the community. Where federal agencies can help is facilitating the production of reports like the one in question by providing a mechanism for scientists in communities that lack agreement to apply for funding and engage a community meeting on the site of the funding agency to create standards.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Protein Databank (PDB) was created in 1971 and thus has 38 years of experience in becoming a standard within the structural biology community. It is funded by international agencies with hubs in three countries. A PDB “accession number” is a precondition for publication in computational biology, meaning your data is available in PDB. Phil Bourne, one of the founders of the PDB, has noted that some tweaking in the policy may be in order now since some researchers appear to be tempted to get the accession number very early in their work and then feel they then have a license to publish. One remedy I might humbly suggest is the inclusion of the concept of reproducibility of published results: accession numbers for the data the results were derived from, and another resource locator for the code that will permit others to replicate the published results, using the data.

The PDB has enabled a new field of statistical studies and molecular dynamics research around the deposited structural data, impossible without access to the data as part of each publication.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

I believe the most effective tool is convening meetings as described in the answer to question 10. Funding can be provided to bring international community members into the discussion.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

We could experiment. If linking is required for publications arising from federally funded research for, say, the next 5 years, and repositories provided for researchers without access to communities repositories, this approach can be tested. This is straightforward for funding agencies to verify, by randomly checking the links in articles published from federal funds.

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

Submitted by the Data Preservation Alliance for Social Science (Data-PASS)

January 12, 2012

Introduction to Data-PASS

The Data Preservation Alliance for the Social Sciences (http://Data-PASS.org) is a broad-based voluntary partnership of data archives dedicated to acquiring, cataloging, and preserving social science data, and to developing and advocating best practices in digital preservation. The Data-PASS partners collaborate to acquire data at risk of being lost to the research community; to develop preservation practices; and to create open infrastructure for collaborative cataloging and preservation.

Collectively, the founding partners have over 200 years of combined experience in social science data archiving. These partners include the Inter-university Consortium for Political and Social Research, The Roper Center for Public Opinion Research, The Howard W. Odum Institute for Research in Social Science, the Electronic and Special Media Records Service Division, National Archives and Records Administration, the Institute for Quantitative Social Sciences at Harvard University (which contains both the Harvard-MIT Data Center and the Henry A. Murray Archive), and the Social Science Data Archive at the University of California, Los Angeles (UCLA).

Thus far, the partnership has identified thousands of at-risk research studies (collections of data) and acquired many of these for permanent preservation. These range from data collections created under NSF (National Science Foundation) and NIH (National Institutes of Health) grants, to surveys conducted by private research organizations, to state-level polling data, to data records created by governmental research or administrative programs. [Gutmann, et *al*, 2009]

The preservation of quantitative data has a more extensive history and more well-established practices than in most other disciplines. Social science continues to rely heavily on data in its traditional forms, such as opinion polls, voting records, surveys, and government statistics and indices. On the other hand, although most large data sets are in public archives, most data produced by and used in social science research is neither publicly available nor preserved by an archival organization. And digital content is evolving into more forms than can be preserved readily. Changes in technology and society are greatly affecting the types and quantities of potential data available for social-scientific analysis. Any data describing human activity may be a subject of social science research. Taken as a whole, the evidence base of social science is shifting [King

2011], and consequently, approaches to curating this evidence, or data, is shifting as well

A National Digital Stewardship Alliance Founding Member, the Data-PASS partnership works to archive social science data collections at-risk of being lost; to catalog and promote access to data collections; to establish verifiable multi-institutional collaborative replication and stewardship of data; and to develop and advocate best practices in digital preservation.

Supporting Long-Term Access to the Scientific Evidence Base

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to digital data that result from federally funded scientific research. When applied, these values support the practical collaboration of private, public, and governmental memory institutions to support long-term access to research data.

Institutional Collaboration

The Data-PASS partnership is based on institutional collaboration, in which multiple organizations and virtual organizations adopt joint stewardship of collections. Partnership, is of course, an ancient approach, but the revolution in communication technology has lowered the barrier to widely distributed partnerships, and the ease of replication of digital content is enabling partnerships to take on a far more direct role in the stewardship of content.

Many threats to long-term access can be effectively ameliorated only when collections are replicated, geographically distributed, and audited by independent institutions. Independent replication and auditing reduce the risks of loss from software, hardware, and physical failure. Moreover replication across a diverse set of institutions that use a variety of business models and operate under different legal regimes insures against many forms of institutional risks, such as curatorial error, institutional mission change, and loss of funding. [Rosenthal, et al. 2005]

Building Mutually Reinforcing Infrastructure and Archival Practice

Institutional collaboration for long-term access requires institutions to establish mutual trust. And it is critical that there be a sound basis for this trust. Data-PASS is committed to good archival practice based on criteria for trustworthy organizations and partnerships that provide solid evidence. One widely recognized example of good archival practice is TRAC, the Trustworthy Repositories Audit & Certification Criteria and Checklist [CRL 2007], which is now in process of being reformulated as an ISO-standard.

In support of good archival practice, Data-PASS has developed open source infrastructure, SafeArchive [Altman & Crabtree 2011], that automates archival replication and auditing polices. The SafeArchive system provides a way to ensure that replicated collections are both institutionally and geographically distributed and to allow for the development of increasingly measurable and auditable trusted repository requirements. Designed as a virtual overlay network on LOCKSS [Rosenthal, et al. 2005], the system provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model. This enables any library, museum, or archive to audit that its content is being replicated across an existing LOCKSS network in conformance with documented archival policies; and to allow groups of collaborating institutions to automatically and verifiably replicate each others' content consistent with a set of expressed commitments. The result is that archives can more easily collaborate to preserve content through geographically and institutionally replication; which mitigates against technical and organizational threats to preservation.

SafeArchive and LOCKSS are exemplars of open community infrastructure that is designed explicitly to support long-term preservation and access. Another exemplar, developed by Partnership members, is the DataVerse Network. An exemplar of collaborative community efforts is the Dataverse Network project [King 2007] recently described by the National Research Council of the National Academies as the "State of the Practice in Data Sharing." [National Research Council 2011] The DVN is a data preservation and dissemination system that is based on open standards, supports open protocols, and integrates with systems such as LOCKSS to enable institutionally distributed replication and stewardship.

Interoperability through Open Standards

Long-term access to data, and effective collaboration to ensure it require the development of open protocols and information standards. For example, the Data-PASS shared catalog is based on open standards for metadata and metadata exchange. In the social sciences, many data producers and data archives have converged on the Data Documentation Initiative (DDI) metadata standard. [see http://www.ddialliance.org] The DDI specification provides a mechanism to document data in a structured, machine-actionable way. Combining information standards such as DDI with with open protocols such as the Open Archives Initiatives metadata harvesting protocol [Lagoze, et 2002], enables institutions to more effectively collaborate to manage and provide access to data.

Standards are equally important, and most often absent, for the use of data in scientific publications. Data-PASS actively promotes citation standards for research data. Accurate citation of data will promote more and better science. It will make data easier to find, to replicate, and to manage for the long term. Moreover it will make it much easier to trace the influence of data on social science. We advocating a simple baseline standard:

title, author, data, and a persistent identifier (of any widely recognized type, such as URN's, handles, DOI's). [See, for an example, Altman & King 2007]

Summary of Major Recommendations

- Support institutional collaboration to provide short and long-term access and stewardship of data, and to develop related standards and infrastructure

- Establish policy that requires auditable replication of data across multiple institutions that have demonstrated capacity and commitment to long-term access.

- Leverage substantial national and international efforts to develop archiving, metadata, and citations standards.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI.

In addition we fully endorse the recommendations of the National Digital Stewardship Alliance of which Data-PASS is a founding member. These recommendations may be found here (and are also attached with this submission):

http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf

Finally, institutional members of Data-PASS have submitted responses on behalf of their institutions. (Copies of these may be found here: http://www.data-pass.org/node/95) We support the principles of embodied in these responses and recommend that they be carefully considered.

References

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital Preservation Through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist*, 72(1).

Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

Gutmann, M., Abrahamson, M., Adams, M., Altman, M., Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R., Timms-Ferrara, L., & Young, C. (2009). From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data. *Library Trends*, 57(3).

Hedstrom, Margaret, Jinfang Niu, Kaye Marz, (2008). "Incentives for Data Producers to Create "Archive/Ready" Data: Implications for Archives and Records Management", *Proceedings of the Society of American Archivists Research Forum*.

King, G., (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, 36(2), 173-199

King, Gary. "The Changing Evidence Base of Social Science Research." In The Future of Political Science: 100 Perspectives, edited by Gary King, Kay Schlozman and Norman Nie. New York: Routledge Press, 2009.

Carl Lagoze, Herbert Van de Sompel,M. Nelson, M., & S. Warner, "The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0.", (2002). http://www.openarchives.org/OAI/openarchivesprotocol.html

National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

David S. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. "Requirements for Digital Preservation: A Bottom-Up Approach", D-Lib Magazine 11 no. 11 (2005)

NDSA
NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

Submitted by the National Digital Stewardship Alliance (NDSA)

January 2, 2012

Introduction to the NDSA

The National Digital Stewardship Alliance (NDSA) was founded in July 2010 to extend work begun in 2001 by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress. The Alliance has over 100 members from educational institutions, non-profit organizations, businesses and local, state and federal government agencies, as well affiliations with international organizations. Its mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. [1] Members of the Alliance are taking action to preserve access to our national digital heritage by:

- broadening access to our nation's expanding digital resources
- developing and coordinating sustainable infrastructures for the preservation of digital content
- advocating standards for the stewardship of digital objects
- building a community of practice around the management of distributed digital collections
- promoting innovation
- facilitating cooperation between government agencies, educational institutions, non-profit organizations, and commercial entities
- fostering the participation of diverse communities and relationships across boundaries
- raising public awareness of the enduring value of digital resources and the need for active stewardship of these national resources.

Supporting communities of practice for preservation and access

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally-funded scientific research. When applied, these values support the practical development of communities of practice capable of gaining consensus to support preservation and access to digital data. The shared expertise and common experience of these communities result in stakeholder buy-in and adoption of policies and

standards. The National Digital Stewardship Alliance member organizations are bound as a community by the following values.

Stewardship. Members of the NDSA are committed to managing digital content for current and long-term use. The members of the NDSA are actively ensuring sustained access to the digital content that constitutes our national legacy and empowers us as leaders in the global knowledge economy. Individually, these organizations support the management of digital resources; the Alliance is committed to protecting our nation's cultural, scientific, scholarly, and business heritage.

Collaboration. Collaborative work is the centering value of the Alliance; it is a value shared by all members and a priority in work with all organizations and associations. Approaching digital stewardship collaboratively allows the NDSA to coordinate effort, avoid duplicate work, build a community of practice, develop new preservation strategies, flexibly respond to a changing economic landscape, and build relationships to increase capacity to manage content beyond institutional boundaries.

Inclusiveness. The NDSA is a collaborative effort to preserve a distributed national digital collection for the benefit of current and future generations. We value the range of experience, the potential for innovation, and the fault-tolerance that heterogeneity brings. We believe the preservation of digital information is a pervasive challenge and that engaging across different communities strengthens the nation’s digital preservation practices and increases the likelihood of preserving content now and into the future.

Exchange. Members of the Alliance encourage the open exchange of ideas, services, and software. This leverages the commitments of each member to increase the capacity of the entire stewardship network. Participation and engagement result in innovations and benefits that can be shared by all. The Alliance is committed to transparency and all products generated or produced by the Alliance will be circulated under open licenses.

Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

Community-based approaches to the challenges of rapid change and high volume within the data domain have proven to be the most successful in the long term. The Blue Ribbon Task Force on Preservation and Access recommended that for research data “Each domain, through professional societies or other consensus making bodies, should set priorities for data selection, level of curation, and length of retention.” [2]

The report validated experience over the last ten years of digital preservation work. A study of the networks developed through the NDIIPP program indicated that participating institutions bring to the network their own resources, interests, and organizational culture. Under the auspices of a neutral convener and honest broker, natural networks emerge over time through participation in shared activities and problem solving. As these networks form, the larger network becomes more complex, but also stronger and better able to withstand stresses and strains. [3]

The Opportunities for Data Exchange (ODE) project supported by the Alliance for Permanent Access and the European Union also takes a cross-cutting community approach to preservation and access to digital data. “The potential answers to grand challenges of our times require…the inclusion of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-infrastructures…All stakeholders in the scientific process must be involved in the design of this layer; policy makers, funders, infrastructure operators, data centers, data providers and users, libraries and publishers…” [4]

An exemplar of collaborative community efforts is the Dataverse Network project [5] recently described by the National Research Council of the National Academies as the “State of the Practice in Data Sharing.” [6] The Dataverse Network is “unique in being designed to explicitly support long-term access and permanent preservation. To this end the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats.” [7]

Foster public values and support for stewardship of digital data beyond mandating data management plans.

Policy should assert the value of research data and provide mechanisms to support the preservation, discoverability and access. To relieve frustration and confusion about actions the policy should provide a clear direction for funders, researchers and stewardship organizations. The Blue Ribbon Task Force recommended “Funders should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving. They should explicitly recognize “data under stewardship” as a core indicator of scientific effort and include this information in standard reporting mechanisms.” [8]

Leverage substantial national and international efforts for common practices that support interoperability.

Substantial efforts have been made to pave the way for interoperability, re-use and re-purposing. Emerging practices for data citation, licensing and protocols for data sharing and sustainable re-use are becoming enough to adopt more broadly. Notable in these areas are work on the Data Seal of Approval by the Data Archiving and Networked Services that promotes sustainable access to digital research and provides training and advice about archiving and reuse.[9] LOCKSS is a community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content. [10] The

Data-PASS organization promotes collaborative, institutional stewardship of research data, permanent data archiving, and citation that permits results to be verified and re-purposed. [11] DataCite collaboratively addresses the challenges of making research data visible and accessible through data citation.[12] The Creative Commons project, Science Commons, has focused on protocols for sharing scientific data that includes licensing and mitigating legal barriers.[13]

Summary of Major Recommendations

- Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Establish policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Foster public values and support for stewardship of digital data beyond mandating data management plans.

- Leverage substantial national and international efforts for common practices that support interoperability.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI. The responses below exemplify how the principles can be applied to the individual questions, and highlight relevant NDSA activities in these areas.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The most effective policies in this regard would mandate data deposit into publicly accessible repositories. In the absence of such a policy, there are already cases of data which have been lost. The Federal policy framework should move public access to data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use.

Many members of NDSA provide repository services at low cost or through cooperative arrangements. Members of the NDSA also provide repository services that provide legal, technical, procedural and statistical controls necessary to protect data confidentiality while ensuring long. And the NDSA provides a model of institutional collaboration that supports stewardship, discovery and accessibility An example of a free access service is ViewShare.org, a platform for empowering curators, archivists, and librarians to provide access to the digital collections they are preserving through a shared interface. This

service provides the dual benefit of making data more broadly available and accessible while also making it easy for end users to copy and make use of the data in other environments. [14] The NDSA content working group is also working toward developing a clearinghouse for at-risk digital collections to help match data to potential preservation partners.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Each domain and discipline should be empowered to set priorities for data selection through, level of curation, and length of retention, through professional societies or other consensus making bodies.

Notwithstanding, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. For most data, "open access" is needed not only for the short term, but for the long term. And scientific disciplines have focused primarily on short-term access. There are critical standards for metadata exchange, fixity information and verification, and persistent citation that can support long-term access to data, preservation, and the long-term reproducibility of public results. Such baseline standards should be applied all scientific data. Among the range of important new standards for preservation and access there is still little knowledge about which standards are being implemented in which situations. The NDSA Standards working group is working on inventorying these standards and exploring how they are currently being used by NDSA member organizations. More than advocating the need for standards there is a clear need to understand which standards are being used in which situations and use that information to promote the usage of standards that are leading to results.

(5) *How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., the Data-PASS archives). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation being used loosely without associated evidence of meeting specific requirements. Memory institutions such as archives, libraries and museums have an extensive track record with these functions and collaborative organizations such as NDSA could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions. In this respect, work from the NDSA innovation working group toward developing a "Neighborhood Watch" system for repository quality assurance could serve as the basis for establishing clear, externally verifiable reporting. [14]. The group has identified a pressing need for

an objective, repeatable, independently verifiable and simple way for an external agent to periodically retrieve content, verify its bit level integrity and publicly announce the results. This is a clear example of how assertions about data management could be tested and certified.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The most important step would be to communicate that the real costs of preserving and making digital data accessible are indeed legitimate and necessary costs of the overall research enterprise. Researchers routinely include publication costs within their research proposals -- the costs of ensuring long-term access reuse of data should be treated in the same way.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Many NDSA partners are leaders in this area. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

References

[1] The National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa

[2] Berman, Francine, and Brian Lavoie, et al. 2010. *Sustainable Economics for a Digital Plant: Ensuring Long-term Access to Digital Information.* Final Report of the Blue

Ribbon Task Force on Sustainable Digital Preservation and Access supported by the National Science Foundation, et al. Washington, DC: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Library of Congress. 2010. *Preserving our Digital Heritage: The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report*. Washington, DC: http://1.usa.gov/hmw2lj

[4] Alliance for Permanent Access. 2011. "Opportunities for Data Exchange (ODE) Project": http://www.alliancepermanentaccess.org/index.php/current-projects/ode/

[5] King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173-99.

[6] National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

[7] Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

[8] Berman et al. 2010.

[9] Data Seal of Approval: http://www.datasealofapproval.org/

[10] LOCKSS: http://lockss.org

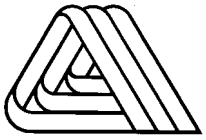
[11] DataPass: http://data-pass.org/

[12] DataCite: http://datacite.org/

[13] Creative Commons project, Science Commons: http://creativecommons.org/science http://wiki.creativecommons.org/Science

[14] ViewShare: http://viewshare.org

[15] Abrams, S, Cruse, P, Kunze, J, Minor, D, Smorul, M. 2011."Neighborhood Watch" for Repository Quality Assurance. Presented at Designing Storage Architectures for Preservation, Washington, DC: http://1.usa.gov/uXj2Mf



American Educational
Research Association

Response to Request for Information (RFI): "Public Access to Digital Data Resulting from Federally Funded Scientific Research," Office of Science and Technology Policy (OSTP)

76 Federal Register 218, pp. 70176-70178, November 10, 2011

American Educational Research Association
Felice J. Levine, Executive Director (flevine@aera.net)

January 12, 2012

About AERA

The American Educational Research Association (AERA) is the major national scientific association of 25,000 members dedicated to advancing knowledge about education, encouraging scholarly inquiry related to education, and promoting the use of research to serve the public good. Founded in 1916, AERA as a scientific and scholarly society has long been committed to knowledge dissemination, building cumulative knowledge, and promoting data access and data sharing.

For more than 20 years, AERA under its Grants Program has fostered the use of federally supported data sets, especially those of the U.S. Department of Education's National Center for Education Statistics (NCES) and the National Science Foundation (NSF). This long-term project has led to important scientific discoveries and methodological advances and has contributed to a culture of building scientific knowledge cumulatively through analyses of such data. In 2009, with continued support from the National Science Foundation, AERA expanded its efforts and is now working with principal investigators of NSF-funded research on sharing and archiving data from completed studies on education and learning. In collaboration with the Inter-University Consortium for Political and Social Research (ICPSR), AERA is providing support and technical assistance in data archiving to projects with potential for multi-investigator use and will be holding a small grants competition to stimulate use of these data. Through this initiative, AERA is actively engaged as a leader and partner with a federal agency (NSF), the world's largest archive of social science data (ICPSR), NSF research investigators, and potential scientific users on a model project directed to nurturing and promoting the advantages of data sharing and respectful, responsible use.

Because of its special interest in data sharing, AERA has been at the forefront among research organizations in promoting and exploring ways to promote data sharing. AERA engenders this culture within its own policies, procedures, and practices. The revised AERA *Code of Ethics* adopted in February 2011 mandates data sharing and appropriate acknowledgement of data use and takes account of the potential for data use under restricted access provisions that may be necessary to protect privacy rights and the confidentiality of information. (See Appendix A to this response.) Authors in AERA journals and other publications are guided to cite data in their reference lists so as to acknowledge data as contributions in their own right. And, in AERA's 2008 NSF-funded study of education research doctorate programs in U.S. universities (being undertaking in collaboration with the National Academy of Education), a data archiving and data management plan were an integral part of the submission.

In word and deed, AERA strongly supports the goals that led to this RFI and is pleased to share its perspective on the questions posed by OSTP. We shall do so by responding in question order.

Response to RFI Questions

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Federal policy must promote not just the production, sharing, and preservation of digital data, but also the sharing of data collection instruments, the production of transparent and comprehensible metadata, the health of data repositories and similar institutions, and the development of software that will foster or enable data sharing. In addition, federal policy on digital data sharing should not be limited to federally fund "scientific research" but should extend to federally funded "scientific data" of all sorts, with "scientific data" being broadly defined. This is particularly important in the area of education where administrative and other operational data are often routinely collected and can be of great value in advancing learning and education science. We suggest that the following policies would contribute to the RFI's data preservation and access goals:

- There should be a strong presumption that scientific data collected with federal support along with relevant instruments and related metadata will be preserved as specified in a data management plan and made accessible to others. Strong consideration should be given to data archiving requirements as the most effective and efficient way of promoting data use, ensuring data preservation, and ultimately creating a culture of inquiry that values and acknowledges data products and their use. While a presumption of data sharing should not be absolute, any limitation should be very narrowly defined and carefully scrutinized in advance as part of an agreed-upon data management plan, and every effort should be given to the

feasibility of data sharing under restricted conditions or after passage of time even when there may appear to be substantial privacy and confidentiality concerns.

- A data management plan should be a part of federal grants and contracts that fund data collection, and evaluation of that plan should be part of the review process. The cost of data management should be regarded as an essential cost of the research and be evaluated for adequacy and reasonableness along with other proposal costs. Special justification and guarantees of future accessibility should be required if the data management plan does not include data archiving requirements in a publicly accessible research data repository.

- Data repositories, like the ICPSR, that archive and disseminate data from multiple sources greatly facilitate data preservation and sharing. As data sets have become more numerous, larger and more complex, such repositories are likely to prove necessary for any data sharing system to work. Federal policy should foster the development of data repositories, working to improve and sustain them in fields where they now exist and to create repositories for fields that currently lack them. Federal funding should be available to these ends as institutional start-up assistance and as grants or contracts to support innovations in data acquisition and dissemination technologies and procedures, including research on issues that affect the data sharing enterprise such as protecting subject privacy and ensuring that data uses are in accord with informed consent. In addition, support of ongoing operations is desirable, perhaps as a function of the amount of archived federally funded data and its usage rates and/or as add ons to grants to be used to pay fees for data archiving and dissemination services. Archiving data with approved repositories can be further encouraged by providing that depositing data is sufficient to meet a data provider's responsibility for ensuring that subject privacy, confidentiality, and informed consent interests will be sufficiently protected as the data are stored and disseminated.

- Data management should be recognized as a scientific profession in fields where it is not now adequately recognized. Federal policy can support this by supporting data manager education and research in ways similar to the support that it provides students pursuing education and careers in other science fields. Federal policy and funds could also support meetings of data repository managers and others involved in data archiving and dissemination to ensure that their treatment of data and metadata is mutually compatible and to maximize the feasibility of working with data drawn from different repositories.

- The demands and challenges of data management, including archiving and dissemination, change regularly as new technologies develop and new policies, like revised privacy rules, are put in place. Federal agencies should be encouraged to contribute funds for periodic National Academy of Science studies such as the National Research Council's 2005 report, *Expanding Access to Research Data:*

Reconciling Risks and Opportunities, or the 2009 report on *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. Also, investments in initiatives like the NSF-funded AERA/ICPSR project that enables investigators and their teams to implement plans for data archiving and use or address challenging issues can have high payoff and long-term impact for relatively modest cost. State of the art and consensus conferences centering on issues of standards for data and metadata and consistent forms of data citation are also important short-term priorities and could usefully inform further federal policy on data sharing, access, and preservation.

- Where data are collected by a federal agency the data should be available for sharing to the widest extent possible as determined by federal law and administrative rules. Some agencies, in particular some federal statistical agencies like NCES, have been leaders in this effort for quite some time, but this issue is worthy of consideration federal-wide. In some instances, it may be appropriate for federal agencies to form their own plans and systems for access to digital data, but the existence of such plans and systems should not preclude making all or portions of federally collected data available in data repositories even if they can be also acquired from the government. Indeed, multiple systems of availability should be encouraged since this broadens access and different providers may create tools that make working with the data easier and more cost effective.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

First, the intellectual property interests of all involved should be specifically defined. The definition should make clear that, if scientific data have been collected with federal funds (unless the conditions of funding provided specifically to the contrary), those data and related metadata are in the public domain. That said, data collected and prepared by researchers are a product that merits and deserves appropriate credit by those engaged in their use. Citation of such data with appropriate attribution should both facilitate tracing advancements enabled by data resources and also offer a vehicle for giving credit and measuring interest and use.

Scientists understandably want to have the opportunity to analyze the data and make contributions to knowledge that follow from their conceiving of the project and the data collection effort. Such use is an appropriate incentive and reward for engaging in data collection. The time period for exclusive use should be specified as part of a data management plan and be approved or modified by the federal funder. Depositing the data in an archive for dissemination need not and ordinarily should not await the end of the exclusive use period so long as the dissemination of the data to others is embargoed until the exclusive use period has expired. Creators and licensees of works that use the

data, such as reports or articles analyzing the data, should have the usual intellectual property protections for the products of their use unless the conditions of funding the research or licensing the work provide otherwise.

Problems may exist when collected data are, like some business data, proprietary or the intellectual property of the person or entity supplying the data. When this situation exists, the rights of the data provider must be recognized and protected by the data collector even if the collector has no property rights in the data. In such circumstances the person or entity collecting the data should attempt to draft an agreement that calls for the widest data sharing arrangements that the data provider will agree to, and should propose confidentiality, data source masking, or other arrangements to facilitate sharing. In making funding decisions, federal funders determining the likely value of the proposed research should take into account the likelihood that the intellectual property or proprietary rights of data sources will limit reuse of the data.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

The basic scientific principles and value of data sharing may vary less by discipline and field than it might seem, although in practice the levels of experience and exposure to data sharing and archiving vary. Thus, much of the challenge may be less in the policies derived than in the federal plan for investment in implementation and education so that sharing approaches are specified consonant with the different forms of data and methodologies for data acquisition. As social/behavioral scientists, education researchers collect and use the full spectrum of data at the individual, social, and institutional levels employing systematic and rigorous methods for data acquisition and analysis. Increasingly in fields like ours, there is growing use and attention to physiological and biological information, and we can anticipate both greater use of biomarker data in the social/behavioral sciences and greater use of social/behavioral measurements in the biomedical and biological sciences. Since many of these data collections are large-scale and longitudinal, the discussion of data sharing and access has already commenced (see, for example, the 2010 NRC report on *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biadata*).

When it comes to establishing policies regarding data sets that integrate various forms of human and organizational data, federal agencies have considerable social and behavioral science resources to draw on. In setting data sharing and management policies, they should be encouraged to utilize and build upon this expertise end experience. First, there is the ICPSR, now 50 years old and a pioneer in gathering and facilitating the further use of data that has informed original research across social science fields. ICPSR has had to confront such issues as obtaining usable data, data security, subject privacy protections, allowable use, legitimate access limitations, and requirements for metadata among other matters. They have considerable wisdom to

share and their experience will be highly instructive. Second, there are in the social and behavioral sciences a number of important longitudinal surveys whose purpose is to provide data to broad user communities, often giving no user priority to the data collectors (e.g., the Panel Study of Income Dynamics, the National Longitudinal Study of Adolescent Health, or the Health and Retirement Survey). Their experience, particularly in providing data in transparent user friendly forms can also inform federal policies and requirements. Third, there are professional associations, like AERA, that have given considerable attention to issues of data sharing as a professional responsibility. Their consideration of these issues can inform federal agencies of the data sharing behavior that is coming to be regarded as normative within our fields and can provide a professionally acceptable starting point for establishing data sharing policies. Fourth, there are federal agencies like the National Science Foundation (NSF) and the National Institutes of Health (NIH) which have established data sharing policies that can serve as models for other agencies. The obligation to share data collected with NSF funds and the commitment to a data management plan, although only recently reaffirmed and formalized as an agency-wide requirement, was initiated as far back as 1987 in the then NSF Social and Economic Science Division and 1989 in an NSF policy statement on data access and data sharing. Finally, there is an extensive body of knowledge through books, articles and reports, like the NRC Reports mentioned above, that deal with issues relating to justifications for, problems posed by, and the mechanics of social and behavioral science data sharing. This learning should be consulted.

We do not know whether similar resources exist to aid agencies in establishing data sharing policies for other kinds of data, but to the extent they do exist we believe federal agencies should take advantage of them. We also believe that federal agencies should eschew "one size fits all" policies when establishing rules and recommendations for data sharing. Instead, the characteristics of the kinds of data collected and used in different sciences should be examined, and policies should reflect the experiences and knowledge specific to different disciplines and fields.

4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

We are not clear on whether this question is concerned with the mechanisms for considering the cost-benefit issues it raises or with the standards and policies that should be applied. However, in either case consulting the sources of aid and information identified in our response to Question 3 is advisable. We would add that the cost-benefit issues raised by this question are important. The fact that a data set is created with federal funds does not necessarily mean that it is worth preserving and sharing over the longer run. At a minimum, however, data should be preserved and made available to others until a reasonable period of time after publications and reports drawing on the data have appeared. This allows for verification of results, examination of alternative hypotheses and questions, and consideration of these data to address

other issues or problems. Beyond the overall value of data preservation, judgments regarding data retention and dissemination should be based on the quality of the data, the range of issues that the data address or could address if combined with other data, the importance of the issues the data address, and foreseeable future uses of the data. In addition, when data have been available for some time in a publicly available repository, the demand for the data should be taken into consideration in any data culling decisions. We suggest that the federal agencies that fund data collection will most often be in a poor position to judge these issues, and that these determinations should be delegated to scientific advisory committees on data acquisition and retention attached to different approved data repositories and similar organizations. In addition, we suggest that no data should be removed from an archive and/or made unavailable to researchers if an agency wants the data retained and is willing to pay the costs of retention. We believe that any group determining what data should be kept should begin its evaluation process with a presumption that favors data retention. And even data that appears of little current value, as might be the case if it has been exhaustively analyzed, should be retained if there appears to be a real possibility that future access to the data will prove scientifically important.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Representatives from stakeholders like these can serve as participants in conferences and on advisory bodies when issues regarding data retention and dissemination policies and practices are being raised. It should, however, be recognized that in many instances the views of those who represent such organizations will be the views of consumers of data services and not the views of data stewards or management experts. In addition, research communities, universities, and research institutions can contribute to the implementation of data management plans by making clear to their members and/or employees that such plans should be a routine part of any grant application and that there is value in working with data repositories in providing data for further use. In particular, professional associations, research institutions, and universities should support the ethical standards of data sharing and responsible use consonant with human subjects research protections. Publishers may contribute by cooperating with journal editors to establish policies for citing data by a persistent digital identifier, such as a digital object identifier (DOI), thus encouraging data management plans that call for the prompt contribution of data to repositories. Finally, universities and research institutions could do more to recognize the scientific status and role of data managers.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Costs of preserving digital data and making the data accessible should be an accepted or even a required element of any grant or contract seeking federal funds which includes a

data management plan. Federal support might also be made available to recognized data repositories, perhaps in relation to the amount of federally funded data stored and demands for the data's use. In addition, special grant programs might be developed to support research aimed at innovations in data management technologies and practices as well as seed money or ongoing support for a consortium of data repositories and for technological linkages and periodic conferences that facilitate communications among different data managers and data management organizations. Finally, support in the form of educational grants and internships would indirectly reduce the costs of preserving data and making them accessible by expanding the pool of those with careers dedicated to these ends.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Agencies could require that final reports on grants and contracts demonstrate compliance with federal data stewardship and access policies, and could provide that compliance is demonstrated by providing a persistent digital identifier showing that the data together with relevant metadata have been deposited in a recognized data repository. Problems might be expected, however, because final reports in many instances are likely to be due before it would be expected to deposit data. In such cases, final reports could be accepted conditional on the filing of an addendum demonstrating data policy compliance by a set date.

It might also be appropriate to recognize data repositories available and appropriate for data archiving and preservation. One possibility would be establishing broad federal-wide standards for recognized repositories coupled with a requirement that any repository that wished to be so recognized file a statement with the agency that the standards are met. We suggest that the determination of standards of adequacy and revisions in them be made in the first instance by a group of those organizations that can today be recognized as well-functioning data repositories, with the federal role limited to endorsing or rejecting the standards. As new repositories are recognized, the group setting standards for recognition could expand accordingly.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Anything that increases the availability of data and ease of access and use can be expected to stimulate the use of publicly accessible research data in new and existing markets. To this end federal agencies could (a) support the development of clear and consistent standards for metadata and require those data collectors it funds to meet these standards as part of their data management plans; (b) support for certain important and complex data sets the development of software tailored to the data set

to make it easier for users, including the relatively unsophisticated, to find relations of interest to them, and (c) create or support the creation of a partially catalogued and searchable library of all data preserved in data repositories that have associated persistent digital identifiers.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Some mechanisms, already discussed, for example, in our response to Question 5 above, must be left to the private sector, including universities and professional associations, but federal agencies could call relevant issues to the attention of non-federal organizations and encourage them to consider steps that promote appropriate attribution and credit. Federal agencies could, in addition, set a valuable example by attaching citation information and persistent digital identifiers to data they create and make available and by requiring those who submit grant or contract proposals or who report on their scientific activities to cite sources of data referenced in them and provide persistent digital identifiers if available.

Questions 10-12 require technical knowledge that others are better positioned to provide than we are.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

We have discussed many of the policies, practices, and standards that are needed to support this linkage in our responses to the questions posed above. To summarize, policies and incentives are needed so that data sources will be cited and, when cited, will be accessible to those interested in examining the data more closely. This means that data must be available, preferably in publicly accessible data repositories, and that the data must have associated with it citation information and a persistent digital identifier that encompasses relevant metadata along with the data. The federal government can lead both in its own data management practices and in the standards it sets for its grant and contract seekers. In addition, through the sponsorship of conferences or other means, the federal government can encourage professional associations, publishers, universities and research institutions to adopt as a matter of policy or professional ethics those standards that will promote the linking of publications and associated data.

Appendix A

Excerpt on Data Sharing
Code of Ethics
American Educational Research Association
(Code of Ethics Adopted by AERA Council, January 2011;

14.06 Data Sharing

(a) Education researchers share data and pertinent documentation as a regular practice. Education researchers make their data available after completion of the project or its major publications for verification or other analyses by other researchers, except where proprietary agreements with employers, contractors, or clients preclude such accessibility or when it is impossible to share data in any useful form.

(b) In sharing data, education researchers take appropriate steps to protect the confidentiality of the data and the identity of research participants. When appropriate future use necessitates access to identifiable data, researchers take steps to ensure that the data are accessible under appropriate restrictions where the confidentiality of research participants can be secured. See also 12.04(b) and 12.08(c).

(c) Education researchers anticipate data sharing as an integral part of a research plan whenever data sharing is feasible.

(d) Education researchers share data in a form that is consonant with research participants' interests and protect the confidentiality of the information they have been given. They maintain the confidentiality of data, whether legally required or not; remove personal identifiers before data are shared; and, if necessary, use other disclosure-avoidance techniques. When data are shared with personally-identifiable information, education researchers take steps to ensure that access is provided only under restricted conditions where users agree to protect the confidentiality of the data consonant with prior commitments.

(e) Education researchers who do not otherwise place data in public archives keep data available and retain documentation relating to the research for a reasonable period of time after publication or dissemination of results and share data consonant with 14.06(a).

(f) Education researchers who use data from others for further analyses explicitly acknowledge the contribution of the initial researchers.

Thu 1/12/2012 11:50 PM
Response to RFI

Michael Carroll; mcarroll@wcl.american.edu

Professor of Law and Executive Director; Program on Information Justice & Intellectual Property, American University, Washington College of Law

Meredith Jacob; mjacob@wcl.american.edu

Assistant Director; Program on Information Justice & Intellectual Property, American University, Washington College of Law

We appreciate the opportunity to submit the following comments to the National Science and Technology Council's Interagency Working Group on Digital Data in connection with the Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research published in the Federal Register on November 4, 2011.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The question implies that changes to federal data sharing policies would be needed to improve access and reuse of data produced or collected from federally-supported scientific research. This implication is correct. As a general matter, existing federal data policy is uncoordinated, underspecified, and, frankly, incoherent. Notwithstanding the laudable goals articulated in the America Competes Act and in OMB Circular A-130, data produced or collected with federal support is subject to a range of possible rules regarding public access and terms of reuse.

We conducted a review of publicly accessible policies from agencies supporting scientific research to ascertain what, if any, data sharing requirements recipients of federal funds agreed to with respect to non-classified research.

1. What mandates are imposed by federal law through statute, regulation or policy on recipients of federal funds to provide public access to scientific data generated by federally funded research? (Both intramural research and grant or contract funded research will be examined.)

2. In the absence of a federal mandate for data sharing, what efforts do agencies take to promote or provide public access to data produced or collected through federally funded research?

3. What, if any, restrictions or requirements do agencies place on recipients of federal funds who make their research data public to use technological protections or contractual terms of use that limit reuse, reanalysis or redistribution of such data?

In sum, our results show that there are very few federal mandates requiring data-sharing, that agencies by and large have adopted an ad hoc approach to promoting data-sharing by federally-funded researchers, and that no policy that we could find recognized or addressed the common practice among federally-funded researchers to impose terms of use on data made public over the Internet without any federal input into the presence or substance of these terms of use.

We recommend that federal data sharing policy should be consistent across all agencies so that data availability is useful and predictable. This policy needs to be clearly set out, easily accessible online, and consistently enforced.

Federal funding for scientific research comes through a number of routes: through the direct employment of researchers; through grants to researchers at not-for-profit institutions; and through partnership and contract agreements with for-profit corporations. Whenever possible, the data that results from this federally funded research should be made available to the scientific community and to the public.

Ideally, data produced by employees, grantees, or contractors could be made available online to the public in a searchable, standardized form without any artificial barriers or limitations on reuse. Unfortunately, many Federal agencies only meet a minimum standard, complying with OMB Circular A-110 that requires that any data that were used for rulemaking be made available in response to a FOIA request. Additionally, A-110 gives the Federal Government right to "(1)obtain, reproduce, publish, or otherwise use the data first produced under an award; and (2) authorize others to receive, reproduce, publish, or otherwise use such data for Federal purposes."

Notice that this policy does not prevent the recipient of federal funds from imposing reuse limitations on other users via contract.

In practice, two agencies, the National Institutes of Health and the National Science Foundation impose the broadest requirements on grantee data-sharing. In other agencies, specific projects or institutes will impose data-sharing requirements on grantees, such as the Genomics:GTL project within the Department of Energy. Finally, many agencies, such as the Environmental Protection Agency and NASA have invested significant resources into making data publicly available, but neither has a policy of mandatory access to data generated by grantees. Data policy and location should be standardized across agencies so that researchers, policymakers, and lay people can local data and rely on its continued availability.

With rare exception, federal data policy defers to investigator preference as to whether data will be shared with the public. We recommend that this policy be revised because investigators may have a conflict of interest with the public interest and engage in competitive withholding for personal gain or may undervalue the reuse potential of research data by researchers in other disciplines.

While agencies may need to establish limited criteria to opt-out of data sharing, these criteria should be specific and should not simply rely on researcher choice. Federal policy on access to digital data should not use the NSF model that relies solely on the discretion of the researcher to determine whether data is made publicly accessible.

(2) **What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?**

Some care should be given to delineate the "intellectual property interests" referenced in the question. Copyright does not apply to factual data that are arranged in an unoriginal manner. Most data themselves are not patentable inventions. Data can be treated as a trade secret, but there is circularity in this determination. Information can only be a trade secret if it is not "readily ascertainable", and it is a matter of federal policy as to whether data should be made readily ascertainable. As a consequence, there is only an "intellectual property interest" to be protected if federal policy is that there should be such a private interest rather than a right of public access.

The issue that needs to be addressed is to what extent researchers may hobble or encumber access to data arising from federal funds through contractual restrictions or technological protection measures. Neither of these is an "intellectual property interest," but each can be an effective means for undermining the public's interest in data access and data sharing.

However, even when researchers use private databases, policies can protect public access to the results. In one example from the National Center for Environmental Economics:

> **(d) Data Plan (if applicable).** Provide a Data Plan (2 single spaced page limit) to make available to the public all data generated from observations, analyses, or model development (primary data) collected under an agreement awarded as a result of this RFP. The plan should describe how the applicant plans to make all data resulting from an agreement under this RFP available in a format and with documentation/metadata such that they may be used by others in the scientific community. This includes both primary and secondary or existing data, i.e., from observations, analyses, or model development collected or used under the agreement. *Applicants who plan to develop or enhance databases containing proprietary or restricted information must provide, within the two pages, a strategy to make the data widely available, while protecting privacy or property rights.* (emphasis added) National Center for Environmental Economics, Grant Solicitations *available at* http://yosemite.epa.gov/ee/epa/eed.nsf/pages/GrantSolicitations.html#bmk50

This example illustrates options that could be included in other guidelines to acknowledge the interaction between public and private databases, and the need to allow contribution to the private database while protecting public access.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Federal agencies can take into account inherent differences in digital data across disciplines by allowing data deposit in discipline-specific repositories while also requiring metadata necessary for indexing and search to be submitted and maintained in a central database that makes it possible to local all digital data resulting from federally funded data from a single central search.

Discipline-specific repositories can thrive with the support of a dedicated research community, but we should not lose the value in making that data also locatable and useable by the larger scientific community and the public.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

The presumption should always be in favor of long-term stewardship and dissemination because the future utilization cannot be anticipated. Federal data access and preservation policies should protect the reuse of digital data for both researchers outside the original field and for future researchers even when the potential for reuse is not obvious within the field. Digital data is a resource which can be an input into a range of innovative activities, and it would be unwise to assume that we can predict the value of data as technological capacities evolve and as public access to research increases.

Discipline -specific repositories lack a central directory or access point for lay members of the public, or for researchers outside the field. Data.gov. or another central portal should provide a central search to locate data sets, even if they are deposited in discrete repositories. This would allow specialized repositories if necessary to adapt to the needs of a specific scientific community, while still insuring broad public access for novel or crosscutting research.

Currently, though large amounts of digital data are available online, there is no system for determining either data sharing policies, or data repository location, across agencies. Even within agencies, such as discussed below for the Department of Health and Human Services, data sets are scattered across agency websites, without a central index

If the public funds the cost of data collection and storage, then it is imperative that we have an efficient central index for locating these data sets across repositories so that they get the most possible use. Clear policy and a functional central search index helps the research community and the public get the highest possible value for the effort of data collection and preservation.

Finally, long term data preservation allows for the measurement of change over time, even when that was not the initial intent of the data collection. One example of this is Library of Congress-led project on the preservation of historical geospatial data that was initially intended for mapping and geologic studies, but can be used for other environmental, economic, and social research when preserved over a longer period.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Research communities can contribute to standards and best practices that allow collection, standardization, and deposit of high quality data.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The preservation, deposit, and hosting of digital data should be addressed throughout the research funding process. Funds should be allocated in proposals for data collection and management. Clear criteria should be given to reviewers for the evaluation of funding proposals. Finally, completion of data deposit with a public repository should be an enforceable requirement of Federal grants.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The burdens of compliance are actually reduced as the policy of requiring open access to digital data is standardized.

A key step to encourage compliance with Federal data stewardship and access policies would be to build in a focus on data access throughout the grant process. For the preservation and deposit of digital data to thrive, it must be seen as a core deliverable of a grant or contract. This focus on data preservation and data sharing should begin at the grant review phase, where clear guidelines should be given on how to evaluate data management plans. It should be followed by an approach, such as the one currently in place at National Institutes of Health (NIH), that views failure to implement the data management plan as grounds for enforcement, and as a barrier to future grants.

While agencies may need to established limited criteria to opt-out of data sharing, those criteria should be specific and should not simply rely on researcher choice, and the interests of researchers and the public are not completely aligned. Researcher may not be able to see the applicability of data sharing to fields other than their own and also may have self interest in delaying the publication by competitors in the field.

If this became a standard part of scientific research, systems could be developed to reduce inefficiency and automate the deposit of data in open repositories. The more data deposit is standardized and automatic, the lower the cost of enforcement and verification.

(**8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy**?

Data made available through data sharing should not contain any contractual preclusion on reuse. Repositories used for public access should not contain any terms or conditions that limit the free reuse of data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The Federal government should support initiatives to develop standards for data citation and data attribution.

Summary

The following principles should guide Federal data policy:

- Federal data sharing policy should be consistent across all executive branch agencies. Consistency across agencies is valuable for researchers and the public so that data availability is useful and predictable. Specific policies can be implemented at the agency or project if need be.
- Federal data policy and any agency or project level modifications should be clearly available online and specifically set out the location of data indices and repositories.
- A central index of all data or data repositories should be established, i.e. data.gov
- Federal data policy should require data sharing as the default for all federally funded research. While agencies can establish criteria to opt out of data sharing, these criteria should be publicly available as part of the data sharing policy. Researcher election alone should be insufficient criteria to opt out of data sharing.
- Data sharing guidelines should be built into the research grant process from proposal evaluation to completion of the grant. Data sharing should be seen as a enforceable requirement of the grant.
- Federal data sharing policy should recognize the value of research outside the original field, as well as the high potential for future, unanticipated use of data.

- Federal contracts should require the same data sharing policies as grants to non-profit institutions, unless they fall within criteria established by the contracting agency.
- Data made available through data sharing should not contain any contractual preclusion on reuse.

Best regards,

Michael W. Carroll
Professor of Law and Director,
Program on Information Justice and Intellectual Property
American University, Washington College of Law
4801 Massachusetts Ave., N.W.
Washington, D.C. 20016
vcard: http://www.wcl.american.edu/faculty/mcarroll/vcard.vcf

Research papers: http://works.bepress.com/michael_carroll/
http://ssrn.com/author=330326
blog: http://www.carrollogos.org/
See also www.creativecommons.org

January 12, 2012

Brian Westra, Science Data Services Librarian, University of Oregon, bwestra@uoregon.edu
Eugene, OR

Comments in response to the Office of Science and Technology Policy Request For Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research
Federal Register Doc No. 2011-28621
submitted electronically to: digitaldata@ostp.gov

These comments include text and distill discussions and ideas from a number of data curation scientists and librarians at other institutions and the University of Oregon. The response also reflects the opinions of the compiler, Science Data Services Librarian at the University of Oregon, Brian Westra.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

A clear, government-wide policy supporting open data should be the foundation on which other policies are based. At a minimum, it should require that research data be made available via open access repositories and data centers. Allowing data to become a restricted access commodity held by commercial entities will only serve to enrich those commercial data providers, while impeding scientific research and economic development in all sectors. “Open data access with appropriate ethical restrictions can be viewed as a new core principle for developing a global data infrastructure” (Parsons, 2011). The requirement by the NSF to include data management plans in all proposals has been a good step in encouraging researchers to actively plan for and engage in practices that will enable sharing of research data. More specific guidance and requirements for sharing data will help. One example would be to directly tie data stewardship and open access to data, to future research funding.

Of course, data sharing requirements can only be met if there are sufficiently robust and diverse resources to accommodate the variety of data sets collected and generated by researchers across scientific domains. Funders should support the infrastructure for research data, from data centers and repositories for the data, to the metadata standards, and discovery and access tools necessary to find and use those data sets. Policies that minimize barriers to sharing should also attempt to address other issues, such as insufficient local and national infrastructure. For instance, see the NSB (2005 - http://www.nsf.gov/pubs/2005/nsb0540/) report on long-lived digital data: "Participants

agreed to a considerable extent on the main policy issues, even though there is one stark difference between NSF and many other agencies: the vast majority of long-lived data collections supported by the NSF are managed by external research organizations, while other agencies, such as the National Aeronautics and Space Administration (NASA) and the National Oceanographic and Atmospheric Administration (NOAA) focus more heavily on archiving and curating many such data collections themselves."

It may not be necessary for the NSF to hold the data sets themselves, but it is important that stable, long-term funding be made available to maintain data centers and repositories providing open access. Insufficient support for domain-based repositories forces researchers to adopt ad hoc approaches that are not optimal or efficient, and degrade over time. In addition, existing data centers, particularly those that receive federal funds, should be encouraged to accept and curate a broader range of data in their disciplines.

Funding in many cases has been relegated to the traditional term-limited model, which is not compatible with long-term preservation and access to research data. Data preservation and access efforts have been hampered by this instability, and data curated by term-funding in some cases have been put at risk when the repository funding is cut. Secure and stable funding of infrastructure and expertise not only facilitates access, but enables research and development of data tools and services which in themselves can lead to ground-breaking research and economic development. Open access provides new opportunities for commercial development with not only your own intellectual property but also that of others. It opens up opportunities for everyone.

Awareness of data preservation and access problems, benefits and strategies, remains limited amongst many communities. We recommend the addition of basic data management skills into the scientific curriculum in parallel with a campaign to promote a cultural shift in the sciences towards expanding the core principles of reproducibility, transparency of methods and evidence-based assertion.

One of the biggest obstacles to data reuse and preservation is the capture and standardization of metadata. Metadata is at the core of many data reuse issues, from discovery, to trust, and data quality. To realize the goals of growing the economy and improving productivity of the scientific enterprise, we must comprehensively approach metadata through not only technical systems, but also, critically, through social mechanisms.

Lastly, the data science field must continue to be developed, with the recognition that many of the questions in the RFI are the subject of ongoing research.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

We feel it is important that intellectual property considerations be weighted to reflect the contribution of the researcher and the investment of public funds into the research endeavor. Data is fundamentally different from peer-reviewed publications in regards to copyright and intellectual issues. In some cases, embargoes on the public release of data may provide a sufficient accommodation to the needs of the original researchers. It is counter-productive to the research process to grant intellectual property rights over research data to publishers, since that approach discourages free access to the data.

Other steps that can be taken include the development and implementation of strong metadata standards, particularly as they relate to the provenance (history or chain of custody) of the data. Others have pointed out that the Creative Commons CC-BY license, with modification, represents a good foundation for a license for data that facilitates the development of services to maximize the utility of data.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Policy and guidelines should be created with the full input of practitioners in the research field. In addition, data scientists/librarians and curators should also be an integral part of a collaborative effort to generate transparent and workable guidance that can be realistically implemented throughout the data life cycle. Supporting the work of data scientists is important, as they provide a translational layer that brings the data from an isolated study into the context of related data, ready for future work. The services offered by various data centers and communities, lessens the impact of the inherent differences between disciplines, and between the researcher and the data center or repository's curation practices.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Domain-based data centers with an ongoing mandate could help to ensure long-term access to data. Although the long-term value of a particular digital data set may not be known until well after its creation, some metrics, such as the cost of recreating a data set, might be considered as factors. Scientific research at this moment in many fields generates more data than can reasonably be expected to be preserved, and that pace is accelerating, so the research community will need to help establish processes to identify and "promote" data that are recognized as worthy of preservation. Preservation and archiving services and platforms are also evolving, so archivists, data librarians and curators should also be consulted in decisions about assigning costs and values to the facets of data stewardship. Data scientists/librarians and curators should be incorporated into the policy framework for data infrastructure. Policies should recognize the range of relative contributions of and

applications for data sets and support strategic delegation of resources by data scientists and curators.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

In some instances, such as the ICPSR for social science research survey data, an infrastructure already exists. The roles of organizing, annotating and cataloging, preserving, and providing access to information are inherently part of the capacity of libraries. This means that many libraries may have the capability to provide consultation and support for the implementation of data management plans, in collaboration and coordination with other stakeholders. Depending on the resource, these services and skills can be applied to working with local, shared, consortial, or remote/external stakeholders in the process of supporting data stewardship throughout the data lifecycle.

For example, academic libraries and data scientists/librarians and archivists can connect researchers with campus services; collaborate with other stakeholders and the researchers on the development and operation of domain-based infrastructure; promote and support the use of existing infrastructure (assisting researchers in identifying appropriate infrastructure and data repositories, advising on best practices for preparing data and metadata for deposit, etc.); participate in the development of standards; and maintain a current awareness of institutional and funders' policies and assist researchers in meeting them.

For the successful implementation of data management plans, many stakeholders with a diversity of roles, responsibilities and expertise must come together under a common goal of ethical data sharing. Recognizing the need for a diversity of perspectives at the table – including those of the scientist and research team, data curator, technologists, metadata experts, digital archivist, data reuse support personnel, and others – is an initial step towards supporting DMP implementation. Roles and responsibilities must be clearly defined and delegated to the stakeholders with the appropriate expertise. All stakeholders should promote policy compliance along with community-based standards-making.

Specifically: Universities and research organizations - Clarify intellectual property statements regarding data and promote data publication and sharing as part of the tenure process considerations. Recognize and support centralized data management and infrastructure systems and services as valuable institutional facilities for researchers. Offer career paths for data managers and data scientists on campus. Include data training as part of the science curriculum. Research communities – Recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement. Proactively contribute to standards-making discussions and the cultural shift towards data sharing and complete metadata capture. Libraries – Consider data as not only part of the scholarly communication cycle, but as a publication/resource (with equal

weight as traditional resources) to be curated. Apply extensive experience in cataloging and bringing together diverse, interdisciplinary resources to the data paradigm. Publishers – Enable and encourage ethical data sharing and attribution through citations within publications and research documentation.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

In many ways the real costs of preserving and making data accessible are just beginning to emerge. The recognition that data preservation and stewardship are not solely technical problems is one step towards identifying and planning for the costs. The cost of data preservation for reuse includes not only the infrastructure and long-term system maintenance fees, but also the expenditure for capturing and structuring metadata, ongoing standardization work and user support. In cases of observation data from the 'small' sciences, the largest cost involved potentially comes when not only making data accessible over the long-term, but making data *useful* well into the future. The amount and quality of metadata required for reuse can potentially dwarf the metadata required for access and preservation. This is the topic of ongoing research.

A secondary issue that is currently not addressed in most research funding mechanisms is that data stewardship, particularly preservation and access (such as format migration) goes on well beyond the life of the grant. There is also a lack of agreement or established guidelines on how research funds can be allocated toward data curation services, since they involve more than direct hardware or technology costs. Clearer, more comprehensive guidelines and provisions in these areas would have a considerable impact on clarifying how the costs can and should be addressed, particularly at the academic institutional level.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Much as publication submission to an open access resource (i.e., PubMed Central) can be used as a metric for compliance, links to citable and discoverable data sets reported in a final grant report could be a verification mechanism for data sharing compliance. If a minimal set of metadata standards are established that can be harvested from the relevant data center(s) or repositories, that information could provide a very basic and low-cost verification measure. Although this does not directly address many facets of good data stewardship, it would provide at least a starting point. The adoption of data registration services and DOIs for open access datasets are a similar component that could be used to verify compliance, and provide persistent links from publications and reports.

As with other components of data stewardship, these services are likely to evolve, and should be seen as a starting point. Tieing future grant support to compliance, and educational provisions such as the Responsible Conduct of Research requirements are also

mechanisms that will improve compliance. As noted in other sections of these comments, federal support for the infrastructure for preservation and access to research data is necessary to reduce the barriers to compliance by motivated and interested researchers and other stakeholders. Ongoing research support in the areas of data curation will also be key to moving the requisite systems forward to match the volume and complexity of the data being produced.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Research awards can be directed toward innovative re-use of data, and toward proposals that support and promote standardization, normalization, and value-added services and tools. These kinds of approaches are key to leveraging the data deluge, and opening new "fourth paradigm" approaches to research. Collaborative proposal support between Federal agencies, such as the IMLS and NSF, could leverage the strengths of a broader community of partners to address these issues.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Several groups are already promulgating data citation standards (such as the Digital Curation Centre in the UK, http://www.dcc.ac.uk/), data set registration (DataCite, http://datacite.org), as well as author identifiers (ORCID,). Incorporating these into guidelines and as standard mechanisms and best practices will not guarantee appropriate attribution, but certainly will help. If federal agencies work with these stakeholders and others to adopt these practices and require them in reports and other documentation, it will lend credibility to these efforts.

Standards for Interoperability, Reuse and Repurposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Because of the heterogeneous nature of research data across and even within research disiciplines it is unrealistic to expect to employ a single standard for all research data. Community-driven efforts are necessary and important within a domain context, while broader, more general standards are necessary to promote interoperability nationally and internationally. A combination of standards, refined for each data type, is one option for enabling reuse and repurposing. The development of these standards and involvement of the full range of stakeholders and expertise should be encouraged and supported. One

approach would be to spread support across ISO-level standards development while recognizing the role of community-driven standards creation. For top-down, high-level standards to effectively intersect with bottom-up, detail-oriented standards and best practices there will need to be a decentralized model of support and communication with a defined path towards formal integration.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

There are a number of organizations and agencies in other countries that may be further along on the path to coordinated data policies. Within the United Kingdom, the Digital Curation Centre (http://www.dcc.ac.uk/) regularly provides and seeks out collaborations with U.S.-based organizations such as the Coalition for Networked Information (http://www.cni.org/). Some other likely partners are the Australian National Data Service (http://www.ands.org.au/), ISO, Dublin Core metadata iniative (http://dublincore.org/), and groups within the European Commission, for example.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Federal Agencies should support the work of the International DOI Foundation (IDF) and of DataCite and other organizations with the goal of making an unified international standard and support structure linking between publications and associated data. Access and use of this mechanism/standard should be easy for the data generator. There may be a role for a Federal Agency to act as a mediator or an issuer of DOIs.

References and Resources:

Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. GRDI2020.

Response to Request for Information: "Public Access to Digital Data Resulting from Federally Funded Research," November 2011
January 12, 2012

Wendy Pradt Lougee
University Librarian
McKnight Presidential Professor
University of Minnesota Libraries

Thank you for the opportunity to comment on "Public Access to Digital Data Resulting from Federally Funded Scientific Research." These comments are submitted on behalf of the University of Minnesota Libraries. The University of Minnesota is one of the leading public research institutions in the United States, and a key contributor to the entrepreneurial economy of the state of Minnesota, as well as to scholarship both nationally and internationally. We strongly advocate for a policy that ensures public access and long-term preservation ("stewardship") to digital data resulting from federally funded scientific research ("data"). We believe that such a policy would provide immeasurable public benefits far outweighing any costs or burdens such a policy might impose.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Comment 1:
We recommend a Federal policy that would mandate data deposit into publicly accessible repositories as quickly after publication or shortly after the grant-funding period as possible, recognizing limitations that may be imposed due to confidentiality or other legally protected data. This policy step goes beyond the data sharing requirements outlined in the OMB Circular A-110[1], which many researchers interpret as data sharing only by request, would prevent the loss of potentially valuable data, and remove barriers of access due to variability in data managed in local or individual environments.

Data stewardship policies that encourage public access can also position data for re-use through curation and management techniques that ensure long-term access. Characteristics of a sound Federal data deposit policy might include:

- a requirement for a data management plan with all funding proposals that describes how the data will be deposited for public access within appropriate federal, disciplinary or institutional data repositories.

[1] The Office of Management and Budget (OMB) Circular A-110 was revised in 1999 to provide public access under some circumstances to research data through the Freedom of Information Act (FOIA). http://www.whitehouse.gov/omb/circulars_a110

- a post-award review process and merit considerations for future funding based on a successful history of data deposit. The NSF's Computer & Information Sciences and Engineering (CISE) directorate[2] provides a good example.
- recognition that not all data can be open due to security or confidentiality interests, and that some categories of data will need to be appropriately prepared before release or qualify for an exemption. When open access to data must be delayed, deposit with access restrictions could still be required in order to ensure preservation and long-term access.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2:
Intellectual property issues associated with data differ significantly from those associated with publications. Established federal law does not recognize an ownership interest in raw data, reflecting an understanding that data may also be most productive and fruitful throughout the economy when access and use are available to many parties. Intellectual property rights may exist in certain types of data, or compilations of data, but these are well addressed by current law. Federal policies should:

- value and reward data stewardship for re-use -- e.g., by incorporating data stewardship as a consideration at reviews and in applications for future grant awards.
- prohibit constraints on data due to overly tight coupling with related publications, or by publisher-imposed limitations on data access. Requiring open licensing for the data that is covered by intellectual property laws will foster productive reuse while allowing credit to be given to those who did the work.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3:
Recognition of the inherent differences between disciplines is critical, particularly with respect to decisions on archiving, shareability, and discovery. The NSF DMP requirement offers a good model by setting a minimum policy, and yet allowing each directorate to build from the base. The Federal agencies could take this model one step further by outlining a set of data stewardship principles and requiring that each agency provide base-line principles for data management, preservation, and sharing of data in their respective disciplines, such as establishing a disciplinary metadata standard, a shared data repository, appropriate maximum embargo periods, etc. Researchers could then model their data stewardship on these standard requirements.

[2] The NSF CISE Directorate Guidance for CISE Proposals and Awards requires that Data sharing progress and outcomes must be reported in award annual reports, the final report and subsequent proposals by the PI and Co-PIs. Last updated September 15, 2011 at http://www.nsf.gov/cise/cise_dmp.jsp.

Since disciplinary repositories are developed to take into account the unique structures and characteristics of the target data, Federal agencies should support protocols to enable cross-discipline and cross-repository discovery. The University of Minnesota's NSF-funded TerraPop project, a joint undertaking of the Minnesota Population Center and the Institute for the Environment, represents a good example of integration of cross-disciplinary interests (in this case demographic and land-use data over time).

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Comment 4:
In addition to differences between disciplines and associated data, there are differences associated with the data themselves that relate to cost-benefits of management and archiving – e.g., can the data be reproduced, do the data serve a canonical reference value? Consequently, minimum policies about data management, sharing, and preservation may not adequately ensure an overall cost-effective and sustainable data environment. Federal agencies and associated policies for federally funded research could:

- set data retention guidelines based on replicability, importance and potential use
- support establishment of discipline-based repositories with an ongoing deposit mandate to ensure long-term access to data at a scalable cost. See the NSB report[3] on long-lived digital data for examples.
- address the fact that existing disciplinary repositories do not always accept the full range of data generated by researchers in their field. For example, CUAHSI's Hydrologic Information System[4] only accepts geo-referenced data, yet researchers may do lab-based hydrologic research and produce relevant data. Policies should encourage existing data repositories that receive federal funds to accept and curate a broader range of data in their disciplines, and agencies should fund development of new repositories that bridge disciplinary gaps.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Comment 5:
Data Management Plans (DMPs) have had a positive impact on highlighting the issues of stewardship of digital data, however the benefits and strategies involved still remain limited within many research communities. Requiring a DMP for funding applications is a good start, however more specific guidance (see comment 3) and stronger incentives to openly share (see comment 1) would contribute to the implementation. All stakeholders should promote policy compliance along with community-based standards making, specifically:

- Research communities can

[3] NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century (2005) http://www.nsf.gov/pubs/2005/nsb0540/
[4] Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) http://his.cuahsi.org/

    - collaborate on the development and operation of domain-based infrastructure for data stewardship.
    - recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement.
    - create tools for sharing and preserving data and new methods, standards, and tools for metadata capture.
- Universities and research organizations can
    - clarify organizational/institutional intellectual property policies regarding data.
    - promote data publication and sharing as part of the tenure process considerations.
    - recognize and support data management systems as contributing to enterprise-level research infrastructure.
    - require DMPs and data sharing implementation for all doctoral dissertations.
    - offer career development paths for data managers and data scientists on campus.
    - include data training as part of the science curriculum. “Open data access with appropriate ethical restrictions can be viewed as a new core principle for developing a global data infrastructure” (Parsons, 2011)[5]
- Libraries and librarians can
    - curate data as part of the scholarly communication cycle
    - collaborate in development of campus infrastructure for discovery, management, and distribution of data. Provide educational and consulting services about data plans, data management, and access to data repositories.
- Publishers can
    - enable and encourage data attribution through citations within publications and research documentation.
    - eliminate barriers to data access through pay-wall or copyright restrictions to “data supplements” in research articles.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Comment 6:
In many ways the real costs of preserving and making data accessible are just beginning to emerge, with growing recognition that human capital and expertise are often as essential to maximizing the value of data as technical infrastructure. Several areas for agency action include:

- funding research and education on the costs and benefits of data stewardship.
- supporting development of community standards and cost-effective infrastructure such as tools for deposit, management and discovery.
- encouraging greater specificity in data management plans to address sustainability.

[5] Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=e31a1aab-b01e-4e7e-9b10-0fd93d4b710f

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Comment 7:
Standardizing data stewardship policies across granting agencies and providing models of best practices would simplify and streamline both compliance and verification. Agencies have leverage at the time of submission and completion of grants, and mechanisms to track compliance could occur at both points. Strategies might include:

- tying new grant awards to prior data stewardship (for example, the NSF Computer & Information Science and Engineering Directorate Guidance for Proposals and Awards requires that data sharing progress and outcomes must be reported in award annual reports, the final report and subsequent proposals by the PI and Co-PIs).[6]
- enabling better tracking by implementing a data ID registry that is linked to stewardship best practices. For example DataCite is issuing persistent data identifiers for data, however, this is only one component of a verification system. Issuing a persistent and exclusive data ID, that is only given to data in repositories that meet minimum standards for openness and preservation, would provide clear evidence of data stewardship. Models exist, such as TRAC, for trustworthy repository standards that could be adopted.[7]

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8:
Federal agencies can stimulate use of data by supporting the development of public access data archives to enable discovery and download. Public APIs could allow individuals who may not have specialized data analysis software to access and make use of the data, and could automate collection and aggregation for new capabilities and services. Specific agency actions might include:

- providing funds to discipline repositories and data producers to create additional metadata, visualization tools, and augmented holding records to allow users in unrelated fields and without the necessary software or expertise better access and to encourage reuse. For example, the Minnesota Geological Survey proposes to augment their public data collection including surficial geology data that is of interest to citizens and scientists alike who do not have access to the specialized and expensive GIS software to read the professional data. The aforementioned TerraPop project will similarly make data and general/specialized tools available to diverse communities.

[6] CISE Proposal and Award Guidance. Last updated September 15, 2011 at http://www.nsf.gov/cise/cise_dmp.jsp.

[7] Trustworthy repositories audit & certification (TRAC) criteria and checklist. http://catalog.crl.edu/record=b2212602~S1

- establishing awards for web-based visualization tools of existing data –e.g., the successful *Digging into Data Challenge*[8] that NSF has partially sponsored over the past three years. Also, Data.gov[9] has transformed into a "cloud-based Open Data platform for citizens, developers and government agencies" through web-based apps and clear re-use policies.
- promoting standardization of metadata using ontologies or RDF to incorporate data into new domains or use for non-traditional research (i.e. visual arts).
- provide awards for innovative re-use of research data. Alternatively, provide funding to repositories to host a data challenge with monetary incentives,
- stimulating new research in semantic technology, the underlying technology that supports linked open data.
- enabling citizen science efforts which have engaged the public in data gathering or data classification activities –e.g., Galaxy Zoo[10], is a Citizen Science Alliance project that has engaged over 250,000 members of the public in cloud-based scientific discovery.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9:
Strong attribution norms already exist within the academic and professional communities with regard to research publication, but are not as well-developed with respect to data. Policies embracing certain forms of open licensing (such as Creative Commons licenses) might strengthen attribution practices. Federal policies could contribute to developing strong data attribution practices by:

- encouraging development of data citation standards and practices. This would improve capabilities for tracking re-use and also provide impact measure of individual researchers' contributions.
- requiring data citation following the above standards in publications resulting from federally funded research (as opposed to simply mentioning data sources in the bodies of articles.)

Standards for interoperability, re-use and re-purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community driven data standards effort.

Comment 10:
One of the biggest obstacles to data discovery, access and reuse is the lack of standardization of descriptive metadata adoption and practice within a specific scientific domain. This constrains the ability to search effectively across federated data repositories, as well as to ensure search effectiveness within a central index to data assets. Community-driven data standards efforts are essential within a domain context, while more general standards are necessary to promote interoperability nationally and internationally, and at more general levels of discovery (i.e.,

[8] The Digging into Data Challenge is an international collaboration administered by the Office of Digital Humanities at the National Endowment for the Humanities. http://www.diggingintodata.org/
[9] Data.gov - Open Federal Data. http://explore.data.gov
[10] Galaxy Zoo Project. http://www.galaxyzoo.org/

cross-domain searching). The catalog of community-driven domain-based data standards are too numerous to list out here, but prominent examples include standards produced by the Federal Geographic Data Committee (FGDC) for digital geospatial metadata, and Darwin Core, which functions as an extension of Dublin Core for biodiversity informatics applications.

Community-based expertise should be used to develop standards and conventions for data structure and metadata management specific to a discipline's research output. However, certain areas of metadata are of particular cross-disciplinary interest. Measures and representations of time and space are important cross-indexing aspects of many datasets. Researchers and disciplines should be encouraged to ensure inclusion and interoperability on these measures through use of standard models for recording these data. Repositories should also be encouraged to explore further the use of semantic web technologies (RDF and URL-identified entity and relationship vocabularies) and linked data to leverage discovery.

Recognition of a scientific data standards registry, engagement of national and international standards organizations and inter-agency participation in the creation and endorsement of standards are ways to encourage requisite coordination of standards.

In addition to the attention on standardized descriptive metadata, *technical metadata* is of increasing importance and utility. Technical metadata can provide information concerning the instrumentation and methodology used to create the data, and may be critical to ensure validity, reproducibility, and re-use by users not involved in generating the original data set. This should be linked to the data as part of stewardship. Currently some disciplines separate this out and describe data creation methods in a research article, but this information must be linked or kept together in an open data environment for the data to be usable and trustworthy.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Comment 11:
There are examples within various domains that have been successful in producing data standards for interoperability and re-use –e.g., the astronomy community's development of the National Virtual Observatory (NVO) and the associated International alliance. The data standard such as FITS allows for data from different instruments to embed metadata elements that can be read by open source software. The success of this relies on ease (if not automation) of metadata creation, as the instrument or software embed metadata, such as celestial coordinates, into the image. This is similar to the GIS communities' standard of FGDC, but is also software dependent.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Comment 12:
Federal agencies should support community-based efforts that promote and coordinate standards across international communities -- e.g., aligning policies with the processes supported by such bodies as the Committee on Data for Science and Technology (CODATA). Further, support for repository "nodes" of data-sharing infrastructure, such as the NSF DataNet

project DataOne, will bring together disciplinary communities that might not have the capacity to or budgetary incentives to combine efforts.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13:
Linking publications with the underlying or associated data provides significant benefit for research. Dryad is a good example of a discipline repository linking the underlying data supporting journal publications in the biosciences, with plans[11] to integrate data deposit for society and for-profit journals in the field within this publicly accessible data archive. Requirements for unique and persistent data identifiers will aid in this linkage and enable tracking of re-use. Federal agencies should support the work of the International DOI Foundation (IDF), DataCite, and other organizations with the goal of making a unified international standard and support structure linking between publications and associated data.

[11] Dryad Wiki (update January 4, 2012) Submission Integration. http://wiki.datadryad.org/Submission_Integration

Marianne Buehler
Fri 1/13/2012 1:09 AM
open access to research and data response

Dear OSTP,

I am responding to your call for interested parties to comment on the open use of data sets and the resulting research that is created by federally-funded research. Taxpayers are funding this research that is being conducted to improve health, science, medicine, our environment, and more. As a taxpayer, I want this valuable research data and its published outcomes to be available for other researchers to build upon existing scholarship to accelerate the pace of more research to make our country and world a healthier, better, and safer place to live in.

For too many years, commercial publishers have taken advantage of taxpayer funding of research where they charge big money for libraries and researchers to buy back the research that was originally funded for the benefit of the US and beyond.

Please consider a course of action beneficial to the greater good of our country and the global milieu.

Thank you for the opportunity to speak out in favor of open access to research,

Marianne

Fri 1/13/2012 1:26 AM
Personal comments on data preservation

I am Amber Boehnlein, the head of Scientific Computing at SLAC

SLAC is a DOE funded laboratory that hosts three user facilities (Linac Coherent Light Source (LCLS), Stanford Synchrotron Radiation Lightsource (SSRL) and Facility for Advanced Accelerator Experiment Tests (FACET). SLAC and Stanford staff participate as users of these facilities. Additionally SLAC has a particle physics and particle astrophysics program. The SLAC community of staff and users consists of representatives of many different scientific disciplines. This community participates in experimental research at other facilities as well. Our participation in data preservation and sharing include participation in the Study Group for Data Preservation and Long Term Analysis in High Energy Physics (http://www.dphep.org/) for BaBAR long term data analysis (http://today.slac.stanford.edu/feature/2010/babar-prototype-data-servers.asp) and the Joint Center for Structural Genomics "Structure Gallery" (http://www.jcsg.org/). We operate the Fermi-LAT production pipeline that produces data that is publically available. We also host the XLDB conference that brings together a broad community to discuss the issues of 'Big Data' (http://www.xldb.org/)

Our primary interest in data discoverability, preservation and access is to facilitate scientific advances and discovery within the domain communities affordably and with the appropriate level of controls to insure integrity of the results and appropriate attribution.

There are common issues for data accessibility across the domains represented by the SLAC scientific community. The common issues are typically social or technical—scientific challenges with data accessibility tend to be more unique.

Given the stated goal of data preservation accessibility to improve the productivity of the American scientific enterprise, one considers ways to improve the productivity of scientists. Scientists are motivated to transition to new practices when those practices increase their scientific productivity at a cost that they can afford. Stating expectations (rather than policies) for accessibility and preservation of publically funded scientific digital data coupled with Federal investment in reducing the technical challenges, burdens and costs to the research communities is likely to lead to faster change in the area of data preservation and accessibility than wide scale policy mandates will. This approach will be especially effective in domains where standardization of data formats and data is acknowledged to be useful, however, is a low priority for implementation for a variety of reasons. In order to facilitate scientific discovery, from the perspective of the scientists who produce the data and the scientists who would use data that they did not gather, mechanisms for data sharing, data preservation, performing operations on data and understanding the intellectual content of the data have to be intuitive and easy to use.

Many of the issues of digital data lifecycle and the associated technical challenges are legitimate topics of research. It is usually insufficient to preserve the data without also having mechanisms for the preservation and encapsulation of the knowledge and tools that produced the data and can operate on it-these are difficult problems without known technical solutions. Given the growth of 'big' data and associated analytics in private industry, there are untapped areas for public/private partnerships in this area.

As a template for potential Federal investments to reduce the technical burdens, achieving economies of scale, and the power of aggregating expertise and research to serve the needs of the diverse scientific communities, one could look to the success of advanced computation programs. Fifteen years ago, the state of unclassified advanced computation was similar to today's situation with scientific data management. Scientific programs that needed large scale computing built their own computers and developed their own algorithms for their own needs. This methodology was scientifically limiting. To address those limitations, programs to build capacity and capability machines as multiple science computational facilities and the associated research programs in applied math and computer science to enable the science communities to use those resources were developed. As a note, these facilities are now in use for some areas engineering and industrial design— infrastructure built for advanced scientific computation now is used in ways that directly grow the U.S. economy.

Aggregated mid-scale computing resources have been developed during this time within some disciplines. Programs such as Open Science Grid and Earth System Grid have been effective in building communities, providing user support, and aggregating the results of basic research into production solutions. Such programs lead to a "commoditization" that ultimately could lead to more science for fixed costs. Scientific databanks provide additional pilot programs that have been effective in promoting some degree of data preservation and data sharing.

Within the concept of building Scientific Data Management Facilities, existing user experimental facilities (where experimental scientific data is collected, cataloged and sometimes analyzed) and dedicated data centers that could serve aggregations of individual small labs would have a role in working with the scientific communities, provide expertise, developing technical solution, and carry out research projects in partnership with other academic institutions and private sector partners. This could determine informed practices and policies that met the scientific needs, understand the differences in data lifetime and advanced the scientific productivity within an affordable budget.

Fri 1/13/2012 1:46 AM
Response to RFI on Digital Data

Name/Email: Abigail Goben

Affiliation/Organization: I am writing this as a private citizen, though I am employed by the library at a publicly funded university which receives federal grants. Views herein expressed are my own as my employer has submitted their own response already. I am also writing as a blogger at HedgehogLibrarian.com, where I have been exploring data in my Friday blog posts with the advent of the data science movement. Show citation box

City, State: Chicago, IL Show citation box

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The NSF Data Management Plan is an excellent start but it needs improvement. It needs teeth and there need to be gold standardsThere should be immediate adoption of an open access approach to data that is federally funded. This shouldn't wait for a private company to eat up data and hide it from researchersThere needs to be a policy on citation so that researchers creating the data have a way to be recognized for their workMetadata standards should be created so researchers have an idea of what is requiredShow citation box

(2) The primary stakeholder is the taxpayer who is funding this research--certainly not the publishers. That they were listed first, even before the scientists, is really troubling. Federal funding is already paying for creation of this data and as federal employees are not able to claim copyright and IP, these researchers acting as federal employees with the grant money should be similarly bound.
Show citation box

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

There are many leaders in the various scientific fields who, I'm sure, would be happy to consult on developing appropriate needs for the different types of data. Those types should be made as coherent as possible though so that data can move across disciplinesShow citation box

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Again, I'd speak to experts in the field. I would avoid passing those costs to private vendors at all costs though, they're more likely to raise prices and cut off our access down the road. Show citation box

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

From this I'll speak from a librarian perspective, as that is my present role. Librarians have a number of roles that we already play that could immediately assist with this. Librarians are already teaching about data management plans and trying to raise the questions about preservation, storage needs, metadata, short and long term access, open data, etc. I'm preparing a tutorial for my faculty this spring and will be teaching it over the summer. We are already experts in creating metadata and have thousands of catalogers who could assist in tagging data in such a way that removes the burden from the researchers (granted--we would need funding for this, library budgets have been incredibly slashed). Libraries can help with storage, again if there is funding available to set this up and maintain it, and with access. Our missions are about providing access to information, we would be happy to take on data. Scientists and researchers already see us as repositories of books, is it that much of a move to have us be seen as repositories of data? Many librarians are already research partners working with the various scientific fields, so we know what kind of data they are generating. Finally, librarians, because of our role outside of the lab, tend to see the big picture and have excellent ideas about the translation of data, which would promote translational and collaborative science. Show citation box

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Funding needs to be built into the grants, not tried to be scraped together by the institution after the grant is finished. I would suggest working to create hubs based around major public universities that are already working on much of this. If each university takes a subject area, then that could be funded there and data made available without each university or each scientist trying to reinvent the wheel. Show citation box

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Make it as EASY as possibleProvide a federally funded way to clean up the data and apply metadata. If that means the US Government has a division of Data and Metadata, that might be the best thing for it. If the scientist need only turn over their data and someone else will do the cleaning and tagging, they are more likely to comply. Those funded locations would fit nicely into academic research libraries. Show citation box

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Offer awards to young scientists for use of old and publicly accessible data. Too many young scientists are vying against highly published scientists who have moved on from their old data. Find funding ways to encourage people to reuse data rather than only value new data. Encourage academic institutions to grant tenure based on science using old data and/or Find a way to scrape data, particularly medical data of personal information so scientists and researchers can use it without having to go through IRB every five minutes. Show citation box

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Get the experts in the field to develop what standard citation is. Require that publishers verify as part of their work (though they'll just hand it back to researchers)Find a way to convince administrators at R1 institutions that creation of data sets is valued by the federal government. Perhaps offer awards for data sets that have successful reuse? Show citation box

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Open and FreeShow citation box

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?Show citation box

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

While I have no doubt it will be seen as naive, I think we need to lead the way to say that we have good information that we're willing to share. Being open and willing to share will start conversations. If we can establish excellent practices for data storage and access, we can also offer to assist overseas in storing their data and processing it for them. Show citation box

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Publishers cannot be allowed to privatize data or say that by publishing it they have created some intellectual propertyA standard copyright agreement that data funded by a federal grant, while represented by the publisher, is not copyrighted. URIs/DOIs would help

I recommended for further reading this excellent post by Heather Pinowar: http://researchremix.wordpress.com/2012/01/11/nsf-data-vision/

I well recognize my own library bias in this and hope you can appreciate it. Libraries, particularly academic ones, are already trying to solve these problems. We have the expertise but we are struggling with time, funding and federal support. If you could provide those, we'd be more than happy to pick up the ball and run with it. We collaborate well with each other, we work hard to support our faculty, and we're here to provide access.

Thank you for the opportunity to provide feedback and please let me know if I can provide any further information.

Abigail Goben

--
Abigail Goben, MLS

Interagency Working Group on Digital Data
National Science and Technology Council
Executive Office of the President
Washington, DC 20502

12 January 2012

Response to Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research (76 Fed. Reg., 68517, 4 Nov. 2011)

To Whom It May Concern:

With over 400,000 members in over 160 countries world-wide, IEEE is the world's largest professional association dedicated to advancing technological innovation and excellence for the benefit of humanity. In addition to our conferences, standards and other activities, IEEE publishes more than 150 transactions, journals and magazines, which represent more than 30% of the world's annually published literature in electrotechnology, computing and related fields.

IEEE is a strong supporter of public access to government-funded research results within a context that protects and advances other important societal interests inherent in scholarly publishing, including peer review and reserving the intellectual property rights of authors and publishers. IEEE's long-standing policies permit authors to "self-archive," i.e., to post their articles either on their personal web sites or their employers' web sites, consistent with open access practices. IEEE is also experimenting with several open access business models with the goal of supporting sustainable open-access publishing. Our views were outlined in greater detail in a 2007 IEEE Position Statement on "Scholarly Publishing", which is available on-line at:
http://www.ieee.org/documents/IEEE_Publishing_Principles.pdf.

IEEE has responded to the Office of Science and Technology Policy RFI on Public Access to Scholarly Publications, issued in parallel with this RFI on Public Access to Digital Data. Of course, there are qualitative differences between data sets and research articles that would lead IEEE to separate its positions in these two responses:

- The scholarly article is an author's expression of ideas with significant value added by the publisher, primarily in the form of peer review, but also in copy editing and formatting, and in other ways as well. Data sets, in and of themselves, are not normally considered part of a research article, but rather supplemental material sometimes made available with an article. However, advances in science and

technology in recent years show the immense potential of well-curated data repositories, and it is appropriate to understand necessary distinctions between published research articles and the data that underlie them.

- Science and engineering benefit from a free exchange of data that allows researchers to use data in the replication of primary research or to build upon prior research. Curation of data sets on behalf of scientific and engineering communities is a valuable service. IEEE supports public initiatives that enable researchers to make their data available to such repositories. To date, scholarly publishers have not made an investment in data archives that would be comparable to their collected volumes of published research articles.
- Although case law is developing for copyright in compiled databases, data are generally accepted as non-copyrightable facts, and therefore more amenable to free exchange without encumbrance of copyright. This is not to assume copyright does not or will not play a role in data management, only that a tradition does not exist in the same way that copyright is clearly established for collected research articles.

Ongoing maintenance of a growing number of repositories of data sets will eventually require a financial model to assure sustained support. Such models exist for research publications, including those made universally available via open access. Private-sector companies have readily exploited databases on a sustained and profit-making basis. Public databases supported by government funding are familiar, and institutional repositories are gaining experience in archiving data sets. In some cases, not-for-profit organizations have made data sets publicly available for free, while charging fees for value-added services to defray costs of archive maintenance.

IEEE strongly supports formal collaboration among stakeholders in public-private working groups. As with the policy-making process for public access to scholarly publications, we would be pleased to provide representatives to serve on collaborative panels convened to deal with these issues.

Please see detailed responses to OSTP questions following this letter.

Sincerely,

Gordon Day
IEEE President and CEO

With respect to the specific questions posed in the Request for Information, IEEE is pleased to provide the following input:

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

The US Federal government may play a most beneficial role in partnership with standards-setting organizations that can develop and promote responsible sharing of research data. Public access to data presents a great benefit to users who have the resources and knowledge to exploit such data, either to further advance the research or to develop commercially viable products. As such, public access opens the digital archive to a global market, with no special advantage to American researchers or businesses. US job creators may therefore share a resource that can serve an expanding economy, but policy makers should recognize that the benefits cannot be confined to the US scientific community.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Data associated with scholarly research represent a class of intellectual property that can and should be distinguished from peer-reviewed research articles. Data represent information created in the course of research and subject to researchers' interpretation in the reporting of results. Traditionally, authors have transferred copyright in the peer-reviewed and copy-edited articles that report results in the publishers' chosen formats; publishers have protected copyright in these articles of record. Generally, research data have resided with the authors, who may share the data freely, or with conditions. Open access publishing models frequently recognize the author's right to retain copyright in research articles, while inviting users to freely reuse data. However, most authors will also expect attribution and credit for their primary research, including the underlying data.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Somewhat similarly to the life cycle of research articles in various disciplines, data in certain fields may have more compelling near-term value than data in other fields. This is true, for instance, in biomedical fields compared to other types of technical study. It may be reasonable to assume that the demand for publicly accessible data would parallel that of published research.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Public access policies must recognize that the archiving and curation of data carry costs, just as publishing of research does. Consistent with the Principles of Scholarly Publishing, identified above, the IEEE believes that public access models should be sustainable, whether by assured government funding, or by opportunities for repositories to charge for value-added services to offset the cost of curating publicly accessible data sets. Government funding is appropriate to creating infrastructure and propagating data interchange standards, but offsetting revenue models are desirable as a way to avoid overreliance on government support.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Institutions with the appropriate resources may offer services to researchers and authors to curate data sets on their behalf and for the good of scientific communities at large. Standards for interoperability among content repositories will ensure discoverability of data.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Recognizing that ongoing funding is needed to support hosting and maintenance of data sets, as well as assuring their availability in future formats, research grants should provide for some level of support for data archiving.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The National Science Foundation in 2011 introduced a requirement for grant recipients to file a plan for sharing of data created as a result of NSF funding. Such a plan can readily be checked by the agency for compliance. We suggest that this approach provides a sufficient process for grantees to describe their measures in a way that can be verified; allowing for grant support to promote compliance would assure participation by researchers.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

As with public access to scholarly publications, care should be taken to see that small and medium-sized businesses are able to benefit from public access to funded research data. However, also in parallel with access to publications, public access to data should be sustainable through a business model that assures continued revenue to support costs of maintaining and curating a data collection. One means of support could be continuing US Government grants to support content repositories; however, government funding is less than ideal for sustainability. Encouraging private sector investment in fee-based services offered along with free access to data repositories could generate revenue that can be channeled to support the repositories themselves.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

If appropriate models are adapted from traditional forms of scholarly publishing, it should be possible to achieve at least the same level of attribution and credit in reporting secondary results, but this can be validated only with experience. Attribution and provenance are critical to participation. The issue of ambiguity is an obstacle. The IEEE and other scholarly publishers are actively pursuing ways to disambiguate author names; such projects, along with ORCID (see response to Q. 11, below) will help ensure that the researcher is identified with their data.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The global organization DataCite has already been successful in registering Digital Object Identifiers (DOIs) for data sets. The same continued industry collaboration that has successfully introduced the use of DOIs for articles should be brought to bear for data sets. Public access is furthered by the use of metadata that identifies US Government funding sources for research data that can be made publicly available from any content repository.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

CrossRef, a not-for-profit association founded by scholarly publishers in 2002, has pioneered the development of the Digital Object Identifier (DOI) as a unique identifier to locate every published scholarly work. IEEE and other scholarly publishers would support an initiative to enhance article metadata with information clearly identifying the agency responsible for funding the described research. Agencies would save considerable effort and expense by supporting improvements to DOI metadata that would automatically capture this essential information.

Similar collaboration between public and private sectors will lead to success in efforts to provide identifiers for data sets that are created in connection with funded research. Examples include DataCite (www.datacite.org) and the NISO/NFAIS Working Group on Supplementary Journal Information (www.niso.org).

Another example of a collaborative approach among publishing partners is the Open Researcher & Contributor ID (ORCID) project (www.orcid.org), a successful public-private partnership with 275 participating organizations. This project addresses name ambiguity among individual authors and the resulting difficulties in consistent author attribution.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

The 2009 report of the Interagency Working Group on Digital Data of the NSTC has already proposed that funding agencies require data management plans for funded projects. While recognizing that some agencies, such as the National Institutes of Health, may already have detailed policies in place, and that each agency will have unique considerations, care should be taken to approach data management in a uniform and consistent way across the spectrum of Federal agencies.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

As with public access to scholarly publications, interoperability among content repositories would be essential to successful public access to digital data. Standards-making for data collections may be informed by the processes used to achieve interoperability for content repositories of published articles, and indeed, the same stakeholders are likely to have similar interests in the parallel cases. Government sponsorship of public-private working groups and task forces is an appropriate and productive way to bring parties together for sustained collaboration. Establishing the appropriate identifier for the publication, person and data will help facilitate the link. CrossRef is a great example of the impact the DOI has had on discovery and access.

In conclusion, IEEE appreciates the opportunity to provide input to the Working Group's deliberations and stands ready to answer questions and provide additional information as needed. If we can be of any further assistance, please contact Kenneth Moore, Director, IEEE Book & Information Services (e-mail: k.moore@ieee.org; tel: 732-562-3954).

Thu 1/12/2012 12:46 PM
Pat Lambert
Public access to federally funded scientific research

We paid for it. It is public information. It is useful information.
Jefferson: There is not a truth existing which I fear or would wish unknown to the whole world.

To:
The Office of Science & Technology Policy

From:
Erich S. Huang, MD, PhD
Director, Cancer Research
Sage Bionetworks

Re:
OSTP RFI: Public Access to Digital Data Resulting from Federally Funded Scientific Research

Transparency in biomedical research isn't just good science, it's good medicine.

Why? Disease's complexity requires many eyes coming from many angles. This is clear on the front line where the practice of medicine can be a humbling process. I have had cases where a combination of knowledge, training, and luck have vanquished suffering, and also cases where the complexity of affliction has confounded me at every turn. This is equally true in our investigation of disease. No one has a monopoly on insight:

> *"Every intellectual has a very special responsibility. He has the privilege and opportunity of studying. In return, he owes it to his fellow men (or 'to society') to represent the results of his study as simply, clearly and modestly as he can."*
>
> –Karl Popper

Karl Popper's above quote is fitting. While you maintain confidence and decisiveness to effectively treat your patients, you are ever conscious of the repercussions of hubris when your patients' lives and well-being are at stake. A good physician-scientist is ultimately an intellectually "modest" doctor—you know that you don't always have all the answers.

With this in mind, the Office of Science and Technology Policy's Request for Information on Public Access to Digital Data Resulting From Federally Funded Scientific Research is particularly timely. The cloudburst of complex "big" biomedical data precipitated by genomic technologies is so promising, yet so difficult to contend with, that only an open exchange of data and ideas can unravel its complexity and maximize its benefit to our patients. This free market is threatened today by legislation such as HR 3699 that seeks to shut the door on free access to NIH-funded publications. Publicly-funded in the public interest? Throw the doors open.

I have seen the complexities of such data humble an institution. At Duke, where I received my medical and scientific training, a mixture of hubris and opacity in data and methods resulted in tenuous science. Institutionally, intentions were good—who wouldn't want to figure out how to use genomics to match the best therapy to a patient's disease? But the status quo (and this is *everywhere*) for managing Federally-funded data made open review by the scientific community difficult, if not impossible. While there are data-sharing plans mandated in the Federal grants process, they are unenforced.

Academic research has a cultural problem. There is no doubt that taxpayer funding of basic and translational research catalyzes progress in treating human disease and the basic discoveries that fuel advances, but there is also no doubt that funded individuals are conflicted.

As with any human endeavor, progress and discovery in the life sciences is driven by a complex mixture of desire to serve society and self-interest. And here we scientists must remind ourselves how we got to our positions: Taxpayer money. My doctoral training was Federally-funded. My stipend as a surgical resident was Federally-supported. My first faculty grant was Federal. It is easy to say "my grant", but really, it's the *public's* grant. My faculty development award's real purpose was to put me in the position to contribute meaningfully to the NIH's mission of mitigating disease. A *public* mission. Yet it is easy to see that many faculty see Federal grants as the mandatory offsets for covering their salaries, their research groups and, ultimately, their tenure. And surely university administrators often see things the same way.

So incentives are all in the wrong place: the currency of academic promotion is research grants and publications. If you're an Assistant Professor and are fortunate enough to obtain a 5-year Federal research grant that funds valuable and interesting cancer research, 4 years into "your" grant—with your eye cocked on competitive renewal (because you probably need two or three of these to obtain tenure), and publishing your findings in a prestigious scientific journal—why would you share these data? Why wouldn't you maximize your competitive advantage over all the other Assistant, Associate, and Full Professors out there?

We need to recalibrate how we do science. There's plenty of room for individuals to demonstrate their great ideas in a world where publicly-funded science is openly accessible. You can see a vibrant exchange of ideas in the open source software world on Github (www.github.com). This web service allows programmers to store their software projects in a frictionless open system that allows others to see their code, make copies of code, suggest improvements, or even write improvements themselves. For Facebook, Twitter and LinkedIn programmers, a job candidate's Github repository is considered far more valuable a record of their mettle than their resume. Even in this open, free-flowing milieu, programmers do pretty well for themselves; and their value can be measured by their contributions.

In science, the simple fact is that where the funding goes, scientists go too. Therefore an easy first step would be a stipulation that publicly-funded research be placed in an open-access repository and that standards for data quality, usability and accessibility be vetted by a neutral party. No open access? No renewal. Besides idealism, there are pragmatic reasons for doing this. See the Harvard Provost's response to the related RFI on open access publications for a cogent analysis of the economic benefits of open science. Another consideration is that public investments in research should provide maximal return on investment. We paid for it, right? Why not let the community (hackers, biotechs, colleagues, even Big Pharma) figure out how else to unlock a dataset's full value rather than leaving it to be skimmed by an individual marquee principal investigator? Really, we should see these data as *vital infrastructure*, like Eisenhower's Interstate Highway System and treat them as such—accessible to all.[1]

We owe this to our patients.

In the end, the scientific enterprise is about sharing and communicating "simply, clearly, and modestly". Why? Because none of us knows all the answers.

[1] And the economic benefits of our interstate system has far exceeded Eisenhower's investment.

This is a response to the Office of Science and Technology Policy's "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." Note that our responses cover questions 1-9. We do not respond to questions 10-13.

Request for Information:
http://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific

Responses written by:
Matthew Mayernik – mayernik@ucar.edu
Mary Marlino – marlino@ucar.edu
Karon Kelly – kkelly@ucar.edu

Affiliation:
NCAR Library
National Center for Atmospheric Research (NCAR) / University Corporation for Atmospheric Research (UCAR)
Boulder, CO

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

RESPONSE:

Federal policies can strongly influence how data resulting from federally funded scientific research are managed and preserved. Such policies should focus on creating institutional structures within and across disciplines such that researchers organize their research practices around data sharing and re-use.

Any new policies must recognize that providing access to and preserving digital data is a profoundly human process. Technologies facilitate the collection or creation of digital data, as well as the discovery, transmission, and preservation of data across space and time. But digital data can only be collected, accessed, and preserved through the purposive actions of individuals and organizations across the public and private sectors.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

RESPONSE:

Intellectual property interests around digital data should focus on the creation of social norms within particular communities, not new legal protections. Facts, such as "the temperature in Boulder, CO, was 62 F on Jan. 10, 2012," are not copyrightable, thus most

forms of scientific data do not fall under copyright control. "Data organization," on the other hand, can be put under copyright licenses (Stodden, 2009). Free/Open copyright licenses, such as the GNU General Free Documentation License (GFDL) and the Creative Commons licenses, can thus be applied to data organization systems. Applying such licenses to data, however, will inevitably complicate data sharing and integration efforts of any large scale for a couple of reasons. First, there is an amorphous line between what can and cannot be copyrighted within databases. Within a database, what is an uncopyrightable fact and what is a copyrightable arrangement of facts? Second, different copyright licenses have different usage requirements. The implication of this is that querying across ten databases may return results with ten different usage licenses. The data user is then put in the difficult position of navigating complex legal regimes before bringing data together and releasing subsequent results.

Because of these difficulties in navigating intellectual property issues around data, the Science Commons project, an off-shoot of the Creative Commons organization, recommends that scientific data be assigned to the public domain rather than being placed under copyright of any form (Wilbanks, 2008). Their "Protocol for Implementing Open Access Data" (http://sciencecommons.org/projects/publishing/open-access-data-protocol) outlines how the public domain is the most appropriate way to enable the widest use of scientific data. Putting data in the public domain eliminates data use restrictions, it enables data integration in that data from disparate projects all have the same legal status, and it encourages non-legal means for resolving problems related to data use. In lieu of copyright-based methods of controlling data use, federal policies should promote norms within scientific communities as to how data should be made available, used, and attributed. Scientists should check with data centers for data use and attribution policies, and work with collaborators to ensure that the usage and attribution of others' data meets with community accepted practices. For example, the 2007-2008 International Polar Year (IPY) project required that data be "made available fully, freely, openly, and on the shortest feasible timescale…equitable, non-discriminatory access to all data preferably free of cost, but some reasonable cost-recovery is acceptable" (IPY, 2008, pg. 3). Similarly, the seismology community has developed a norm in which data are released to the broader community after a specified period of time. This norm is codified within the NSF Division of Earth Sciences policy: "For those programs in which selected principle investigators have initial periods of exclusive data use, data should be made openly available as soon as possible, but no later than two (2) years after the data were collected." (NSF Division of Earth Sciences, 2010, pg. 2).

Not all data can be assigned to the public domain. Data collected about individuals, medical data, classified data, and other sensitive data (such as the locations of endangered species), are, and should be, withheld from subsequent use unless measures have been taken to ensure their compliance with ethical and legal considerations, such as anonymization, declassification, or removing sensitive data by other means.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

RESPONSE:

Disciplines do exhibit differences in their data management practices. For example, some disciplines have higher levels of adoption of data and metadata standards than others. However, recent studies have shown that the variation in data management practices is often as large within individual disciplines as it is between disciplines. This intra-disciplinary variation in data management practices can be seen in astronomy (Wynholds, et al, 2011), ecology (Mayernik, Batcheller, & Borgman, 2011), and the quantitative social sciences (Pienta, Alter, & Lyle, 2010), among others.

When looking at data management practices from an institutional perspective, however, it is possible to see that many important data management challenges span the academic disciplines. Many questions important to data management and preservation are discipline agnostic: What data management institutions exist (or do not exist) for particular disciplines? How well are they known by researchers within those disciplines? Do institutions exist that create and promote data format, transmission, and preservation standards? Do data centers/repositories/archives exist? Does the discipline have a tradition of working with trained data management experts within library and/or computing institutions? Is data management/sharing valued by the institutional structures that reward achievements within a discipline, such as graduate student advancement, and faculty tenure and promotion decisions?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Please see responses to question #3, #6, and #8.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

RESPONSE:

Data management plans are the first step in the data management process. *Plans help collaborators orient each other to their data management needs and options, but only lead to effective data management down the road if data are actively managed as an ongoing process.*

The stakeholders listed in this question can engage with researchers from the beginning of the research planning process to promote and facilitate data management. Many university and research libraries are now offering data management planning services, in which they work with researchers to develop a data management plan for a research proposal. These planning services build relationships between researchers who create or collect data and the library and university institutions that have the expertise and (ideally) the capacity to ensure that data are made available and preserved over time.

As an enforcement mechanism, funding agencies and universities can reduce/withdraw funding from Principal Investigators if data management plans are not carried out. Similarly, universities and funders can deny funding future projects based on a Principal Investigator's insufficient data management actions in the past. An incentive-based approach to promoting data management would reward researchers for reusing and repurposing existing data collections, thereby increasing the demand for quality data collections.

Publishers can build relationships with data archives to build pipelines that enable researchers to deposit data with data archives as a part of the publication process. Publishers can also request that researchers provide citations to the data that were used to produce a publication.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

RESPONSE:

Funding agencies should work with existing data archives to understand different cost and sustainability models for data management and preservation work. Numerous data archives exist in many disciplines. These organizations understand the costs involved in collecting data, organizing and providing access to data, maintaining data over time, and preserving digital resources. How much does it cost for them to do their work? Understanding current data archiving practices would greatly inform future funding for long-term preservation and access of data.

Another way to assess data management and preservation costs would be to fund select (but diverse) pilot projects where the economics of data preservation and accessibility are explicitly studied. For example, the NSF could explicitly study the costs of data management and preservation within the forthcoming National Ecological Observatory Network (NEON, http://www.neoninc.org/), or the Advanced Cooperative Arctic Data and Information Service (ACADIS, http://www.aoncadis.org/home.htm;jsessionid=B7A0C3ABCA80E00C83D4A316D76DE570), which is being managed by NCAR and the National Snow and Ice Data Center, both in Boulder, CO.

Critically, any attempt to assess data management, preservation, and long-term access must take a long-term view. The costs of data preservation and access cannot be quantified by looking at a two-to-three year window. The largest costs of data management and preservation are ultimately related to the long-term (and often open-ended) commitment required to ensure that data resources will continue to be available into the future.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

RESPONSE:

Representatives from funding agencies should promote compliance by knowing the institutional landscape for data stewardship. What data centers are relevant for a particular project? Should a funded project be working with a particular data center, or using a particular data standard? Individual investigators may not have a wide enough view to know where their data might be submitted for long term preservation. Funding agencies can create relationships that may not exist yet between individual investigators and data centers by making introductions and providing financial support for researchers to prepare and submit their data to relevant data centers.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

RESPONSE:

Within the research communities, agencies can develop funding programs that explicitly invite/require researchers to make use of existing data. That is, agencies can release calls for proposals wherein money is earmarked specifically for proposals that will leverage existing data. Currently, it is much easier for researchers in any discipline to receive funding for projects that produce new data. To stimulate innovate use of existing data, data reuse must be financially supported in the same way as original data production.

Second, agencies can develop and support education programs across the disciplinary spectrum that promote "data science" as a viable career path. This can include graduate and post-doctoral fellowships for data-related research and development in ecology, sociology, atmospheric science, library science, biology, etc, as well as educational initiatives that bring researchers from different disciplines together in order to foster collaboration and cross-discipline sharing of knowledge, technologies, and research opportunities. Funding for the development of educational programs that cross the information and scientific disciplines could serve to introduce data management techniques and practices into disciplinary curricula. Data management and curation workshops for undergraduate and graduate students might also bring more such activities within disciplines where those topics are not regularly addressed.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

RESPONSE:

Within NCAR, we are promoting data citations as a way to assure that data producers are given appropriate attribution and credit for the use of their data for secondary and integrative purposes. Our data citation initiative is part of a broader movement in scientific and public policy circles. The interest in data citations is coming from many research stakeholders, including funders, policy makers, professional societies, research organizations,

and individual researchers (NAS, 2011; AGU, 2009; ESIP, 2011; Parsons, Duerr, & Minster, 2010), and is stimulated by the availability of new tools for identifying and linking to data in a web environment (Van de Sompel, et al., 2004; Bizer, 2009).

Data citations promote transparency in research by offering a direct pathway to the data so that the research can be validated or easily carried forward from a known starting point. They also raise the profile of data, that is, to make data as valued and rewarded in scientific settings as peer-reviewed publications. The benefits to scientific communities of data citations include: 1) formal citations give credit to scientists for their work in collecting and creating data, 2) formal citations will allow data center managers to track the use of data sets and gain the benefits of documenting their services and creating a foundation to design better services, and 3) formal citations will help accelerate scientific progress by tightly coupling scholary publications and data, so that two-way discovery and access are common.

In order for data citations to serve these desired roles, however, *there must balanced support for citation-linking technology, promotion of data citations within research settings, improved bibliometric measurements of data citations, and greater acceptance of data citations as an indicator of scientific impact within research organizations.*

References:

- American Geophysical Union (AGU). (2009). *AGU Position Statement: The Importance of Long-term Preservation and Accessibility of Geophysical Data.* http://www.agu.org/sci_pol/positions/geodata.shtml
- Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE Intelligent Systems* 24(5): 87-92. http://dx.doi.org/10.1109/MIS.2009.102
- Federation of Earth Science Information Partners (ESIP). (2011). *Interagency Data Stewardship/Citations/provider guidelines.* http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines
- International Polar Year (IPY). 2008. *International Polar Year 2007-2008 Data Policy.* http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf
- Mayernik, M.S., A.L. Batcheller, & C.L. Borgman. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. In *Proceedings of the 2011 iConference* (iConference '11). New York, NY: ACM (pg. 417-425). http://doi.acm.org/10.1145/1940761.1940818
- National Academy of Sciences (NAS). (2011). *Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop.* http://sites.nationalacademies.org/PGA/brdi/PGA_064019
- National Science Foundation (NSF). (2010). *Division of Earth Sciences Data Policy.* http://www.nsf.gov/geo/ear/2010EAR_data_policy_9_28_10.pdf
- Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos Transactions, AGU,* 91(34). http://dx.doi.org/10.1029/2010EO340001
- Pienta, A.M., Alter, G., & Lyle, L. (2010). The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. Presented at the *BRICK, DIME, STRIKE Workshop, The Organisation, Economics, and Policy of Scientific Research,* Turin, Italy, April 23-24, 2010. http://hdl.handle.net/2027.42/78307

- Stodden, V. (2009). The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. *Computing in Science & Engineering* 11(1): 35-40. http://dx.doi.org/10.1109/MCSE.2009.19
- Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Magazine* 10(9). http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html
- Wilbanks, J. (2008). Public domain, copyright licenses and the freedom to integrate science. *Journal of Science Communication* 7(2). http://jcom.sissa.it/archive/07/02/Jcom0702%282008%29C01/Jcom0702%282008%29C04/Jcom0702%282008%29C04.pdf
- Wynholds, L., Fearon, D.S, Borgman, C.B., & Traweek, S. (2011). When use cases are not useful: data practices, astronomy, and digital libraries. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (JCDL '11). New York, NY: ACM (pg. 383-386). http://doi.acm.org/10.1145/1998076.1998146

AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS

An International Geological Organization

Dr. Paul Weimer
President

University of Colorado
Department of Geological Sciences-CB399
2200 Colorado Avenue, Benson Building
Boulder, Colorado 80309-0399
Phone: (303) 492-3809
Fax: (303) 492-2606
E-mail: paul.weimer@colorado.edu
www.aapg.org

January 12, 2012

To: Office of Science and Technology Policy
digitaldata@ostp.gov

1. The following comments are submitted by the American Association of Petroleum Geologists, a not-for profit scientific society whose mission is to advance the science of petroleum geology. Publishing is one means of fulfilling that mission. Our publications are based on research stemming from a combination of industry, and in some cases government supported research – such as NSF, USGS, and DOE.

2. As a publisher AAPG is not in the business of running a digital research data repository, and is not viewed as such by authors. AAPG as a publisher offers an online "Data Sharing Option" for our journal, the AAPG *Bulletin*. This allows authors, at their discretion, to include backup digital data which may be pertinent to their original research or to the published paper, to be posted at an online location with a linkage back to the published paper. Very few of our authors take advantage of this option.

3. The mission of many publishers, and of AAPG, is to advance science through publishing, and other means, but not in providing a repository of raw research data. As a publisher, AAPG never requires and only occasionally receives access to any of the digital data, Policies which are developed to insure the preservation and discoverability of digital data should recognize that a researcher's raw digital data is entirely separate from the publishers publication process.

Comments to specific questions in the RFI are outlined below with the specific question outlined in **Bold.**

> **(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?**

4. For AAPG, and I believe most publishers, the intellectual property retained by us as a publisher rests with the publication itself, not the raw data which may have served to develop the publication.

Research data may be a combination of government, private, or industry sourced data. Policies which are developed to insure the preservation and discoverability of digital data should recognize that this effort is entirely separate from the publication process. Privately funded data, which may also be linked to the publication, would need to be excluded from the requirement of openly accessible archiving.

> (**3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?**

5. Policies for digital archival should include a definition of digital data to be provided. For example the access policy should outline whether: all raw gathered data; data which has been digitally processed or enhanced to some degree; only final tabulations of results, such as spreadsheets are to be archived. In the field of geology, for example, geochemical or other analyses may be conducted on rock samples. Several levels of analysis might be conducted on the raw geochemical measurements, ultimately leading to a tabulation of analyses, prior to interpretation and publication. Sample retention would likewise be a question. In short a policy should outline whether all or just some of these levels of data need to be retained.

6. Digital data within each scientific discipline, not to mention, across disciplines, are of uniquely different formats and standards. Data silos would need to be created within a repository, with each holding a distinct type and format of data. For example, within the field of geology, the data formats of raw geochemical analysis are vastly different from those of seismic measurements, or well log information, or analysis of a satellite image. These of course are vastly different from data sets in the fields of medicine, physics, or engineering. Perhaps a solution might be to require maintenance of only tabulated summary data in spreadsheet format, as an example; this implies no raw data but rather final processed information of some type.

> **(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?**

7. Policies for compliance of data stewardship should be outlined at the point of grant award, and to the individuals who have received that grant, and not to publishers or other third parties not involved in the research. This implies that a government coordinated or approved site would be available for the researchers to archive the digital data prior to or perhaps simultaneously with the publication process.

> **(13) What policies, practices, and standards are needed to support linking between publications and associated data?**

8. Scientific publishers widely utilize a DOI number (digital object identifier), which uniquely identifies that particular paper. According to the International DOI Foundation, over 55 million DOI names have been assigned by the DOI System since its inception, (http://www.doi.org). Government grants also have unique grant numbers, so this linkage would seem reasonable. A published paper with one DOI may tie into multiple government grants. Likewise one grant may result in multiple DOIs. Policies designed to link publications to data should focus on DOIs.

Sincerely,

Paul Weimer

Dr. Paul Weimer
AAPG President

MICROSOFT CORPORATION RESPONSE TO OSTP REQUEST FOR INFORMATION: PUBLIC ACCESS TO DIGITAL DATA RESULTING FROM FEDERALLY FUNDED SCIENTIFIC RESEARCH (FR DOC. 2011-28621)

SUBMITTED JANUARY 12, 2012

SUMMARY

Microsoft believes that curating, preserving, and using the digital data that result from federally funded scientific research are critical for advances in scientific discovery and for building a strong, innovative economy. We support the good work done to date by the research community and Federal agencies to define the challenges and outline possible solutions. In particular, we cite the report of the National Science Foundation's Advisory Committee for Cyberinfrastructure's Task Force on Data and Visualization (http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf) and the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (http://brtf.sdsc.edu/). Experts from Microsoft participated in the drafting of these reports, and we remain committed to their conclusions. We also agree with many of the challenges described and conclusions reached in the National Science Board's draft Data Policies Report released on January 5, 2012.

As the National Science and Technology Council's (NSTC's) Interagency Working Group on Digital Data proceeds with deliberations to inform Federal policies concerning access to digital data resulting from federally funded scientific research, Microsoft would like to draw your attention to two areas—**Economic Models** and **Software Tools and Online Services**—and offer three recommendations in these areas.

Economic Models. The nation must create an environment in which innovation can occur around the critical elements that enable data sharing, retention and use, and costs can be distributed among the various groups that receive benefits from the data and associated discoveries. Challenges include the long-term nature of the problem (costs and activities in this space will extend over timelines greater than a typical research grant), and the need to make choices around what data should be preserved and shared. A wide variety of groups will create and use the data and they all must share in the costs and decisions about how data is preserved and shared. These participants include scientific communities and research groups of various sizes, universities, Federal laboratories, commercial service providers, and both Federal and state governments. They also include the consumers of the data, who may be outside of the research community, but who will have a stake in defining what data is of value and a responsibility to contribute to costs.

The economic models deployed around the data ecosystem will vary by discipline, but approaches should incent sharing and should provide support for the individuals and organizations that create, make available, and maintain high value scientific data collections. Exploration of different business models is critical, and the full variety of information technology (IT) infrastructures available for the various stages of data preservation and use should be exploited.

Recommendation 1 – Assessment of Economic Models around Data Retention and Sharing: Assessment projects should be undertaken to evaluate the *economics* associated with supporting and facilitating the long-term hosting and use of data. These projects should analyze potential economic models, including factors such as cost effectiveness, opportunities and risks for businesses and research institutions, and potential for value-added software tools and services. The results of these analyses should provide options for policies and programs through which the Federal Government might successfully foster a stable long-term ecosystem of service and data providers and consumers.

Recommendation 2 – Broaden the Range of Supported Information Technology Infrastructures around Data-Related Activities: Research communities and institutions, as well as information technology service providers, would be better able to explore different models for data sharing if there was clarification of Federal policies for support of information technology infrastructure, including computing services such as cloud, around data-related activities during and after research grants. In particular, it is important that Federal policies focus on the desired outcome (e.g. data sharing to advance science) and enable a variety of approaches to accessing the necessary IT hardware and software capabilities, whether through a purchase as part of a research grant, as an ongoing service from a commercial provider, or as an institutional or community resource.

Software Tools and Online Services. Simple, easy-to-use software tools and online services for data archiving, dissemination, discovery, and analysis are critical to maximizing the ability of researchers, entrepreneurs, companies and others to extract understanding and value from data. Also, metadata standards are key to facilitating the development and use of broadly applicable tools. Such tools are particularly critical for science conducted in increasingly multi-disciplinary and international environments. Some scientific disciplines have made progress in this regard, but more attention to this problem is needed.

Recommendation 3 – Support for the Development and Sharing of Software Tools and Online Services: Financial support is needed for the purchase, development and deployment of software tools and online services for data archiving, dissemination, discovery, and analysis. This support may take the form of Federal or foundation grants to universities or domain research collaborations, and such investments should include tools and services that can be shared across and customized for multiple research communities.

This issue relates back to Recommendation 1, as the role of tools and services is critical in evaluating potential economic models for data sharing. For example, a tool or service infrastructure that enhances the value of the data may allow the provider to monetize access to the data at a level sufficient to cover the investment made in creating or maintaining the data archive.

Additional Issues. In addition to the specific issues and recommendations discussed above, there are a number of other important policies that could contribute to supporting the effective long-term stewardship of publicly-funded research data. Examples include incentives to change the scientific culture around hoarding of data; development and implementation of domain-specific metadata, standards, formats, and protocols; rules around timing for and constraints to access to data by other researchers; ways to allow citations to data and credit to data creators and sharers; clarification of research agency expectations around domain-specific retention policies; and assigning responsibilities for long-term data curation. Although we do not discuss these topics in this document, we support the analysis and comments on them in the existing reports referenced above.

ECONOMIC MODELS

To maximize the opportunity for scientific discovery and innovation, it is favorable for data to be accessible and usable by a range of stakeholders including academic researchers, industry laboratories, members of the general public and international research collaborations. Sharing data advances scientific discovery, but also has positive economic impact, informs policy formulation, and provides educational and other societal benefits.

The growing amount of data created and used in scientific research is well documented, and the value and impact of making this data available has been widely discussed.[1] However, where the data (including metadata, images,

[1] For examples of how using new computing capabilities to explore massive datasets will advance multiple scientific domains, see *The Fourth Paradigm: Data-Intensive Scientific Discovery* (http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx).

and software tools) will be stored, how the use of this information will be facilitated, and who will pay to develop, deploy, maintain, and improve the relevant resources and capabilities remain a fundamental challenge. It is unlikely that a single model will work across all disciplines, research team sizes, institutions, and countries.

All Data are Not Equal

It is not affordable, or even possible, to save all data from all scientific investigations for all time. It is important to understand when data has re-use potential and when it does not, when it is easier or cheaper to recreate data than to store it, and when data that has been saved is no longer worth keeping. A variety of factors will inform such decisions, including how the costs of storage, access, and use are paid. The choices and the associated costs are not obvious – cheaper storage media, with limited bandwidth access might be an option for infrequently used information, while more intensely used information can need a variety of additional services, such as low latency storage system, replicated storage at multiple sites, high bandwidth connections, and tools and computing power to search, analyze, and access the data.

Another key variable in data retention and sharing discussions is the source of the data. Not only will different scientific fields have different cultures, priorities, and expectations around data management, but different research models also will require different approaches. For example, for data generated by single investigators or small groups, the processes and infrastructure for sharing may be a significant burden on the researcher. In this case policies and tools could focus on methods that allow the research to move the data into an existing curated collection that has a well sustained business model. For larger, multi-disciplinary or multi-institution collaborations, some level of shared data storage, access, and analysis is likely to exist as part of the collaborative process and sustainability rests on the economics of the discipline. For example, high energy physics and astronomy have models in which there is long-range government funding. In other areas, the users of the data may come from outside the scientific community generating the data – from other research fields or from commercial entities. Biology and chemistry have both nationally supported archives as well as the potential for public-private partnerships. In these cases, decisions about what data to keep, how to disseminate data and associated services, and what parties bear which costs requires negotiations beyond the community of the original data creators.

International partnerships are a special case. Different countries have different cultural norms, different policy mandates, and different economic models around the responsibilities of governments, universities, and other organizations. Flexibility will be necessary to craft solutions that balance different requirements, and up-front planning for the systems, costs, and policies of data retention and access will be critical.

Enabling Different Models – Facilitating Public-Private Partnerships

In addition to being cognizant of the variety of services and support that could be associated with a given data collection, it is important to recognize the goals, resources and priorities of the individuals, communities, or institutions that are either producing the data, using the data, re-using the data, storing the data, providing tools around the data, or paying for these various activities. Consequently it is favorable to enable economic models that have the flexibility to allow different groups to provide different services to different audiences at different times and costs.

There are a number of models that provide support at different stages of the data lifecycle (see below), but allowing market forces to operate can be of help in understanding how to preserve data and what kind of access patterns need to be supported. In particular, flexible pricing models can be important for gathering quantitative information about which data sets are being used, by whom, and how. Gathering such information facilitates informed evolution of choices about retention and pricing and could be used, in concert with evaluation of scientific needs and directions, to determine when data should be moved to cold storage or expunged.

A variety of approaches to providing and maintaining data ecosystems can be imagined or observed. For example, Federal agencies can create and operate systems for storage and access, or pay third parties directly for those services. Research organizations, including universities or scientific societies, can provide organization-wide access to services funded by fees or supported out of indirect costs. Individual researchers can fund data dissemination from a specific research grant, and can pay for access to data or tools on an ad hoc basis. Examples of approaches already deployed in various fields include:

- The University of Michigan Inter-university consortium of political and social research, asks customer institutions, such as universities and research laboratories to subscribe on behalf of their researchers.
- The data from the experiments on the Large Hadron Collider is managed by an international, multi-tier distribution system which is funded as part of the project and is provided for free to the participating physicists.
- For a number of geosciences and life sciences data sets, there is already a marketplace for providing access to data based on modest subscription fees which cover the cost of maintaining high quality tools to search, analyze and access the data as well as storage costs for the data. Examples include Datamarket.com and Windows Azure DataMarket, LifeSpan BioSciences, Inc.

No specific model is correct for all situations, but the most important factor in ensuring successful data impact is enabling various organizations to bring their skills and resources to different elements of the data lifecycle. Public-private partnerships will be an important component. For example, in some situations, federal agencies may directly fund the research that generates data, but later only indirectly fund the storage and access to that data by allowing other researchers to pay fees for access and tools developed and maintained by the original researchers, other researchers, a scientific society, or a for-profit entity. An additional market value is created when organizations can develop and deploy value-added services on top of free data, or data from other organizations.[2]

The Role of Cloud Computing and Storage

Cloud technologies, which are being developed and deployed for a variety of business, government, and consumer applications, are relevant to the data challenges in a variety of ways.

Move the Analysis to the Data: Today, a scientist can store or download modest amounts of data to a local computer for limited analysis and study, but increasingly the size of the data sets or the computing power required for analysis will make this inefficient or impossible. It will become necessary to move the analyses to where the data is. Using the large data centers that have been built to support massively parallel analysis of resident data, scientists can conduct research on petascale data archives in ways that are not possible on local facilities.

Environment for Collaboration: Cloud computing services potentially provide an information technology environment that facilitates both collaboration and effective data sharing. This may be particularly valuable for multidisciplinary and/or multi-organization collaborations. Cloud computing may also be a platform to support the ecosystem of data sharing and use – different parties can come together in the cloud to provide different elements of the tools and services needed (from the data, to the storage, to the applications and tools, to the computational power). Microsoft maintains a cloud based marketplace for data access. While some of the data is subscription based, there is a great deal of public data that is provided free of charge. Other large cloud providers, such as Amazon and Google, offer similar services.

Build on Other Investments: The scale of cloud deployments, and the rapidly evolving ecosystem around cloud applications for business, government, and consumers, as well as science, are likely to facilitate the evolving

[2] The European Union emphasized the potential economic value of commercial re-use of data in announcing its proposed Open Data Strategy in December (http://www.zdnet.co.uk/news/regulation/2011/12/12/reuse-of-public-data-to-get-easier-under-new-eu-rules-40094628/).

development and deployment of technology that can potentially reduce the costs of data storage, access, and analysis for research.

Recommendations (Economic Models)

An important step in creating a vibrant environment for data sharing is facilitating people and organizations' ability to experiment with different approaches for different research communities. In particular, it is critical to enable *an ecosystem in which different actors can contribute relevant materials, tools, and products for the different elements of the data lifecycle*.

Recommendation 1: Assessment of Economic Models around Data Retention and Sharing
To encourage the development of an ecosystem of services supporting data retention, access, and use, Federal agencies should support targeted economic assessment projects. The goal of the projects would be to explore the economic viability of a variety of support models for the longer-term hosting and use of scientific data including both academic use and any potential commercial exploitation that could be used to supplement (or completely pay for) the costs the data access and retention for academic research. The results of these analyses should provide options for policies and programs through which the Federal Government might successfully foster a stable long-term ecosystem of service and data providers and consumers.

Types of Projects: The assessments should explore a variety of disciplines and consider the roles and needs of single or small group investigators, multi-disciplinary/multi-institution or public-private collaborations, and data collected for scientific and operational consumption across a variety of sectors. For example, one assessment might explore the use of the cloud by a multi-institution collaboration for its own data analysis in the short-term as well as dissemination of data and associated tools to the larger community in the long-term. Another might look at the development and deployment of tools that cost-effectively allow single investigators to manage the data lifecycle and workflow from creation to archival. Another might evaluate the business models for how weather data is used by climate researchers, government operations, and commercial entities.

Questions to Be Explored:
- how users as well as the disseminators of data are currently supporting the associated costs of data management – to give a clear understanding of the current cost 'baseline';
- tracking of who is using shared data and for what purposes (including access patterns and derived value), as well as how the users discovered the data – to understand current practices and successful collaborative models;
- existing or potential inflexion points in data management costs due to economies of scale (e.g. data or access volumes which suggest a more cost-effective transition to cloud-based service providers) – to understand the criteria for when/where different types of service are applicable for different volumes of data or where access volumes might experience different service levels/support costs;
- cost-effectiveness of different service models and comparison between cost structures across the different phases of the data management life-cycle – to understand whether/where there are particular models which work for specific parts of the data access/use/retrieval lifecycle and how these differ between scientific disciplines;
- potential for value-added software tools and/or services – to understand what scope there is for software or service infrastructure might be applicable/available to different types of research data and, specifically, whether there would be opportunities for commercial exploitation of data which could supplement or completely cover the costs of data retention and access by the research community;

The gathered data and assessments could inform Federal decisions on what sorts of programs would enable data ecosystems. In addition, the domain-specific information could help inform research groups considering long-term data management plans of various relevant business and information technology models.

Recommendation 2: Broaden the Range of Supported Information Technology Infrastructures around Data-Related Activities.

The fiscal role of the Federal government in enabling access to data can take many forms, including direct and indirect support of various elements of the data sharing ecosystems. Research communities and institutions, as well as information technology service providers, would be better able to explore different models for data sharing if there was clarification of Federal policies for support of information technology infrastructure, around data-related activities during and after research grants.

In particular, it is important that Federal policies focus on the desired outcome (e.g. data sharing to advance science) and be flexible about which IT infrastructures are used to achieve the outcome. Conducting research, creating data, preserving, sharing and reusing that data can require a wide array of IT hardware and software capabilities, and these capabilities can be achieved in a variety of ways—purchased through an individual research grant, provided by a university as an institutional resource, obtained through a community resource shared across a scientific domain, or acquired on an as needed basis from a commercial service provider. When architecting any new Federal policies, it would be advantageous to avoid discriminating against any particular approach and/or presuming a favored solution.

For example, in determining what IT infrastructure is allowed to be used in the conduct of a research grant, a selection should take into account not only the resources necessary to carry out the specific research, but also whether the choice will smooth the transition for data to be shared. This could include support within the grant for usage (and fees) for community resources, or payment for commercial storage, computing, or software services.[3]

Software Tools and Online Services

The development and deployment of software tools and services to enable sharing, discovery, and analysis of data is key to realizing the ultimate goal – increased scientific and societal impact of data. Unfortunately, beyond a few tools used within a few narrow scientific subdomains, there are no standard software packages that scientists can use today. Metadata standards are equally scarce. Researchers, government agencies, companies, and others have a role to play in creating and supporting these tools. The deployment of on-line services that provide these essential capabilities increases the value of the data and creates a possible market and business model to sustain it.

Software tools are critical at every stage of the data sharing lifecycle. From the start, tools that simplify the steps of preparing data and associated metadata for sharing, perhaps integrated into the data generation and capture process, facilitate data preservation, annotation, and sharing. If the data workflow is well managed and cataloged from creation, then archival requires much less effort.

Increasingly, the value of data extends beyond traditional disciplinary boundaries, and ensuring data access and the ability to correlate data from multiple disciplines requires appropriate metadata, protocols, and interoperability standards. Tools for analysis that are deployed in concert and coordination with specific data sets are particularly valuable in supplementing metadata and other annotation of data sets. Specific technical issues that must be addressed include robust, long-term secure digital storage, reliable techniques for predicting storage media aging, mining large-scale data collections to provide useful information, and visualization tools.

---

[3] In particular, purchase of IT equipment and purchase of IT services may be treated differently under Federal grant regulations (OMB Circular A-21) in terms of whether they will be subject to indirect costs, providing a potential fiscal disincentive to utilize services. However, in certain circumstances the use of services, such as cloud computing, may be a more effective approach to meeting goals for timeliness of access to resources, flexibility in scaling storage or computing power up and down, enabling of long-term data retention, etc.

In addition to coupling data with analysis tools, great value can also be obtained by connecting data with research publications. As these publications become digital artifacts, it should become easier to trace the provenance of a research result back to the supporting data collections and analysis tools. This capability will facilitate repeatability and reproducibility in scientific experiments and transparency around the context when data and analyses are being used in policy discussions.

Recommendation 3: Support for the Development and Sharing of Software Tools and Online Services

Software tools are essential for facilitating data archiving, dissemination, discovery, and analysis. Federal programs and policies should facilitate access to and use of such tools and online services. This should include financial support for the purchase, development and deployment of software tools and associated online services, which may take the form of Federal or foundation grants to universities or domain research collaborations. It also should include policies designed to minimize duplication of existing resources and encourage the sharing and reuse of tools and services.

- *Use of Existing Capabilities*: Federal research programs could give priority to research proposals that describe how they will make use of commercially available data services and software tools or community developed and supported tools in cases where these can cost-effectively provide storage, analysis capabilities, visualization, pre/post processing and/or data handling/manipulation capabilities.
- *Focus on Sharing*: Priority for Federal investments could focus on tools and services that can be shared across and customized for multiple research communities and domains of science. Funding tool development separately from domain science would encourage a focus on capabilities relevant to multiple fields.
- *Stable Long-Term Deployment and Use*: Proposals for tool development support should include consideration of the potential deployment models, including the opportunity for a sustainable economic model for the maintenance of the tools. Emphasis should also be placed up front on what metadata will be required for effective use of the tools.

Examples of the types of tools and services that are important for researchers in multiple fields to have access to include tools that simplify the data lifecycle management including the steps of preparing data and associated metadata for sharing, easy-to-use search, visualization, and analysis tools, and tools for that allow limited access and analysis of data to maintain privacy preservation[4] or intellectual property constraints.

[4] A potentially significant barrier to data sharing in certain research areas, such as biomedicine and some social sciences, is concerns about maintaining compliance with related privacy and data integrity rules. Examples of the challenges include the potential for de-anonymization of data when multiple related data sets are shared, or the need to comply with different country-specific regulations. Tools to constrain users to acceptable analyses, or methods to build data sharing approaches around providing analytical results rather than raw data, are needed to facilitate the realization of the economic and societal benefits of data sharing and reuse.

Matthew Cockerill
Managing Director BioMed Central
236, Gray's Inn Road
London
WC1X 8HB
UK

tel +44 20 3192 2000
www.biomedcentral.com
info@biomedcentral.com

Office of Science and Technology Policy
digitaldata@ostp.gov

11th January, 2012

Response to FR DOC 2011-28621

Dear Sir or Madam,

On behalf of BioMed Central Ltd, I am writing to respond to the OSTP's RFI on *Public Access to Digital Data Resulting From Federally Funded Scientific Research*.

BioMed Central is a leading open access publisher. Since its launch in 2000, BioMed Central has demonstrated that commercially viable business models exist which allow scientific publishers to make the peer-reviewed research articles they publish immediately and freely available online in their official form, with costs typically covered via a publication fee. BioMed Central is a founder member of the Open Access Scholarly Publishers Association (http://www.oaspa.org/) and since 2008, has been part of Springer Science+Business Media, the world's second largest publisher of scientific, technical and medical journals (STM).

One of BioMed Central's key objectives as a publisher has been to help researchers to share not only the final results of their work, in the form of a research article, but also the data which underlie that work. To that end, BioMed Central has taken an active role in many data sharing initiatives, and has created a Publishing Open Data Working Group involving funders, researchers and publishers to help identify best practices to encourage data sharing, and to identify steps publishers can take to facilitate such sharing. See http://bit.ly/n4U348 for additional information.

Responses to the specific questions in the RFI are given below:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from Federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

There are already examples of scientific communities where policies requiring open data sharing have been widely adopted with great success, and which could serve as useful models for a wider range of data resulting from Federally-funded research, and which demonstrate the economic benefits of such data sharing. Specifically, since the early days of the human genome project, general consensus in the genomics community has led to a policy of immediate open availability of all genetic sequence data from publicly-funded research, with little or no embargo period. Community norms give data creators priority in terms of publications and credit resulting from data analysis, but with a relatively short period of exclusivity (6-12 months) [Contreras 2011: http://www.sciencemag.org/content/329/5990/393.short].

Economics and astronomy are other examples of fields where community norms exist requiring data sharing after some limited time period of exclusivity. The appropriate timeframe for such an embargo period is likely to vary by field, but it is important to realize that some fields can learn from one another, and that in fields in which data-sharing has been slow to take off data-sharing may need to be incentivized, and obstacles to such sharing eliminated.

In the field of clinical trials, there is strong public interest in ensuring that results and data are made freely available as soon as possible following the completion of the trial. [Gøtzsche 2011, http://www.trialsjournal.com/content/12/1/249]. Not only are positive results of vital interest to patients, but negative results are also of vital importance. If they are not reported, then the resulting selective reporting of clinical trial results can lead to a systematic bias in the scientific literature, undermining the validity of the evidence-base for the effectiveness of treatments, and ultimately leading to detrimental effects on the quality of healthcare.

Since 2007, the deposit of summary results for clinical trials has been Federally mandated (http://clinicaltrials.gov/ct2/info/results). However, while such summary result sharing is beneficial, a great deal of additional data is captured as part of the clinical trial which could have great value if shared. For example, sharing such raw data can facilitate more accurate meta-analysis of the results from multiple trials. Great care is needed in designing ethical data sharing policies for clinical trial data, however (Hrynaszkiewicz et al. 2010. http://www.trialsjournal.com/content/11/1/9). Informed consent is vital, as is a suitable mechanism to protect the privacy of individual patients. For complex datasets, watertight anonymization may be difficult to guarantee, and for this reason full public access to all raw data may not be possible. In such cases, public access may be given to a limited subset of

data, while the full dataset might be maintained in a suitably protected repository, with access provided only for specifically approved research uses.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from Federally-funded scientific research?

Intellectual property interests of publishers are likely to be less of an obstacle when developing policies on access to research data than when considering policies on access to published research articles. The ALPSP and STM publishing associations issued a joint statement on access to data in 2006 [http://bit.ly/wrPwph] recommending that: *"raw research data should in general be made freely available. When data sets are submitted along with a paper for consideration in a scholarly journal, the publisher should not claim intellectual property rights in those data sets, and best practice would be to encourage or even require that the underlying research data be publicly posted for free access."*

To encourage sharing of data from the private research sector, it may be beneficial to identify types of dataset which relate to "pre-competitive" research work, and which companies may agree to share to create an "information commons" to facilitate new knowledge discovery. Sage Bionetworks [http://sagebase.org/] is one example of such an information commons, formed as a non-profit spin-off from Merck.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

By developing policies in conjunction with the scientific communities the policies are intended to serve, while also sharing experience and best practices between different domains. For example, in ecology and evolutionary biology, researchers in the community have worked with a consortium of different journals in the field to successfully implement a joint data archiving policy (JDAP), which requires the data supporting peer-reviewed publications to be publicly available [Whitlock et al., 2010: http://dx.doi.org/10.1086/650340].

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from Federally funded research?

It is not economically feasible to indefinitely preserve all research data and so data archiving policies need to take account of the likely long term value of data when determining what should be permanently archived. For example, because of the huge amount of genomic data being collected, much of it may only be kept for the length of project, with only the most relevant parts being preserved long term.

To make data part of the permanent citable scientific record, we need confidence that it will remain available long term, so ensuring that robust and sustainable models are in place to achieve this should be a vital part of any Federal data archiving and dissemination policy.

DataCite (www.datacite.org) has helped with one aspect of managing digital permanence for datasets, by associating digital object identifiers (DOIs) with datasets stored in digital repositories, so that even if the dataset moves to a new location, the DOI can still be used to locate it.

Stability of funding for data repositories is the other major concern, when it comes to delivering digital permanence. Currently many data repositories survive on short term project funding, which creates an increased risk that data will be lost.

It may be that dataset archives will need to adopt some of the same approaches to funding used by research journals to achieve long terms self-sufficiency, perhaps by charging a fee for deposits. One possibility is that such a data deposit fee could form part of the fee paid by the author from their research funds, for publication in an open access journal.

The Dryad repository (http://datadryad.org/) – currently publicly funded – has a long-term sustainability plan, which includes deposit fees for data sets associated with peer-reviewed publications. Funders should consider providing explicit, ring-fenced, funding as part of grants, to cover data archiving and data publication costs, as many already do for open access publication fees.

Commercial services offering data archiving exists, such as LabArchives (http://www.labarchives.com/) and Amazon, with usage-based subscription models, and which may have a role to play. While they may not be able to guarantee long-term preservation, they are well placed to help the scientific community by making archived data conveniently available to researchers 'in the cloud'.

International collaborations between data archives to mirror each others' content provide one approach which seems likely to increase the likelihood that content will be preserved long term, especially if the various archives have independent sources of funding.

It is important to recognize that Federal policy could encourage and even mandate data sharing without needing to provide all infrastructure for such data-sharing centrally. In developing data-sharing policies, Federal agencies should also consider distributed options.

A 2008 report by John Houghton and colleagues, for the Joint Information Services Committee, concluded that data archiving offers an excellent return on investment for funders [http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf]. However, in order to motivate individual researchers to comply with data sharing policies, specific,

relevant case studies – success stories – for data publishing should be catalogued to demonstrate the benefits of data sharing. Examples include secondary use of clinical trial data for a future systematic review, or identification of adverse effects of a medication.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Incentives (including penalties for non-compliance) may be needed to ensure widespread compliance with data sharing policies set by funders, because many researchers may be understandably cautious about sharing their hard won results with others, whatever the wider benefits to society, and all the stakeholders listed can play a role in ensuring such incentives are in place. Making data sharing a condition of publication has proven to be viable in the case of DNA sequences, protein structures and clinical trial registration, and the Dryad example shows how such policies can be successfully introduced in new fields, if a critical number of journals can be persuaded to participate. Journals and publishers can also support open sharing of data by enabling citation of data and providing journals and publication formats for publishing and describing published data sets. New measures of research impact, which go beyond traditional Impact Factor measures, are evolving (http://total-impact.org/). Also, existing tools need to be identified, or new tools developed, to enable efficient sharing and management of data. At Oxford University, the DataFlow project provides an open source data management infrastructure (http://www.dataflow.ox.ac.uk/), which aims to make it easier for research groups to manage data. LabArchives also provides online lab notebooks offering the ability to publish data and assign DOIs.

Many funders do already require grantees to create a data management plan. Unfortunately, in practice the existence of such a plan has done little to encourage the real-world sharing of data. Anecdotal evidence suggests that requests sent to labs requesting access to data under the terms of data management plan are often ignored, or rejected on spurious grounds. Therefore we would recommend that funders do not place too much reliance on such plans, but rather, consider explicit mandates relating to data deposit.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?
See response to question (4).

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Ultimately, economic incentives/sanctions could be applied by agencies, but hopefully this would not prove necessary.

The examples of clinical trial registration, and DNA sequence/protein structure databank deposition demonstrate that compliance with a specific data sharing/publishing policy can potentially be easily verified by publishers, by requiring the relevant accession number or other permanent identifier to be provided by the author, as a condition of publication. With appropriate cross-publisher support, this approach could be applied to many other types of data.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Agencies should ensure data are available in formats and under licensing terms which facilitate integration and re-use. Offering convenient Application Programming Interfaces (APIs) for access to data will also help to stimulate use (See for example: http://www.guardian.co.uk/open-platform). The use of open formats should be encouraged, though support for widely adopted proprietary file formats is also often a pragmatic necessity. Data should be available under licensing terms which remove any concerns about legal barriers to data integration and reuse (see below). Value-added and enhanced products and services can be built on open content, and this can drive the discovery of new knowledge.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

To eliminate potential legal impediments to integration and re-use of data and to help enable long-term interoperability of data an appropriate license or waiver specific to data should be applied. There are a number of licenses for open data, of which the Creative Commons CC0 license is perhaps the most widely recognised. Under CC0, an author waives all of his or her rights to the work worldwide under copyright law and all related or neighboring legal rights he or she had in the work, to the extent allowable by law.

CC0 overcomes the challenge that the CC-BY Attribution license, widely used by open access publishers including BioMed Central, is not always suitable for data reuse, because a derivative built on data may take work from many thousands of sources, making the attribution requirement extremely burdensome and often simply not feasible.

So, in the case of data reuse, rather than relying on copyright law to ensure credit is given, it seems more appropriate to rely on the established academic cultural norms as to when attribution (citation) of another scientist's work is appropriate. See: http://pantonprinciples.org/faq/ In most cases attribution will be required to ensure reliability and validity of the secondary work. For example, a systematic review of data from several clinical trials would not be publishable if it did not cite and attribute its data sources.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Data standards are important for ensuring the interoperability of data between different research groups and platforms, and for enabling data to be reused efficiently. There are numerous digital standards for scientific data,which are being catalogued by the BioSharing group at Oxford UK (http://biosharing.org/?q=standards) in partnership with the journal *BMC Research Notes* (http://www.biomedcentral.com/bmcresnotes/).

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies might do this by becoming involved in international working groups on data sharing and standardization. For example, we would welcome NIH participation in the Publishing Open Data Working Group led by BioMed Central which already includes representatives of several major European funders.
Collaboration with agencies in other countries on data archiving/mirroring can also be a productive approach, as the National Council for Biotechnology Information has demonstrated through its international mirroring/data exchange agreements covering the Genbank DNA sequence database, and the PubMed Central open access literature archive.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?
To reliably hyperlink between datasets and the associated peer-reviewed journal publications, the associated dataset must be permanently and persistently available, either alongside the journal article as an associated file, or preferably in an appropriate data repository. The issuing of persistent identifiers, such as DOIs, can help to ensure that links to data can remain functional long term. Citation of datasets, following the standards developed by DataCite, should be encouraged or required by journals and publishers. Space issues are sometimes used by journals to justify stringent restrictions on the number of citations that may be included in an article, which may discourage data citation, but in an

online environment such limitations are unnecessary and should generally be avoided. Explicitly citing datasets is an important mechanism to ensure that visible academic credit is gained for data publication and sharing, removing one commonly-perceived barrier to sharing and publishing of data. Journals and publishers should also provide additional tools for consistent linking between publications and the supporting datasets. For example, a number of BioMed Central journals now require the inclusion of an 'Availability of supporting data' section which clearly points readers to the location from which they can obtain the raw datasets supporting an article.
See: http://www.biomedcentral.com/about/supportingdata and http://bit.ly/zbsPRp

Journals which publish data papers – where the primary purpose of a publication is to publish a description of a dataset, rather than methods and results – are also an important means of earning academic credit for data sharing. *BMC Research Notes* is one such journal: http://www.biomedcentral.com/bmcresnotes/authors/instructions/datanote

Finally, sharing data online using appropriate standard formats and technologies to make it machine readable and semantically meaningful can help to achieve Tim Berners-Lee's vision of Linked Data on the web. See: http://en.wikipedia.org/wiki/Linked_data

Yours sincerely,

Matthew Cockerill

Managing Director,
BioMed Central Ltd

Association of College & Research Libraries
50 E. Huron St. Chicago, IL 60611
800-545-2433, ext. 2523
acrl@ala.org, http://www.acrl.org

TO: Office of Science and Technology Policy digitaldata@ostp.gov
FROM: Association of College and Research Libraries
RE: Recommendations on Public Access to Digital Data Resulting from Federally Funded Scientific Research
DATE: Thursday, January 12, 2012

The Association of College and Research Libraries (ACRL) writes in response to the request for information issued November 10, 2011, by the Office of Science and Technology Policy (OSTP) regarding recommendations on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research.

ACRL, a division of the American Library Association, is a nonprofit professional organization representing more than 12,000 academic and research librarians and interested individuals. ACRL is the only individual membership organization in North America that develops programs, products and services to meet the unique needs of academic and research librarians. Our initiatives enhance the ability of academic library and information professionals to serve the information needs of the higher education community and to improving learning, teaching and research. ACRL publishes scholarly, peer-reviewed journals in the field of library and information science.

ACRL appreciates the opportunity to comment on increasing public access to digital data that result from federally funded scientific research. Many of our individual members and their libraries will also submit detailed comments to OSTP. ACRL has long believed that ensuring public access to the fruits of federally funded research is a logical, feasible and widely beneficial goal. We have endorsed "The Federal Research Public Access Act of 2009" (S. 1373) noting, "It reflects ALA policy regarding access to federal government information by providing for the long-term preservation of, and no-fee public access to, government-sponsored, taxpayer funded, published research findings."

ACRL offers comments on the first nine questions posed in the request for information, as we would most like to express our position regarding policy for preservation, discoverability and access. We are refraining from addressing the last four questions on standards for interoperability, reuse and repurposing, which seek specific examples of good data management. Our comments to the specific questions follow:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

- The most effective federal policy to encourage public access to, and preservation of, research data would be to require that data generated in the course of federally-funded research be deposited into publically accessible repositories. The National Science Foundation (NSF) requirement of a data management plan is a laudable step toward awareness of the need to manage data, but a mandate will be required to create the critical mass of available data that will support rapid scientific innovation and encourage the commercial reuse of data that can underlie economic growth.

- The NSF approach, which has been to set an expectation but not to require a specific method of implementation, is the right one. Because of the variety of approaches and types of data across different disciplines, flexibility in compliance is called for, even within the context of a mandate.
- This flexibility can help the scientific community come to view data preservation and sharing as an issue of principle, necessary for good research and scientific accountability, rather than as merely a burdensome compliance issue.
- A policy mandating data deposit will need to be accompanied by the development of standards and services that make data sharing economically feasible and data reuse as accessible as possible. Incentives are as important as requirements if the goal is to make usable data available for scientific verification and commercial reuse. By creating systems that are as simple, standardized and open to reuse as possible, the maximum potential of economic growth will be achieved. A useful metric for public access to data is whether someone, or some computer, can discover, access, interpret and use the data without having to contact the original data producer; such access is both economically beneficial and less burdensome to the data producers.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

- It is important to recognize that raw data is not subject to copyright under U.S. law, and that rights even to protectable collections of data usually remain with the data producer, rather than being subject to transfer to publishers. So the rights issues are not as complex for data as they may be for publications.
- The basic premise of federal policy should be that more openness is better, and restrictions should be applied only when genuinely necessary, for example, when the data makes it possible to identify a particular person involved in the research study. Basically the default, which is currently that data is manage locally (if at all) and idiosyncratically, should be changed to openness and standardization.
- Theykey to convincing data producers, who are also the holders of whatever IP rights exist, to participate is to provide easy roads to compliance and incentives, usually in the form of norms and expectations within their disciplinary communities, to comply.
- The federal government could assist in making data preservation and sharing as seamless as possible by working with publishers and other stakeholders to reduce the burden of handling "supplemental materials" and make it easier to integrate data into their publishing platforms and access systems.
- There is a reasonable argument for embargoes, in some cases, based on the unique effort exerted by the data producers or original scientific research team. Although effort alone cannot justify copyright protection (based on the Supreme Court's 1991 decision in *Feist Publications v. Rural Telephone Service Co.*), the need to protect data for some short period of time while the team or lab completes its own analysis could be respected by allowing a fixed-term period of exclusive access. Such an arrangement, however, does not preclude the deposit of the data into a certified repository even during that embargo period, particularly so that archiving activities can begin.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

- Since there is ample evidence that different scientific disciplines present a variety of requirements for the management of data, the policies and principles that underlie a data sharing mandate should be relatively general, while the practices within each discipline will need to be specific. Data sharing policies should be viewed as flexible requirements that remain open to modification as problems arise or best practices emerge from within specific communities of scientific practice.
- Some baseline conditions or requirements, especially related to archiving and preservation, can be applied across the board. This is a vital place to begin, since many scientific disciplines have focused on access or discovery rather than preservation, yet the latter is key to fostering efficiency and innovative reuse.
- In some disciplines, a funder requirement will serve as a first step toward creating awareness of the fundamental need for data management. We have seen this take place among working scientists as awareness of the NSF data management plan requirement has spread, and further mandates will facilitate this awareness.
- Funding agencies should be willing to provide funding for data management expertise that is available locally at researchers' institutions, and/or through disciplinary repository services (such as the DRYAD repository at the National Evolutionary Synthesis Center). Such support will assist researchers in applying data management approaches that are appropriate to their specific disciplines.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

- The particularized and ad hoc nature of so many approaches to research data until now makes it difficult to assess relative costs and benefits for different disciplines. It may be most useful to think in terms of baseline services that should be supported across all disciplines (i.e. archiving) and more particularized secondary services (such as specialized query capabilities). Agencies might consider the relative emphasis that is appropriate in the area of research that that agency funds, and what areas are appropriate for local institutions to assume responsibility for. Thus an agency might provide seed funding to institutions for preservation, but recognize the need for ongoing funding to a scientific community to develop secondary services.
- A potential technique to establish a baseline cost would be to set an allowable cost for data management for funding requests, then analyze, after several rounds, what approaches have been applied and how effective they have been based on metrics such as use statistics, the verifiable integrity of the data over time and third-party costs to discover, retrieve and use the data. If funding is provided to disciplinary repositories, reports based on these metrics should be required.
- The benefits of shared data will also be difficult to measure, but they are nonetheless real. Accountability and the ability to verify scientific results are vital, but hard to quantify. Other benefits, such as the support provided for reuse by different teams of researchers or by commercial enterprises, will be easier to track. The opportunities for innovation and commercial exploitation of shared data will be evidenced by increased growth within a sector of the economy.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

- It is important to keep researchers focused on research; this is vital if a data sharing requirement is going to support innovation and growth and not hinder it. The stakeholders named thus have the important role of providing the services, standards, best practices and infrastructure that make data sharing simple and efficient. Insofar as agencies can provide funding and other incentives to support those functions, they will contribute to the implementation of data management plans.
- The best approach is to build on existing infrastructures and practices, learning from what works well while being sensitive to disciplinary differences.
- While successful practices should be the model for policy implementation, it is important that success be demonstrated and not merely asserted. Each agency, as part of its data sharing mandate, should identify metrics that are important within that field by which the success of a plan or services can be measured. Those metrics will evolve over time, but with a clearly articulated set of requirements it will be possible to identify how various stakeholders can contribute to the successful implementation of data management plans.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

- Perhaps the most important step would be to acknowledge and communicate that the real costs of preserving and sharing digital data are indeed legitimate and important costs of the overall research enterprise.
- It is important to recognize that not all costs associated with good data management will be directly attributable to specific projects. As data management expectations become more widespread and routine, an increasing proportion of the costs will need to be considered indirect costs. While some disciplines or projects may present exceptional needs, many other research projects will likely rely on baseline services provided by institutions or disciplinary groups that need more general formulas for funding.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

- Reporting metrics should be developed and applied to early efforts at improving data stewardship, and the results shared broadly. As best practices emerge and community norms support good data management, researchers will have an incentive to preserve and share their data.
- Compliance should be verified through systematic approaches, which can be much easier and efficient for the agency and less punitive for researchers. Most researchers pay special attention to two milestone events in the research process – the grant proposal and publication. Policies and metrics that are embedded at these points will get the attention of researchers and make compliance more likely.
- Agencies should develop guidelines for those who review grant proposals that highlight what to look for in a well-developed data management plan within the specific discipline.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

- The use of open data licenses and platforms that facilitate sharing in standardized ways will make it easier for other researchers and industries to reuse data and increase the return on investment for funded research projects.
- Support for well-documented APIs that allow individuals and machines to develop new capabilities and services is key to fostering innovation.
- One of the benefits of the broadest possible access and opportunity for reuse is that federal agencies could help build on "citizen science" efforts, which have up until now largely focused on data gathering and classification. Open licensing and usable APIs will ensure that the maximum number of creative imaginations are looking for innovative ways to use research data.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

- Support should be given to ongoing efforts to develop data citation standards (such as the DataCite project) and author and institutional identifiers (such as those being developed by ORCID).
- Agencies should require disclosure of data sources using common data citation and researcher identification standards in order to build community norms that reward good attribution practice, as is the case for research articles.
- Nevertheless, it should be recognized that existing attribution standards for published articles will not translate seamlessly into the world of research data, especially given the importance of machine-based access and reuse. As in so many other areas, this is a case where standards will have to develop as reuse and innovation grows, and agency mandates should remain flexible while publicizing and encouraging best practices.

Joseph Serene
Treasurer/Publisher

Phone: (301) 209-3220
Fax: (301) 209-0844
Email: serene@aps.org

American Physical Society
One Physics Ellipse
College Park, MD 20740-3844
www.aps.org

The American Physical Society's response to the OSTP's request for information on "Public Access to Digital Data Resulting from Federally Funded Research," FR Doc. 2011-32947

Introduction

The American Physical Society (APS) was founded in 1897 with the objective to advance and diffuse the knowledge of physics. While this objective is now understood to include physics education and outreach, public affairs, scientific meetings, and international collaborations, the publication of significant advances in physics has been central to APS since 1913, when we became the publisher of the *Physical Review*, a journal founded at Cornell University in 1893. Since that time, APS physics journals have grown tremendously. We now publish ten journals: *Physical Review A-E* (each journal dedicated to a particular subject area in physics), *Physical Review Letters, Reviews of Modern Physics, Physical Review Special Topics – Accelerators and Beams, Physical Review Special Topics – Physics Education Research,* and *Physical Review X*. The last three are Open Access journals whose peer-review and other operating costs are covered by contributions or publication fees. Our other journals are available through subscriptions held by a variety of individual institutions and consortia around the world. The *Physical Review* journals and *Physical Review Letters* (our flagship journal) allow authors or their sponsoring institutions to pay an article-processing charge to have an individual article made freely available. All APS Open Access articles are available under the Creative Commons Attribution 3.0 License, and we no longer hold copyright to these articles. All APS journal content (back to 1893) was made available online by the end of 2001, making us one of the very first publishers to put our entire corpus online.

The APS journals are broadly international in scope. Only about 30% of our submissions (and published articles) come from authors within the U.S. Similarly, only about one third of our subscription revenue comes from the U.S. The remaining submissions and revenues are roughly equally divided between Europe and the Asia-Pacific region.

The APS has a long history of support for Open Access initiatives. In 1998, we became the first fully "green" publisher when we amended our copyright transfer statement to explicitly allow authors to post their manuscripts (both new and previously published) on e-print servers, such as arXiv.org, and to post PDFs of their APS-published articles on their home pages or institutions' web sites. Indeed, the recent content of one of our journals, *Physical Review D*, is essentially completely available on arXiv.org because of submissions by the authors themselves. *Physical Review Special Topics – Accelerators and Beams* was one of the earliest "gold" Open Access journals. It started publication as an Open Access journal in 1998 and is supported by contributions from accelerator laboratories around the world. Authors and readers incur no fees for this journal. In November 2009, the APS Council adopted the following statement:

> *The APS supports the principles of Open Access to the maximum extent possible that allows the Society to maintain peer-reviewed high-quality journals, secure archiving, and the Society's long-term financial stability, to the benefit of the scientific enterprise.*

In keeping with our objective, APS also recognizes the importance of making the research published in our journals as widely available as possible, even to the general public. We believe that it is essential for the general public to have access through our web site to the full, final, peer-reviewed content of all APS journals. This ensures that the public sees the official "version of record," including any updates or corrections. Thus, in July 2010, we pioneered a program that allows any U.S. public library to sign up for free subscriptions to all of our journals. After a librarian at a public library completes a simple online form, agreeing to straightforward terms and conditions, we grant access promptly (usually in one business day). Any person visiting a participating public library can access the full content of our journals dating back to 1893. We are pleased that the Library of Congress was the very first public library to sign

up under this program. This program was subsequently extended to all U.S. high schools, and to date well over 500 libraries from around the country have taken advantage of this opportunity.

Finally, APS prides itself on subscription prices per article and per page that are among the lowest in the industry. We were the first publisher to introduce tiered pricing, allowing smaller institutions with little research activity to pay substantially less than the leading research institutions. Subscription prices for our highest and lowest tiers currently differ by more than a factor of two, and we continue to increase (gradually) this ratio. Our article-processing fees for Open Access articles cover the actual cost of reviewing and publishing an article (without charging for submissions not accepted for publication), plus a very small margin that supports our the education and outreach activities. Twice in the most recent decade we have actually decreased our subscription prices as our expenses decreased (most recently in 2009). When we increase our prices the primary driving forces are inflation and growth in the number of manuscripts submitted for review (by 3-5% annually for many years).

OSTP Questions

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Not all scientific data has long-term value, but the costs of preservation (at least of "bits") are already very low for all but the very largest datasets. A number of centralized data-centers for scientific information have sprung up in recent years, some of them subject-oriented (for example PANGAEA for earth science data at http://www.pangaea.de/) and some of regional scope (for example the Australian National Data Service at http://www.ands.org.au/).

Each of these services has a commitment to preservation and public access (some may allow for an embargo period) so the first helpful federal policy in this regard would be to recognize and provide funding to support scientific data centers of these sorts in the U.S. In fact a number of them are already in place – Genbank, the Protein Databank, the NIST scientific and technical database, the National Nuclear Data Center, the OSTI DOE Data Explorer, etc. One beneficial policy would include some sort of certification process to ensure that these centers are meeting at least minimal standards for data preservation and access, coupled with continued federal support and sponsorship for the creation of centers covering new subject areas.

A second beneficial federal policy would be encouraging or requiring researchers receiving federal funding to deposit their research data in a certified data center, to ensure preservation and access. To allow for reuse, deposited data must be accompanied by sufficient context – instruments used, the meaning of data fields, geographic location, time stamps, etc. – so developing standards for that context could also be a federal role (but it is surely very field-dependent).

There may also be a need for federal funding for deposit and discovery tools, so that these processes are not a burden on researchers but do actually help to improve scientific productivity. These tools have great potential and some of that potential has been realized already in fields such as genetics; it could surely be more broadly realized to improve research productivity across many other fields.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

We believe that the most important step here is not to mandate any specific policy, but rather to allow the choice of a certified data center that has access policies most amenable to the stakeholders involved. If the scientists are most comfortable with a one-year embargo on public access while they work on their own data, then leaving them free to choose a data-center that allows for that embargo period should be acceptable. Private data-centers that retain copyright and provide access only to subscribers may be another option that researchers could be free to choose unless federal policy insists on public access after a period of time less than the standard copyright term.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Leaving the choice of (certified) data-center to the researchers should be sufficient for this.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

The bias should be for preserving everything. The difficulty is going to be with extremely large (particle accelerator) or diverse (bench-top science) datasets. Some filtering may be needed but long-term stewardship costs should continue to decline over time, widening the scope of what should be preserved.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

All should encourage the use of (certified) data-centers and development and use of appropriate tools for data deposit and discovery. Standards for metadata (data context) should be developed by the research community involved rather than by other parties.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Federal research agencies should recognize and directly fund the existing data centers, and should support the creation of new ones to cover fields poorly represented at present.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Every data deposit should come with an identifying citation, possibly a DOI; citations for deposited data should be provided in the same way that citations for publications are when reports are provided on grants and other funding. A project that produced data but provides no data deposit citation should be queried for compliance on the issue.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Agencies should build awareness of the research data that has been made available, including tools for discovery (providing relevant hooks for search engines like Google, for instance). Tools for data manipulation, aggregation, and conversion to different formats and other overlays may need to be developed; a commercial market for innovative applications could arise.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data citations should include the names of the producers, or they should be provided as part of the data context. Standards for data citation are being developed by organizations such as DataCite (http://datacite.org/).

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

[No comment]

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

[No comment]

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

[No comment]

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Standards for citation of data are the most important here - in particular the work of DataCite and assignment of DOI's (digital object identifiers) to individual citable datasets should be encouraged. Then publications can link to their associated data in much they same way they link to one another now, through their reference sections with appropriate citation practice.

Response to Request for Information - FR Doc. 2011-28621

Dr Cameron Neylon – U.K. based research scientist writing in a personal capacity

Introduction

Thankyou for the opportunity to respond to this request for information and to the parallel RFI on access to scientific publications. Many of the higher level policy issues relating to data are covered in my response to the other RFI and I refer to that response where appropriate here. Specifically I re-iterate my point that a focus on IP in the publication is a non-productive approach. Rather it is more productive to identify the *outcomes* that are desired as a result of the federal investment in generating data and from those outcomes to identify the services that are required to convert the raw material of the research process into accessible outputs that can be used to support those outcomes.

Response

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Where the Federal government has funded the generation of digital data, either through generic research funding or through focussed programs that directly target data generation, the purpose of this investment is to generate outcomes. Some data has clearly defined applications, and much data is obtained to further very specific research goals. However while it is possible to identify likely applications it is not possible, indeed is foolhardy, to attempt to define and limit the full range of uses which data may find.

Thus to ensure that data created through federal investment is optimally exploited it is crucial that data be a) accessible, b) discoverable, c)interpretable and d) legally re-usable by any person for any purpose. To achieve this requires investment in infrastructure, markup, and curation. This investment is not currently seen as either a core activity for researchers themselves, or a desirable service for them to purchase. It is rare therefore for such services or resource need to be thoughtfully costed in grant applications.

The policy challenge is therefore to create incentives, both symbolic and contractual, but also directly meaningful to researchers with an impact on their career and progression, that encourage researchers to either undertake these necessary activities directly themselves or to purchase and appropriately cost third party services to have them carried out.

Policy intervention in this area will be complex and will need to be thoughtful. Three simple policy moves however are highly tractable and productive, without requiring significant process adjustments in the short term:

a) Require researchers to provide a data management or data accessibility plan within grant requests. The focus of these plans should be showing how the

project will enable third party groups to discover and re-use data outputs from the project.

b) As part of the project reporting, require measures of how data outputs have been used. These might include download counts, citations, comments, or new collaborations generated through the data. In the short term this assessment need to be directly used but it sends a message that agencies consider this important.

c) Explicitly measure performance on data re-use. Require as part of bio sketches and provide data on previous performance to grant panels. In the longer term it may be appropriate to provide guidance to panels on the assessment of previous performance on data re-use but in the first instance simply providing the information will affect behaviour and the general awareness of issues of data accessibility, discoverability, and usability.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

As noted in my response to the other RFI, the focus on intellectual property is note helpful. Private contributors of data such as commercial collaborators should be free to exploit their own contribution of IP to projects as they see fit. Federally funded research should seek to maximise the exploitation and re-use of data generated through public investment.

It has been consistently and repeatedly demonstrated in a wide range of domains that the most effective way of exploiting the outputs of research innovation, be they physical samples, or digital data, to support further research, to drive innovation, or to support economic activity globally is to make those outputs freely available with no restrictive terms. That is, the most effective way to use research data to drive economic activity and innovation at a national level is to give the data away.

The current IP environment means that in specific cases, such as where there is very strong evidence of a patentable result with demonstrated potential, that the optimisation of outcomes does require protection of the IP. There are also situations where privacy and other legal considerations mean that data cannot be released or not be fully released. These should however be seen as the exception rather than the rule.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

At the Federal level only very high-level policy decisions should be taken. These should provide direction and strategy but enable tactics and the details of implementation to be handled at agency or community levels. What both the Federal Agencies and coordination bodies such as OSTP can provide is an oversight and, where appropriate, funding support to maintain, develop, and

expand interoperability between developing standards in different communities. Federal agencies can also effectively provide an oversight function that supports activities that enhance interoperability.

Local custom, dialects, and community practice will always differ and it is generally unproductive to enforce standardisation on implementation details. The policy objectives should be to set the expectations and the frameworks within local implementation can be developed and approaches to developing criteria against which those local implementations can be assessed.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Prior to assessing differences in performance and return on investment it will be necessary to provide data gathering frameworks and to develop significant expertise in the detailed assessment of the data gathered. A general principle that should be considered is that the *administrative and performance data* related to accessibility and re-use of *research* data should provide an outstanding exemplar of best practice in terms of accessibility, curation, discoverability, and re-usability.

The first step in cost benefit analysis must be to develop an information and data base that supports that analysis. This will mean tracking and aggregating forms of data use that are available today (download counts, citations) as well as developing mechanisms for tracking the use and impact of data in ways that are either challenging or impossible today (data use in policy development, impact of data in clinical practice guidelines).

Only once this assessment data framework is in place can detailed process of cost benefit analysis be seriously considered. Differences will exist in the measurable and imponderable return on investment in data availability, and also in the timeframes over which these returns are realised. We have only a very limited understanding of these issues today.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

If stakeholders have serious incentives to optimise the use and re-use of data then all players will seek to gain competitive advantage through making the highest quality contributions. An appropriate incentives framework obviates the need to attempt to design in or pre-suppose how different stakeholders can, will, or should best contribute going forward.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

As with all research outputs there should be a clear obligation on researchers to plan on a best efforts basis to publish these (as in make public) in a form that most effectively support access and re-use tensioned against the resources available. Funding agencies should make clear that they expect communication

of research outputs to be a core activity for their funded research, that researchers and their institutions will be judged based on their performance in optimising the choices they make in selecting the appropriate modes of communication.

Further funding agencies should explicitly set guidance levels on the proportion of a research grant that is expected under normal circumstances to be used to support the communication of outputs. Based on calculations from the Wellcome Trust where projected expenditure on the publication of traditional research papers was around 1-1.5% of total grant costs, it would be reasonable to project total communication costs once data and other research communications are considered of 2-4% of total costs. This guidance and the details of best practice should clearly be adjusted as data is collected on both costs and performance.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Ideally compliance and performance will be trackable through automated systems that are triggered as a side effect of activities required for enabling data access. Thus references for new data should be registered with appropriate services to enable discovery by third parties – these services can also be used to support the tracking of these outputs automatically. Frameworks and infrastructure for sharing should be built with tracking mechanisms built in. Much of the aggregation of data at scale can build on the existing work in the STARMETRICS program and draw inspiration from that experience.

Overall it should be possible to reduce the burden of compliance from its *current* level while gathering vastly more data and information of much higher quality than is currently collected.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

There are a variety of proven methods for stimulating innovative use of data at both large and small scale. The first is *to make it available*. If data is made available at scale then it is highly likely that some of it will be used somewhere. The more direct encouragement of specific uses can be achieved through directed “hack events” that bring together data handling and data production expertise from specific domains. There is significant US expertise in successfully managing these events and generating exciting outcomes. These in turn lead to new startups and new innovation.

There is also a significant growth in the number of data-focussed entrepreneurs who are now veterans of the early development of the consumer web. Many of these have a significant interest in research as well as significant resources and there is great potential for leveraging their experience to stimulate further growth. However this interface does need to be carefully managed as the cultures involved in research data curation and web-scale data mining and exploitation are very different.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The existing norms of the research community that recognise and attribute contributions to further work should be strengthened and supported. While it is tempting to use legal instruments to enforce a need for attribution there is growing evidence that this can lead to inflexible systems that cannot adapt to changing needs. Thus it is better to utilise social enforcement than legal enforcement.

The current good work on data citation and mechanisms for tracking the re-use of data should be supported and expanded. Funders should explicitly require that service providers add capacity for tracking data citation to the products that are purchased for assessment purposes. Where possible the culture of citation should be expanded into the wider world in the form of clinical guidelines, government reports, and policy development papers.

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

At the highest level there are a growing range of interoperable information transfer formats that can provide machine readable and integratable data transfer including RDF, XML, OWL, JSON and others. My own experience is that attempting to impose global interchange standards is an enterprise doomed to failure and it is more productive to support these standards within existing communities of practice.

Thus the appropriate policy action is to recommend that communities adopt and utilise the most widely used possible set of standards and to support the transitions of practice and infrastructure required to support this adoption. Selecting standards at the highest level is likely to counterproductive. Identifying and disseminating best practice in the development and adoption of standards is however something that is the appropriate remit of federal agencies.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

There is now a significant literature on community development and practice and this should be referred to. Many lessons can also be drawn from the development of effective and successful open source software projects.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

There are a range of global initiatives that communities should engage with. The most effective means of practical engagement will be to identify communities that have a desire to standardise or integrate systems and to support the technical and practical transitions to enable this. For instance there is a

widespread desire to support interoperable data formats from analytical instrumentation but few examples of bringing this to transition. Funding could be directed to supporting a specific analytical community and the vendors that support them to apply an existing standard to their work.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Development in this area is at an early stage. There is a need to reconsider the form of publication in its widest sense and this will have a significant impact on the forms and mechanisms of linking. This is a time for experimentation and exploration rather than standards development.

AMERICAN UNIVERSITY

WASHINGTON, DC

Jorge L. Contreras
202-274-4124
contreras@wcl.american.edu

January 12, 2012

National Science and Technology Council (NSTC)
Office of Science and Technology Policy (OSTP)
Attention: Ted Wackler, Deputy Chief of Staff
Via email: digitaldata@ostp.gov

Re: OSTP Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research (76 Fed. Reg. No. 214 at 68,517 (Nov. 4, 2011))

Dear Mr. Wackler:

I appreciate the opportunity to share comments with OSTP regarding public access to digital data resulting from federally-funded scientific research. I am a professor of law at American University, prior to which I spent seventeen years as a practicing attorney representing major research institutions, R&D consortia and private enterprises engaged in technical and scientific work. I have recently served as a member of the National Advisory Council on Human Genome Research and currently serve as Co-Chair of the National Conference of Lawyers and Scientists and Co-Chair of the American Bar Association Section of Science & Technology Law's Committee on Technical Standardization. My current research focuses on the production and dissemination of scientific and technical information.

Responses to specific items of the OSTP RFI are set forth below and represent my own views, and not those of American University, Washington College of Law, or any of the other organizations mentioned above.

Preservation, Discoverability, and Access

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Public access to data from federally-funded research has several potential benefits. These include enabling scientists to reproduce and validate the results of their peers, reducing the incidence scientific fraud and misrepresentation, and accelerating the overall progress of scientific discovery. These benefits, however, do not come without cost, and several counter-balancing effects of data release must be considered. These include the exposure of potentially

identifiable personal information of human research subjects, the reduction of intellectual property protection in released data, and the reduction of publication opportunities for data-generating scientists.

In previous work, I have analyzed these costs and benefits in the context of federal data release policies relating to human genomics research, and have traced the development of these policies from 1992 through 2009.[1] Over this period, genomics data release policies have evolved significantly. The so-called "Bermuda Principles", adopted in 1996, required the release of genomic sequence data within 24 hours of generation. Among the purposes of the Bermuda Principles was to limit the ability, both of data producers and third parties, to obtain patents claiming raw sequence data generated by the public genome project. Despite the success of the initial genome project, subsequent projects and policies have seen a measured retreat from the sweeping requirements of Bermuda. Today, the major federal genomics data release policies (e.g., the 1997 NIH GWAS policy, the 2008 ENCODE and modENCODE policies, and the 2009 Human Microbiome Project policy) require that data be released (i.e., deposited into publicly-accessible, federally-managed databases such as Genbank and dbGaP) subject to detailed user agreements and requirements (see also response to Question 7, below). These agreements typically impose an "embargo" period on users of data, prohibiting them from publishing or presenting conclusions derived from this data during a pre-determined period of time (usually 9-12 months).

Interestingly, a number of private biomedical research projects have adopted similar approaches to data release. In some of these cases, however, data is withheld for a period of time (also between 9-12 months), after which it is released without restriction. The fact that both public and private research projects have independently arrived at "embargo" periods in roughly the same range (i.e., 9-12 months) implies that the length of this period (which I refer to as the "latency" period) can be viewed as an equilibrium of sorts. That is, at the latency equilibrium point, the various stakeholder groups negotiating such policies (i.e., funders, data-generating scientists, data-using scientists, and public advocates) are each willing, albeit reluctantly at times, to make data public. A shorter period would not be acceptable to data-generating scientists as it would not adequately compensate them for their effort, and a longer period would not be acceptable to funders and data-using scientists, who have an interest in making such data freely available at the earliest possible time. Thus, as a result of multilateral compromise, an equilibrium latency period acceptable to all stakeholder groups may emerge.

We have seen the development of similar latency periods and equilibria in the area of scientific publishing. In this area, publishers, funders, libraries, universities and scientists have engaged in a series of explicit and implicit negotiations (contractual, administrative and legislative) over the appropriate latency period before which published scientific research may be released to the public free from access restrictions. Again, a latency period between 6-12 months emerges as an

---

[1] *See* Jorge L. Contreras, *Prepublication Data Release, Latency, and Genome Commons*, 329 SCIENCE 393 (2010), *Data Sharing, Latency Variables and Science Commons*, 25 BERKELEY TECH. L.J. 1601 (2010), Jorge L. Contreras, *Bermuda's Legacy: Patents, Policy and the Design of the Genome Commons*, 12 MINN. J.L. SCI. & TECH. 61 (2011).

equilibrium point in several independent contexts. And again, this convergence suggests that a latency period in this range appropriately rewards publishers, on one hand, and adequately meets the access requirements of scientists, libraries and the public, on the other hand.

In my view, these data suggest that "market" forces (i.e., the negotiation and interplay of interested stakeholder groups) can arrive at protective periods that are substantially shorter than default intellectual property rules (20 years in the case of patents, and nearly 100 years in the case of copyright).[2]

Extrapolating beyond the areas of genomics and scientific publishing, I believe that many (if not all) fields of scientific endeavor will exhibit a latency equilibrium point at which the release of data to the public will duly balance the costs and benefits to the relevant stakeholder groups. I do not suggest that this equilibrium point will be the same in all fields. In fact, I believe that different fields could exhibit radically different latency equilibria. It is quite likely that a field such as paleoanthropology would exhibit a substantially longer latency equilibrium point than genomics. Moreover, I do not attempt here to pre-judge whether the release of data prior to or after publication of the associated analysis is preferable in one field versus another,[3] and again suspect that the norms, practices and practicalities of different scientific disciplines would dictate the optimal practice in each such discipline.

Nevertheless, I believe that attempting to discern the latency equilibrium point in a scientific field can yield valuable information regarding the appropriate balancing of intellectual property and other interests among competing stakeholder groups, and I would encourage OSTP to consider this analysis in its future policy development activity.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Improving compliance with Federal data access policies embodies two distinct and significant challenges: monitoring compliance and enforcing (and formulating) penalties for noncompliance.

Monitoring: To date, most violations of data access policies are identified on an ad hoc basis by scientific colleagues, competitors and journal editors (whistleblowers) rather than agency staff. Given Federal budgetary constraints, I imagine it unlikely that an effective inter-agency

---

[2] The fact that these observed latency equilibrium periods are so much shorter than default intellectual property periods also suggests that the default periods may, in some cases, be unnecessarily lengthy (a point that has been made by many others).

[3] Pre-publication and post-publication data release, and the considerations surrounding each, are analyzed in a pair of companion pieces that appeared in *Nature* in 2009. Toronto International Data Release Workshop Authors, *Prepublication Data Sharing*, 461 NATURE 168 (2009) and Paul N. Schofield, Tania Bubela, Thomas Weaver, et al., *Post-Publication Sharing of Data and Tools*, 461 NATURE 171 (2009).

compliance monitoring system could be implemented in the near future. A reasonable alternative might be to coordinate (and incentivize) private whistleblowing activity through a central Federal data oversight board. Such a board could promulgate rules and guidance regarding the reporting of data access/use violations and could encourage relevant stakeholders to report such violations on an anonymous basis. To the extent that investigation of reports is warranted, the board could allocate resources to such investigation or refer the reported violation to the relevant agency.

Enforcement: I have previously written about the somewhat dubious enforceability of Federal data access policies, particularly with respect to third party users outside of the U.S. There are several potential avenues toward improving policy enforceability, both contractually and through international trade mechanisms. However, before embarking on a major enforceability program, it would be prudent to gather data regarding compliance as outlined above. In particular, information regarding overall levels of non-compliance, together with any data that emerge regarding the characteristics of policy violators and the nature, frequency and seriousness of their violations, would be useful to consider. With this information in hand, one could more accurately assess options for improving compliance and potential penalties for non-compliance.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Data exchange standards are ubiquitous in today's technology-driven marketplace. Such standards, which are commonplace in the information, computing and telecommunications (ICT) sector, specify the manner in which products and services offered by different vendors interact with one another. A number of these standards, including WiFi, USB, CD, DVD, PDF and HTML, have become household terms, and thousands of others ensure that a vast array of products and services connect and communicate seamlessly in a manner that is largely invisible to the consumer.

Standardization in the ICT sector, however, has not come without a cost. Over the past two decades, the industry has been plagued by lawsuits brought by participants in the standards-development process as well as by government regulators and affected third parties. Two types of claims generally arise in standards-related litigation: claims that the standards process has been abused to disadvantage one or more companies ('process-abuse' claims) and claims that a participant in the standards-development process has improperly asserted patents against an implementer of the standard ('patent hold-up' claims). Standards-development organizations in the ICT sector have responded to these claims by promulgating rules and policies of increasing sophistication, both to specify procedures designed to avoid abusive activity and to accommodate the requirements of participants who control significant patent assets.

The explosion of data-driven scientific research over the past two decades has led to a surge of interest in the development of interoperability and compatibility standards. These range from standards for genome annotation and controlled vocabularies (ontologies) to data formats and search engine integration. A variety of organizations are involved in these standards-development activities, from large, established standards bodies such as the Institute for Electrical and Electronics Engineers (IEEE) and the Worldwide Web Consortium (W3C) to broad-based industry associations such as the European Bioinformatics Institute (EBI) to narrowly-focused efforts such as the Proteomics Standards Initiative (PSI) and the Functional Genomics Investigation Ontology (FuGO) project. To date, most science-driven standardization efforts have been free of the litigation that has plagued the ICT industry. But with the increasing adoption of standards by researchers and vendors, the issues faced by ICT standards groups will become increasingly relevant.

Today, the large majority of science-focused standards-development efforts are relatively informal and unstructured, and are thus ill-equipped to address or deter process abuse and patent hold-up scenarios. In many cases, the organizations responsible for standards development either lack written policies entirely, or adopt vague, aspirational statements regarding a desire that materials produced be "open" and publicly-available. This informal and minimalist approach not only invites opportunistic behavior, but also leaves aggrieved participants with little legal recourse after abuse has occurred.[4]

Accordingly, I recommend that OSTP encourage science-focused standards-development organizations to review their existing policies and procedures with care. To the extent that they fail to address key points regarding process openness and intellectual property, these policies and procedures should be supplemented.[5] For example, if it is the desire of a group that all scientists worldwide be permitted to access and implement a new scientific data sharing standard without the payment of copyright or patent licensing fees, the group's policy should state this clearly and require contributing participants to commit not to assert copyrights or patents in connection with the standard. Hopefully, such modest prophylactic measures will enable the scientific standards community to avoid the disruptive and costly litigation that has affected the ICT sector.

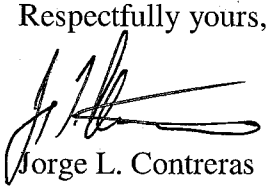
---

[4] *See* Jorge L. Contreras, *Legal Issues in the Development of Biological Research Standards,* 26 NATURE BIOTECHNOLOGY 498 (2008).

[5] A number of resources exist to assist non-lawyers with understanding and developing appropriate standards-development policies. *See* ABA COMM. ON TECH. STANDARDIZATION, STANDARDS DEVELOPMENT PATENT POLICY MANUAL (Jorge L. Contreras, ed., 2007).

Thank you again for the opportunity to offer these comments in response to your inquiries. Please do not hesitate to let me know if there is any additional information that I can provide in support of these matters.

Respectfully yours,



Jorge L. Contreras

Response to Request for Information: "Public Access to Digital Data Resulting from Federally Funded Research," November 2011 January 12, 2012

G. Sayeed Choudhury
sayeed@jhu.edu<mailto:sayeed@jhu.edu>
Johns Hopkins University Libraries
Baltimore, MD

Prudence S. Adler
prue@arl.org<mailto:prue@arl.org>
Association ofResearch Libraries
Washington, DC

Heather Joseph
heather@arl.org<mailto:heather@arl.org>
SPARC
Washington, DC

Summary
Thank you for the opportunity to comment on "Public Access to Digital Data Resulting from Federally Funded Research." These comments are submitted on behalf of the Johns Hopkins University Libraries, the Association of Research Libraries (ARL), and the Scholarly Publishing and Academic Resources Coalition (SPARC). The Johns Hopkins University Libraries have established a leadership role with digital data management through a long-term program of R&D, prototyping and implementation of data infrastructure. ARL is an Association of 126 research libraries in North America. These libraries directly serve 4.6 million students and faculty and spend $1.4 billion annually on acquiring information resources, of which 62% is invested in access to electronic resources. SPARC is an international alliance of academic and research libraries. Action by SPARC in collaboration with stakeholders – including authors, publishers, and libraries – builds on the unprecedented opportunities created by the networked digital environment to advance the conduct of scholarship.
Question 1) Whatspecific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientificresearch, to grow the U.S. economy and improve the productivity of the American scientific enterprise?
Comment 1) The most effective Federal policies in thisregard would mandate digital data deposit into publicly accessible repositories. In the absence of such policies, there are already cases of digital data which have been lost or remain inaccessible or accessible only with high barriers. While laudable efforts such as the NSF and NIH data management plans move the community in the direction of supportingU.S. economic growth and productivity, the reality is that many researcherscontinue to strictly interpret the requirement as sharing data based on specific requests or personal provisions. The Federal policy framework should move public access to digital data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use. Instead of relying upon individual investigators to interpret and support public access through a point to point network (e.g.,researcher provides digital data upon request), Federal policies should ensure that public access can occur through well managed, sustained, preservation archives that enable a legally and policy compliant peer to peer model for sharing. A useful metric for full-fledged public access to digital data is whether someone (or some machine) other than the original data producer can discover, access, interpret and use the digital data without contacting the original data producer. And such infrastructure becomes particularly important as science is increasingly interdisciplinary and global. Finally, fundamental to science is the ability to replicate. Researchers must be able to access data in order to

reproduce results and in the current economic climate, we will not see the same level of research funding so it is imperative that data be shared.

Question 2) Whatspecific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2) Raw data are not subject to copyright. There are fundamental differences between peer-reviewed publications and digital data stemming from federally funded research in the context of copyright and intellectual property. The existing copyright framework and associated business models for peer-reviewed publications are not appropriate for digital data. Unlike publications, data are not processed or reviewed by publishers. Any potential assignment of rights should be applied only to derived datasets for which there is tangible intellectual or creative interpretation or processing. Even in such cases, copyright should be not assigned in an exclusive manner that would curtail or inhibit preservation, discovery and sharing of data. While the Creative Commons CC0 and CC-BY are legallydefensible licenses for publications, they would require augmentation to apply for all types of digital data. Their principles represent an appropriate foundation from which to build a license for digital data that acknowledges potential copyright issues while maximizing the prospects for both people and machines to build services that maximize utility.

If the US Government were to consider the limited use of embargoes, the one reasonable argument for embargoes relates to the unique effort exerted by the digital data producers or the original scientific team. It is true that the original digital data producers exert (often) unique effort that could be acknowledged in terms of a limited, fixed-term,exclusive access to this data. However, it is also important to note that such an arrangement should not confer copyright or preclude the deposit of the digital data into a certified digital data repository even during an embargo period, particularly to initiate archiving activities and (ultimately) subsequent sharing mechanisms.

Question 3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3) There is ample evidence that different scientific disciplines present a variety of requirements for management of digital data. Fundamentally, there are two important considerations in this regard. First, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. While there are notable exceptions, too many scientific disciplines have focused primarily on access or discovery rather than archiving or preservation. There are critical pieces of the digital data infrastructure that can support archiving, preservation or enhanced access (e.g., identifiers, fixity information) that should be applied to all scientific data. For example, the Public Library of Science assigns identifiers to figures within articles. By doing so, it becomes possible to discover, share and preserve data at a more granular level.

Second, while scientific communities are indeed the most qualified to decide regarding appropriate community practices and norms, the reality is that many scientists do not fully understand the implications oftheir preferences or choices or appreciate the choices available to them. In this context, social science and information science research has started to identify implicit and explicit issues that relate to differences and commonalities across scientific disciplinary data practices. As noted in the recent National Science Board report, “Digital Research Data Sharing and Management,” communities of practice should “take responsibility for determining its own standards and conventions for data stewardship and for coordination across the research enterprise.” It is equally important to ensure that when scientific communities identify and recommend data management practices, they are held accountable to the public access concept and focus on scientifically defensible criteria.

Question 4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Comment 4) The reality is that different types of data and community-based requirements will introduce different, relative costs and benefits. However, it is most useful to consider requirements particularly as they relate to baseline services thatapply across all disciplines (e.g., archiving) and secondary services (e.g., specialized query capabilities). Agency policies might consider the relative emphasis between these two categories of services, especially as it relates to distribution of costs and benefits across the full array of stakeholders. For example, a federal agency might wish to fund research libraries to develop the capacity for digital data archives and look to new sustainable funding models for the long-term preservation and access to these digital resources. In this sense, an agency might provide seed funding to develop and establish the preservation infrastructure, provide ongoing funding to a scientific community to develop secondary services and explore new partnerships for long-term preservation and access.

Question 5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Comment 5) There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., Interuniversity Consortium for Political and Social Research for survey-based social science research). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation are being used loosely without associated evidence of meeting specific requirements. Cultural memory institutions such as archives, libraries and museums have an extensive track record with these functions and could serve the essential purpose of developing and/or implementing frameworks that thoroughly test and certify assertions. With a clearly articulated set of requirements, it will become possible to identify how various stakeholders can implement data management plans, noting that these roles will vary by discipline or community.

Since the National Science Foundation and the NationalEndowment for the Humanities announced their agencies' guidance on data management plans, a growing number of research libraries have collaborated with researchers and scientists on developing effective data management plans for proposal submission. It is certainly worth learning from and leveraging thelessons learned from these institutions and in particular, those institutions that directly handle data offer the richest experiences to consider. For example, Johns Hopkins University Sheridan Libraries has acted upon an agreement with the Astrophysical Research Consortium to archive and preserve the Sloan Digital Sky Survey (SDSS) datawhich are considered to be an exemplar collection within the scientific community. SDSS data, which comprise almost 140 terabytes, are also used extensively by citizen scientists who have even helped discover new astronomical objects. On an institutional level, the Johns Hopkins University Libraries have established a data management service that has helped over 35 research teams prepare data management plans for National Science Foundation proposals and committed to archive and preserve the data from these proposals. The Johns Hopkins University Libraries have already acquired the first datasets to be preserved and shared through this data management service.

Question 6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Comment 6) Given the critically important role of digital data to the scientific enterprise, the most important step would be to acknowledge and communicate to federal grantees that the real costs of preserving and making digital data accessible are indeed legitimate and important costs of the overall research enterprise. Researchers do not generally object to including publication costs within their research proposals; it is important to assert that proper data management should be viewed in the same manner.

In addition, agencies could support funding of twenty-first century workforce development. There is some currently underway such as that provided by the Institute of Museum and Library Services but

additional support by other agencies is needed. This was acknowledged in the recent National Science Board report, Digital Research Data Sharing and Management.

Question 7) Whatapproaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Comment 7) One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

There are two milestone events that every researcher cares about deeply: proposal submission and publication submission. These two points of leverage represent the best instances to introduce or implement policies given the heightened attention of researchers. By embedding appropriate policy, license and infrastructure requirements or components into these workflows, the prospects for efficient compliance and verification are heightened considerably. Researchers will likely complain about the additional burden but as their institutions or their communities develop capacity to support and implement data management plans, those "burdens" can be shifted to entities that view such activity as part of their core mission (i.e., do not view them as burdens but rather core business). This was the case with the implementation of the National Institutes of Health Public Access Policy. Compliance is now considered routine and a key component of ensuring future grants will flow to researchers and the research institution.

Agencies could also provide guidelines to proposal reviewers highlighting the elements of a well-developed data management plan, noting that disciplinary or community practices may vary (e.g., National Science Foundation could develop such reviewer guidelines by directorate).

Question 8) Whatadditional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8) Federal agencies could stimulate the development of public access digital data archives that support discovery and download and also support APIs that allow individuals and machines to develop new capabilities and services. In particular, this type of open system would facilitate new opportunities for all types and sizes of businesses including small businesses, perhaps somethinglike an app store for data. Thelicensing arrangements would be critical to ensure that one single entity or group does not secure an exclusive right to generate new business opportunities. By fostering a broader array of participation, federal agencies could help build upon citizen science efforts which, to date, have primarily engaged the public mainly through data gathering or data classification activities. Examples of such successful initiatives that offer rewards to individuals or teams working on projects include:

http://showoffyourapps.challenge.gov/

http://dev.mendeley.com/api-binary-battle

http://appsforscience.com/

Question 9) Whatmechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9) While this topic remains an active area of research and consideration, one of the most important components is author and institutional identifiers (e.g., ORCID) that would support developing attribution and credit processes. Additionally, it seems unlikely that extending existing attribution and credit frameworks or mechanisms can be seamlessly or easily ported into thedata realm, particularly given the importance of machine-based access. Other metrics do exist such as those outlined at:

http://altmetrics.org

By providing all of these metrics through organizations such as ORCID, a greater level of attribution and the impact of the research can be measured.

Question 10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Comment 10) There are many community-driven data standards for digital, scientific data, most of which deal with interoperability or sharing rather than archiving or preservation. One useful activity, perhaps through an inter-agency group or process, would be the development of an inventory of such standards identified or labeled by function. There are too many to list succinctly within this response. An example of a comprehensive list from the bioscience community is:
http://biosharing.org/standards
Additionally, the minimum metadata requirements for DataCite [1] require at least 5 standard descriptors but also feature an optional, additional 17 extra pieces of information that can be added at the discretion of the researcher.

Question 11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Comment 11) There are examples within various domains such as FITS within astronomy and FGDC within the earth sciences. In each of these cases, there are undoubtedly several characteristics or reasons for the success (or alternately reasons why such efforts did not succeed in other cases). Social science or information science research offers the most promising means for rigorously studying such processes, especially toward generalizedlessons that may be applied across domains or toward policy development.

Question 12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Comment 12) While there exist groups that work in this area (e.g. CODATA), it would be helpful for Federal agencies to support community-based efforts that connect nodes of data infrastructure development activities. For example, the European-based EUDAT project has already reached out to projects within the U.S. regarding a Data Access and Interoperability Task Force (DAITF) along the lines of the Internet Engineering Task Force. The National Institute of Standards and Technology could be helpful in this context especially toward an idea of a “data grid” that would operate in a similar manner to the power grid.

Question 13) What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13) There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. The peer-reviewed publication is viewed as the final “snapshot” of the research process and outcome. One of the most important considerations from a policy, practices and standards consideration is thatthere be a requirement to use persistent, unique identifiers for publications, data, authors, figures, etc. These identifiers not only bolster the linking of publications and data, but alsohelp foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

Thank you once again for this opportunity to respond to the Request for Information: Public Access to Digital Data Resulting from Federally Funded Research. Any inquiries related to this response should be addressed to G. Sayeed Choudhury (sayeed@jhu.edu).

_____

[1] http://datacite.org/schema/DataCite-MetadataKernel_v2.0.pdf

UNIVERSITY OF MICHIGAN
Library

Paul N. Courant
University Librarian and Dean of Libraries
Harold T. Shapiro Collegiate Professor of Public Policy
Arthur F. Thurnau Professor
Professor of Economics and of Information

818 Harlan Hatcher Graduate Library South
Ann Arbor, Michigan 48109-1205
734 764-9356 pnc@umich.edu

11 January 2012

Office of Science and Technology Policy on behalf of
National Science and Technology Council
Attention: Ted Wackler, Deputy Chief of Staff
digitaldata@ostp.gov

Re: Response to Notice for Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research (FR Doc. 2011-28621)

Dear Mr. Wackler:

Since 1838 the University of Michigan Library has been serving the research needs of students, faculty and the public. Over its many years of operation the library system has acquired an enormous wealth of diverse resources and continues to be a springboard for research, invention, and learning. Today, the lifeblood of much research and inquiry is data. Scholars, inventors, economists, medical researchers, social scientists, astronomers – all disciplines - look to data to find patterns, make predictions, identify stories of the past and present that may help us make a better future.

American taxpayers invest a tremendous amount in research, reflecting our national commitment to education and fundamental research. The resulting data are paid for by taxpayers and should be made available for further inquiry (along with publications produced as the result of that research). The success of the NIH mandate and PubMed Central as a free, publicly accessible, reliable source for NIH-funded research provides an important practical and philosophical model for making data produced by taxpayer-funded research broadly available.

My response to the questions raised in the Invitation to Comment follows. Thank you for the opportunity to comment.

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

A policy that ensures the ability of scientists, federal agencies, and the public to make use of data produced by federally funded research will foster economic growth and, by enabling other scientists to re-use that data, will improve our scientific productivity and increase innovation.

Data sharing reduces redundancy and eliminates wasteful uses of federal funding, since lack of access to scientific data makes it impossible for scientists and funders to answer important questions without needless duplication of research. It will help to ensure the scientific integrity of federally funded research by enabling the verification of published results and by exposing errors when they occur. It will improve sponsors' ability to allocate funding efficiently by providing better metrics for the measurement of scientific influence and progress.

Finally, funders will gain an enhanced ability to demonstrate the positive impact and value of work they support, helping to ensure the continued availability of federal funding for scientific research.

From the perspective of scientists and their institutions, sharing of scientific data has the potential to promote the development of broad-based metrics for measuring scientific influence. This will improve the current credentialing process, which relies almost exclusively on formal publication and its attendant metrics (such as a journal's "impact factor"), freeing institutions to promote broader means of disseminating the knowledge they create.

A key requirement of a public access policy that functions as a driver of the broader economy is to make sure data are openly licensed in a way that permits widespread reuse, including commercial as well as non-commercial uses. These open licenses could address possible integrity questions for compilations or other bodies of information. That said, data itself are in the public domain and not subject to copyright at all under US law. It is critically important that data be available in a manner sufficiently unencumbered to allow for innovative uses, reuses, and recombinations permitting new insights. This way, the public can benefit from the research that it funds. To be used, data also need to be discoverable, whether by a person or a machine, without the need to contact the original data producer.

Because scientists need strong incentives to share, the policy should encourage the use of appropriate licenses and discoverability-supporting standards through detailed guidance provided to them. Awards must be contingent on verifying that data from past projects are available and discoverable. Finally, and to the greatest extent possible, the policy should be implemented in a consistent way across all funding agencies.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Data licenses must be crafted to protect the rights of scientists who create the work and the rights and interests of the agencies and taxpayers who fund that work. This must be the first priority of any policy.

The policy should assure that patent rights, the right to enter into publishing contracts, and all other intellectual property rights are retained by researchers. It should also assure that the chosen licensing scheme does not limit unnecessarily the ability of other parties to make use of the data.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Current data sharing practices are diverse, making a single, mandated approach to data management difficult. Funders should encourage scientists in each discipline to develop their community's standards and norms. They should also set benchmarks to ensure the preservation and availability of data and to guide scientists toward consensus in disciplines where it has not yet been achieved.

One of the biggest obstacles scientists currently face when contemplating how to share their data is a lack of specific funder guidance. To enable scientists to plan adequately for data preservation and sharing there must be clear guidance provided on certain fundamental aspects of data management. Federal agency guidelines might include:

- Definition of research data
- Data sharing and access policies (including preferred timelines for sharing relative to the time of publication and the conclusion of the award)

- Minimum data retention periods
- Preferred disciplinary repositories
- Preferred file formats for specific types of data
- Preferred metadata standards
- Preferred access mechanisms (modes of data delivery) and licensing schemes
- Admissible exceptions to data sharing requirements (for e.g. privacy or security reasons)

As described under question 4 below, the creation of domain-specific data repositories where none currently exist would go a long way toward ensuring the long-term availability of valuable data, and would even out some of the differences in data sharing and archiving support for various disciplines.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

While it would be ideal if we could consistently and reliably identify data for which the benefits of preservation exceed the costs, in reality it is difficult to know in advance what the value of a scientific dataset might be several years from now. And although we can make assumptions about the long-term costs of data preservation, we cannot know for sure what they will be. So both sides of the cost-benefit equation are unknown.

However, at present the greatest barriers to data preservation and sharing have little to do with our ability to make this determination, so creating the right incentives and support structure to facilitate the sharing of research data must be our first priority.

Facilitating the creation of disciplinary data repositories in areas where they are needed is one way to help achieve this. Another is to make sure that scientists are given enough guidance to be able to identify the communities for which their own data are potentially relevant and to budget realistically for data preservation and sharing. Identifying at-risk datasets seems particularly urgent in light of the difficulties often faced even by major research collaborations in assuring continued access to experimental data that are unique and non-repeatable (see "Data Preservation at LEP" for one interesting case study). Clear guidance from funders will help to ensure that scientists identify communities of interest and budget appropriately for long-term preservation and access.

"Data Preservation at LEP." André G. Holzner, Ryszard Gokieli, Peter Igo-Kemenes, Marcello Maggi, Luca Malgeri, Salvatore Mele, Luc Pape, David Plane, Matthias Schröder, Ulrich Schwickerath, Roberto Tenchini, Jan Timmermans. arXiv:0912.1803v1 [hep-ex].

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

A successful approach to data management requires cooperation among many stakeholders, but here we will limit ourselves to discussing the role of libraries and scientific publishers. Libraries can help fill gaps in the data preservation infrastructure by preparing and archiving data that doesn't have a home elsewhere. We can connect scientists with existing services and in some cases we can manage data locally. We can serve an advisory role, recommending best practices when it comes to basic digital preservation strategies, but we will probably not have the domain knowledge needed to curate every dataset.

Scientific publishers have made significant contributions to the cause of scientific data sharing by requiring authors to make supporting data available, by crafting advice on data archiving options, and by making efforts to link datasets with the published literature. These efforts should be encouraged, but with the understanding that scientists cannot rely on them exclusively for data hosting and dissemination via supplemental materials published in traditional peer-reviewed journals. This would already be problematic simply by virtue of the fact

that most journals are only available to subscribers, rendering these datasets off-limits to most taxpayers. But through the materials made publicly available by the joint NISO/NFAIS working group on journal supplemental materials, it has become apparent that many scientific publishers see data as a liability, not an asset. The group is concerned first and foremost with limiting the scope of what can be considered a supplemental material, a measure aimed at reigning in costs associated with data hosting and review of submitted materials. The publishers in this group have stated explicitly that raw datasets fall outside of their purview, and so cannot be relied upon to curate original data even in the short term. They have also stated that they intend to manage supplemental materials under the same rights regime they commonly apply to published articles, i.e. exclusive copyright is to be signed over to the publisher.

As a result, while we should encourage and support all scientific publishers who are willing to open their archives to the general public to do so, such archives cannot and should not be the only archives or long-term stewards of research data. Those whose mission is to act on behalf of the public interest, be they Federal agencies or universities and colleges acting on their behalf, must have direct involvement in every aspect of assuring public access to data. Publishers currently have few incentives for making data broadly accessible, so creating those incentives and ensuring shared and co-equal ownership, storage, and responsibility for access is essential in any public-private partnership.

NISO/NFAIS Supplemental Journal Article Materials Project. http://www.niso.org/workrooms/supplemental

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The single most important step funders can take to improve scientists' ability to budget for data preservation and access is to assure that agencies set minimum data retention periods (per above, these may differ by discipline, but should be made as consistent as possible across all agencies). Doing so will allow scientists to plan for how long they will need to store their data after the conclusion of the award, and thus estimate the costs of doing so. Funders should also encourage the inclusion in data management plans of an explicit data sharing timetable describing when data will be prepared, deposited, and their availability verified before the end of the award, as well as the use of publicly accessible data repositories whose hosting costs are known in advance.

One way to address the considerable costs in time and effort of preparing data for archiving is to compensate scientists for this labor by providing better incentives for data sharing. In parallel, we must lower the cost of doing this work, a goal that can be achieved in a number of ways. Since paid data curators take much of the burden of preparation off scientists' hands, facilitating the existence of well-funded disciplinary data repositories for most disciplines will lower the costs of data preparation and hosting. Further, encouraging scientists to make good data management practices an integral part of their research process will reduce the burden of data preparation and description at the conclusion of the project. This can be done through specific guidance and through more systematic approaches to verifying compliance at key points during the course of the project.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Implicit in any discussion of a policy or mandate is the need to provide incentives for compliance. Currently, some agencies and directorates have made an explicit commitment to monitor compliance with data sharing policies through annual and final reports, while others have left it an open question as to how or whether compliance will be monitored or acted upon. As with the creation of data management guidelines, mechanisms for verifying compliance don't need to be uniform across all disciplines but they do need to be implemented consistently for all scientists.

However, the best way to ensure that data are shared is to give scientists positive incentives to share. Scientists value data sharing (Tenopir et al., 2011), yet many do not share data because doing so requires significant work and offers little reward. To make data sharing worth their while they must be able to demonstrate the value of their data to funders and to their fellow scientists. Although technologies and standards to support data attribution exist, they are not widely implemented and don't yet carry the weight that publication does. Thus, mandates alone will not create the incentives needed to foster a culture of data sharing; it must be valued sufficiently by the funding agencies and the scientific community at large to justify the effort required. Increased citation to publications as a result of data sharing has been demonstrated, but by itself it's not enough to incentivize scientists; they must receive credit directly. One key to achieving this goal is the development and use of new scientific performance metrics ("altmetrics") that take dataset citation and usage into account.

In many science disciplines the traditionally inseparable communication and credentialing functions of scholarly journals have long since become decoupled, with the communication function shifting to more timely venues such as the preprint archive (Gentil-Beccot et al., 2009). But the continued reliance of the academic credentialing process on formal publication and its attendant metrics has left scientists with little incentive to share data and other research products. A more inclusive and accurate way of measuring scientific progress would benefit all stakeholders, bringing the credentialing function of publication back in line with the communication function. Scientists should be rewarded for doing more and better science, which means sharing data as well as publishing papers. Not only will this benefit science as a whole (through better verification of results and a reduction of duplicated research), it will also improve sponsors' ability to allocate funding efficiently. By the same token, funders will gain an enhanced ability to demonstrate the positive impact and value of the work they fund, helping to ensure the future availability of public funding for research.

To lower the burden of verification and compliance a systematic approach to data management is key. Rather than making data available by request via a wide array of idiosyncratic means, researchers should be encouraged to make use of standard technical components including managed repositories, persistent identifiers, and standard licenses attached to datasets. Metadata could include a reference to a grant number or other means of identifying the sponsor, and could be made available through a machine-readable interface. Such mechanisms would facilitate the automation of compliance and verification tasks, thus reducing costs to sponsors and researchers. Making metadata available and machine-readable will also facilitate the growth of altmetrics and new business models centered around them.

"Sharing Data: Practices, Barriers, and Incentives." Carol Tenopir, Carole L. Palmer, Priyanki Sinha, Jeffrey van der Hoeven, and Jim Malone. *ASIST 2011*, October 9-13, 2011, New Orleans, LA, USA. http://www.asis.org/asist2011/proceedings/submissions/26_FINAL_SUBMISSION.doc

Altmetrics. http://altmetrics.org/

"Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories." Anne Gentil-Beccot, Salvatore Mele, Travis Brooks. arXiv:0906.5418v2 [cs.DL]

Thank you for your consideration.

Sincerely,

Paul N. Courant

Digital data RFI **January 12, 2012**

Submitters: Prof Jason R. Swedlow and Dr. Emma Hill, Open Microscopy Environment 'http://openmicroscopy.org'

Background

Since 2000, the Open Microscopy Environment (OME) has developed data specifications and software tools to handle complex multi-dimensional scientific image data for the life sciences community. OME is an open-source, international consortium of academic scientists building data management tools for life sciences imaging. All resources built and maintained by OME are available at http://openmicroscopy.org.

OME releases OME-TIFF, a specification for an open, multi-dimensional image file format, Bio-Formats a software plug-in library that accesses >120 scientific image file formats, and OMERO an open-source, enterprise-level application. OME works with a large number of commercial imaging companies to provide support for open file formats and their software and for supporting their own proprietary file formats. In 2005, OME founded a commercial arm, Glencoe Software, Inc., to provide opportunities for customising OME software for specific uses. This led to the development of the JCB DataViewer (http://jcb-dataviewer.rupress.org) the world's first online scientific image publication system. The JCB DataViewer is built upon OME's open source foundation and is an example of the delivery and power of open tools for data publication and archive.

Our responses to the RFI reflect over 10 years of work in the field of scientific image data access and management and our expertise in building tools that are useful for scientists. As is clear from our responses, we do not believe that there are single individual standards that can be used for solving the data problems in modern science. We have worked hard to develop standard interfaces that allow access to complex data types, and seen significant success with this strategy. Our experience suggests that this strategy may be deployed more broadly, in domains beyond life sciences imaging.

Different domains require different solutions, however all domains use publication of scientific results as a medium for communication and dissemination. An effective delimiter of what data should be published can be related to the content of a published paper—if the data is directly related to the results presented in a paper, the data should be published. If the data is supplementary or accessory to a publication, it may be published, but this should be optional and decided jointly between the scientist author(s), reviewer(s), and the publisher.

We are happy to follow up questions or discussions on these issues. We are excited that the problem of publishing scientific data is taken so seriously in the United States and will support these activities in any way we can.

Jason Swedlow j.r.swedlow@dundee.ac.uk
Emma Hill e.e.hill@dundee.ac.uk

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Comment 1: Fundamentally we believe that the publication of scientific data is one of the most important parts of the scientific enterprise. We do not believe that all data generated by scientists in most fields should be published. How much data should be published will vary from field to field depending on individual experiments and the existing scientific culture. One common denominator across all scientific domains is the process of publication in peer reviewed journals or online facilities. Publication represents a determination by the scientist authors, reviewers, and editors that a body of work should be delivered to the community for dissemination and consideration. Following this well-established principle, data associated with experiments reported in a publication should be publicly available.

This policy achieves two things. First, it ensures that data associated with a publication is available to the community. Second it provides a convenient definition of what data should be publicly available and what data is probably supplemental or not necessary for publication. Certainly most experiments generate substantial amounts of data that for whatever reason are supplementary or maybe not even sufficient for future analysis. In our view, this less useful data should not be included in the public record, at least in the first instance, simply because we don't yet have the tools to define its status and utility in a convenient and commonly understood way. This overarching policy can be used by individual domains to define what data should be published in each field.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Comment 2: All scientific fields have wrestled with the tensions between publication and preservation of intellectual property. There is no reason to develop any new processes here but simply to use those that already exist to protect intellectual property and to support the publication of science for consideration by the community and the public. Specific licenses can be associated with the data such that scientists and/or publications are always cited if required. This issue has been resolved already for many data types (genes, structures, images at the JCB DataViewer). Data can be accessed, to some extent analysed and also downloaded, and is available for re-use and distribution under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 Unported license.

Funding agencies and scientific research institutions already have policies associated with any intellectual property that results from research they have supported. There are already mechanisms to protect patentable findings if necessary. Data publication can use these same mechanisms.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Comment 3: Different fields have different community standards defining what amounts to an individual publication, and Federal agencies can take their cues from these communities to define what is necessary in each field. If and when a scientist publishes something as a result of their analysis of some digital data, ideally those data should be made available.

Federal agencies can support this diversity (and perhaps slowly drive consensus) by funding development of software tools that access and use this data. This investment will help energise the use of data, and very likely help define what can actually be done with the data.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Comment 4: Consider the costs of a scientist having to run an experiment or create a clone/antibody/etc. from scratch versus the cost of their being able to obtain this from someone who has published some results of that experiment or details of such a clone/antibody. The cost reduction of simply being able to acquire these data is easily apparent. The reduction in up-front costs releases funds for other avenues of research, and the benefits are obvious. In the long-term if there were better locations for all data to be housed and subsequently accessed this would reduce the burden on scientists time and costs associated even further.

An additional consideration is that currently in many research locations the steward of data is the person who produced it rather than any central location within that lab or university. As students and RAs leave laboratories, data is often lost or no longer accessible even to the person who led the research. Submission of data to a central repository serves an important archiving process, and reduces the risk of lost data due to hardware failure and simple inability to properly manage data.

Availability of digital data opens the data up for analysis by people in unrelated fields. For example, computer scientists who work on feature recognition in images might produce advanced new algorithms only if they have good exemplar data sets for development and testing. This applies to both academic and commercial settings, and could serve as a major boost to the US economy.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Comment 5: Stakeholders can best contribute to the implementation of data management plans via the provision of the relevant and necessary infrastructure. Laboratories, Universities and research organisations could resource core facilities to provide researchers with an easy way to access and archive their data. Most scientific publishers already contribute to the implementation of data management by acting as the enforcers of most current policies by making sure data are deposited before publication can proceed. One journal, The Journal of Cell Biology, has also worked to create their own database to house original image data relating to the manuscripts that they publish. While individual journals can do this, it does not seem optimal and as is done for other data types like protein structures etc. this might be better housed in one centralised repository.

Other critical resources that must be developed are tools for accessing public datasets. These are not simply databases, but full-fledged applications that provide access and analysis of data. OME is an example of such a project funded by charity and government research organisations that works with many different entities and develops tools for the community-- thus the investment from funding bodies has been returned to the community for tools for data publication. We note that OME is developed by an active team of scientists and expert software developers. We believe this is critical to the development of tools that are on the one hand well-designed software, and on the other, useful for scientists.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Comment 6: In our experience no single entity can take on all of the aspects of managing scientific data. Laboratories, departments, universities, publishers and funders all have a role to play in this process. To date, significant discussion has gone forward but relatively little action has occurred. However, the most important changes have occurred when funding bodies, e.g., The Wellcome Trust and the NIH, have unequivocally demanded that the outputs of their research be deposited in publicly available repositories. These actions created more subsequent action than any other previous discussion or policy process. They also forced the funding agencies to contend with some of the consequences, e.g., the establishment of PubMed Central. Thus the most important funding mechanism is a strong, definitive and unequivocal policy statement from funding bodies requiring the public release of data. This statement, backed up with resources to develop the necessary tools, is the action that will change the way scientific data will be handled.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Comment 7: As funding agencies review their repertoire, it would be relatively easy to check whether relevant data detailed in any publications resulting from the research funded have been made available. For example, each publication now has a unique DOI. It seems relatively easy to develop DOIs for individual datasets that can be reported by investigators in their

funding reports, follow up studies etc. Generation of DOIs is automated and can be extended to individual datasets.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Comment 8: We believe a 'bottom up' approach is the most effective response for this question. Funding bodies should be ready to invest in the outputs of publicly available research data. The proper investments will appear once the data is available. As an example Google didn't develop their search expertise and then wait for the World Wide Web to appear. First the data was available, second a number of people tried to solve the search problem and failed (rather spectacularly, because the problem was hard), and third a great solution and a great US company appeared with a new method—based on publicly available data—and soon thereafter, a new, transformative commercial model.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9: As mentioned above, use the same existing publication and DOI principles that are already well established across all scientific domains.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Comment 10: These standards must be defined by the community of scientists who generate a certain kind of data perhaps in conjunction with the manufacturers and vendors of the instruments used.

In OMEs experience spending significant effort on defining a single data standard is rarely successful and ultimately self-defeating. The development of new technologies happens so quickly that any data standard is rapidly rendered obsolete. Moreover, many technologies are made available by commercial companies whose compliance with specifications-- even those like MIAME (or DICOM, etc.)--is relatively inconsistent. In OME we have taken the strategy to provide an open specification, OME-TIFF, which many commercial providers now use (however, when examined in detail, their compliance is rarely complete). However, our most successful tools are not common file formats but software that provide a common interface to the wealth and breadth of different data types. These strategies are embodied in our image access tool Bio-Formats and our data management application OMERO. This is a pragmatic and flexible approach that recognises that the change of pace of scientific data applications cannot be limited by a data standard—new technology must be able to record the data it produces in whatever form. Software tools can be adapted to read this new data, and if designed correctly,

used by any processing algorithm to access that data. Thus, a relatively small investment in tools that can access the data and keep up with the rapidly developing data generation systems is probably the best strategy for driving digital data standards. Our mantra: don't standardise the data (it's impossible anyway), standardise the interface to the data.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Comment 11: In OME we have taken an alternative approach to providing standardised access to data. On the one hand we have provided a common data specification, OME-TIFF recognising that the specification will always be behind the cutting edge of data generation. In fact some of the commercial companies who market the microscopes have now also adopted the OME-TIFF format as one of the options for data to be stored in directly from the microscopes. On the other we have built open-source data access libraries that anyone can use. Critically Bio-Formats is built through the contributions of the community: users submit data that should be supported, we reverse-engineer the file formats no matter their source and then-- usually within a day or two-- release software that reads the data. Currently OME holds about 47,000 submitted datasets from around the world. Bio-Formats is installed and running at >37,000 sites worldwide and is started many 1000's times each day.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Comment 12: When driven by the scientific community, these standards already often involve scientific representatives from multiple countries. In our experience the most important contribution of federal bodies would be the commitment to developing tools like Bio-Formats and OMERO that provide standardised access to data as opposed to standardised file formats. This is a relatively small investment of a few FTEs per year who build and maintain these tools that has huge impact that crosses national boundaries.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Comment 13: This already works very well for several kinds of data for which there are already well-developed and supported publicly available databases for data to be deposited into (such as genomic DNA and protein sequences, solved protein structures, etc.). Thus the most important practice is the generation of unique identifiers that define individual datasets using the well-established DOI system. While not ideal it certainly is a system that is established, accepted and can be rapidly deployed. These can then be used for monitoring and verifications.

In our experience with OME and the JCB DataViewer probably the most important practice and policy is to accept that the data publication problem is in fact quite challenging. The community has been discussing data publication and data release now for many years. No individual

solution as built today will satisfy all necessary requirements, but developing and deploying these solutions in steps has multiple benefits. It engages with the community and begins the process of training the community to publish their data. It helps develop new tools and identify the true problems and bottlenecks. It provides the technical solutions to problems as they come up. Most importantly it begins the process of making data available and delivering data to the community.

Greetings!

Please find my response to the above referenced RFI:

******

Patrick Durusau
patrick@durusau.net

Patrick Durusau (consultant)

Covington, Georgia 30014

Comments on questions (10) – (13), under "Standards for Interoperability, Re-Use and Re-Purposing."

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

The goals of interoperability, reuse, and repurposing of digital scientific data are not usually addressed by a single standard on digital data.

For example, in astronomy, the FITS (http://en.wikipedia.org/wiki/FITS) format is routinely used to ensure digital data interoperability. In some absolute sense, if the data is in a proper FITS format, it can be "read" by FITS conforming software.

But being in FITS format is no guarantee of reuse or repurposing. Many projects adopt "local" extensions to FITS and their FITS files can be reused or repurposed, if and only if the local extensions are understood. (Local FITS Conventions (http://fits.gsfc.nasa.gov/fits_local_conventions.html), FITS Keyword Dictionaries (http://fits.gsfc.nasa.gov/fits_dictionary.html))

That is not to fault projects for having "local" conventions but to illustrate that scientific research can require customization of digital data standards and reuse and repurposing will depend upon documentation of those extensions.

Reuse and repurposing would be enhanced by the use of a mapping standard, such as ISO/IEC 13250, Topic Maps (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38068). Briefly stated, topic maps enable the creation of mapping/navigational structures over digital (and analog) scientific data, furthering the goals of reuse and repurposing.

To return to the "local" conventions for FITS, it isn't hard to imagine future solar research missions that develop different "local" conventions from the SDAC FITS Keyword Conventions (http://www.lmsal.com/solarsoft/ssw_standards.html). Interoperable to be sure because of the conformant FITS format, but reuse and repurposing become problematic with files from both

data sets.

Topic maps enable experts to map the “local” conventions of the projects, one to the other, without any prior limitation on the basis for that mapping. It is important that experts be able to use their “present day” reasons to map data sets together, not just reasons from the dusty past.

Some data may go unmapped. Or should we say that not all data will be found equally useful? Mapping can and will make it easier to reuse and repurpose data but that is not without cost. The participants in a field should be allowed to make the decision if mappings to legacy data are needed.

Some Babylonian astronomical texts(http://en.wikipedia.org/wiki/Babylonian_astronomy) have survived but they haven't been translated into modern astronomical digital format. The point being that no rule for mapping between data sets will fit all occasions.

When mapping is appropriate, topic maps offer the capacity to reuse data across shifting practices of nomenclature and styles. Twenty years ago asking about “Dublin Core” would have evoked a puzzled look. Asking about a current feature in “Dublin Core” twenty years from now, is likely to have the same response.

Planning on change and mapping it when useful, is a better response than pretending change stops with the current generation.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The work of the IAU (International Astronomical Union (http://www.iau.org/)) and its maintenance of the FITS standard mentioned above is an example of a successful data standard effort.

Not formally part of the standards process but the most important factor was the people involved. They were dedicated to the development of data and placing that data in the hands of others engaged in the same enterprise.

To put a less glowing and perhaps repeatable explanation on their sharing, one could say members of the astronomical community had a mutual interest in sharing data.

Where gathering of data is dependent upon the vagaries of the weather, equipment, observing schedules and the like, data has to be taken from any available source. That being the case, there is an incentive to share data with others in like circumstances.

Funding decisions for research should depend not only on the use of standards that enable sharing but awarding heavy consideration on active sharing.

(12) How could Federal agencies promote effective coordination on digital data standards with

other nations and international communities?

The answer here depends on what is meant by “effective coordination?” It wasn't all that long ago that the debates were raging about whether both ODF (ISO/IEC 26300) and OOXML (ISO/IEC 29500) should both be ISO standards. Despite being (or perhaps because of) the ODF editor, I thought it would be to the advantage of both proposals to be ISO standards.

Several years later, I stand by that position. Progress has been slower than I would like at seeing the standards draw closer together but there are applications that support both so that is a start.

Different digital standards have and will develop for the same areas of research. Some for reasons that aren't hard to see, some for historical accidents, others for reasons we may never know. Semantic diversity expressed in the existence of different standards is going to be with us always.

Attempting to force different communities (the source of different standards) together will have unhappy results all the way around. Instead, federal agencies should take the initiative to be the cross-walk as it were between diverse groups working in the same areas. As semantic brokers, who are familiar with two or three or perhaps more perspectives, federal agencies will offer a level of expertise that will be hard to match.

It will be a slow, evolutionary process but contributions based on understanding different perspectives will bring diverse efforts closer together. It won't be quick or easy but federal agencies are uniquely positioned to bring the long term commitment to develop such expertise.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Linking between publications and associated data presumes availability of the associated data. To recall the comments on incentives for sharing, making data available should be a requirement for present funding and a factor to be considered for future funding.

Applications for funding should also be judged on the extent to which they plan on incorporating existing data sets and/or provide reasons why that data should not be reused. Agencies can play an important “awareness” role by developing and maintaining resources that catalog data in given fields.

It isn't clear that any particular type of linking between publication and associated data should be mandated. The “type” of linking is going to vary based on available technologies.

What is clear is that the publication its dependency on associated data should be clearly identified. Moreover, the data should be documented such that in the absence of the published article, a researcher in the field could use or reuse the data.

******

Please don't hesitate to contact me if expansions or further explanations would be helpful.

Hope you are having a great day!

Patrick

--
Patrick Durusau
Chair, V1 - US TAG to JTC 1/SC 34
Convener, JTC 1/SC 34/WG 3 (Topic Maps)
Editor, OpenDocument Format TC (OASIS), Project Editor ISO/IEC 26300
Co-Editor, ISO/IEC 13250-1, 13250-5 (Topic Maps)
OASIS Technical Advisory Board (TAB) - member

PURDUE
UNIVERSITY

OFFICE OF THE PROVOST

January 12, 2012

Science and Technology Public Office
National Science and Technology Council
Interagency Working Group on Digital Data

Via Electronic Mail

Dear Council Members,

Purdue University is pleased to respond to the National Science and Technology Council's Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research. This is an extremely important issue for funding agencies and the research community. We at Purdue have made significant contributions to assessing and responding to the need for access to and preservation of digital data and the concurrent impact on advancing research.

Purdue University's mission is to serve global citizens in three ways: learning, discovery, and engagement. We appreciate the opportunity to engage the Interagency Working Group in a collaborative effort to investigate and assist in applying standards and metadata to research data collections. We believe strongly that the dissemination and long-term stewardship of digital data is critical to the academic fabric and that continued progress should be made to further this effort.

We would welcome the opportunity to answer any questions you may have or provide additional information at your request.

Sincerely,

Tim Sands
Executive Vice President for Academic Affairs and Provost
Basil S. Turner Professor of Engineering

Hovde Hall of Administration, Room 100 ▪ 610 Purdue Mall ▪ West Lafayette, IN 47907-2040
(765) 494-9709 ▪ Fax: (765) 496-2031

Purdue University Response to

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Purdue University submits the following to help inform the deliberations of the National Science and Technology Council's Interagency Working Group on Digital Data. We believe that dissemination and long-term stewardship of digital data is not keeping pace with the development and application of research outputs enabled by emerging Cyberinfrastructure (a.k.a. e-Science), and that this disconnect can significantly impact the competitiveness of the United States.

There is an obvious connection between discoverability, access and re-use of data, and these are directly supported and impacted by interoperability and preservation. To be accessed (by machine or human), an object must be discoverable; to be discovered, it must be identified; to be identified it must be described. Files must be documented to allow for or provide operability (and thus interoperability), and they must be preserved to be accessed, etc. An example of a broad approach which seeks to provide discoverability, access and preservation is the Purdue University Research Repository (PURR: https://research.hub.purdue.edu/), developed on the successful HUBzero™ platform (http://hubzero.org/). An example of a successful discipline-specific Purdue-led effort is the "Data Warehouse" of the Network for Earthquake Engineering Simulation (NEES at http://nees.org).

Best practice should include the application of metadata, standards and documentation to facilitate these connections to extend the life and breadth of science, but they too have not kept up with the pace of research accelerated by emerging Cyberinfrastructure. Policies and mandates can go only so far to enforce best practices, let alone conformity—norms come from within a community, not from without. The history, development and adoption of a highly successful data standard Crystallographic Information File (CIF) that exemplifies such communal consent on practice is detailed in *International Tables for Crystallography, vol. G*. This volume serves as primary guide and reference for crystallographers; could it, and the history behind it, serve as an example for other scientific communities?

If and where communities don't form or come together, and even for those who have, librarians are emerging as partners to investigate and help apply standards and metadata to research data collections. Even with such advantageous partnerships a determination must be made of where to focus to "get the most bang for the buck." Libraries have been good at preserving and providing access to content, but preservation without discoverability is basically a dark archive. Addressing these issues will take a collaborative effort among a number of constituents to provide a useful and sustainable approach to discoverability, access, interoperability, re-use/re-purposing and preservation, of one of great national treasures, research data and information.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Researchers have asked why the NSF "mandate" (http://www.nsf.gov/bfa/dias/policy/dmp.jsp) doesn't include specific requirements, standards or criteria for disseminating data. Obviously, there is no one-size-fits-all approach that would satisfy all domain science needs, let alone the demands of various economic needs. If, however, at a minimum, there were guidelines for general attributes (e.g., discovery metadata), a criterion might be developed that could make it easier for entrepreneurs, small businesses and start ups to access research results more easily and readily.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Universities/libraries employ copyright officers who advise on intellectual property and open access policies, as well as other related initiatives. As the experts "on-the-ground", it is quite possible they could provide the best intermediation locally, as state laws—and cases tried in state courts—may well have an impact in this area.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Interagency working groups should engage multiple, diverse, and prominent scientific societies/communities during the development of the policies. The effort should begin by assessing functional examples already developed by diverse communities to identify common classes of digital data and long term issues of accessibility associated with each class. This must be an interagency effort as the existence of multiple agency-specific standards will exacerbate cost and compliance issues.

The diversity of data, data formats, data representation, data reuse, and desirable preservation periods is extremely large across these diverse communities. Common sense asserts, and several examples show, communities that develop their own solutions are successful, though often only within those communities. Compliance requirements and guidelines would likely succeed only if they focus on overarching discoverability (e.g., by developing the equivalent of "Google Scholar" for data repositories and encouraging individual repositories to provide the metadata needed for discovery). It is likely that interoperability, and thus use, would follow when those who need the data developed conversion or emulation tools.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

First and foremost, the cost model of long term stewardship needs to be decided. There may be multiples one, but in a university setting the subsidy model is most likely to be successful. Just as other essential resource costs are allowed as part of the overhead computed in overhead (i.e., F&A), data management must be as well. This would likely require an acceptable increase in what constitutes the percentage allowed for institutional overhead.

It might also depend on how "different types" of data are defined. Large vs small? Numerical vs image? Disciplinary vs interdisciplinary? Not that the funding model would necessarily change, just that it might affect implementation. It will take time to study needs and requirements of long-term stewardship of data—a one or two year study of the cost model would not suffice. ARROW (Australian Research Repositories Online to the World, http://arrow.edu.au/) provides a historical example of how repository projects grow and evolve, but even this stellar example has not been able to measure cost benefit in the short time it has been functioning (since 2008). It would probably take a "decade-of-data" project, monitoring the cost model for research data for ten years, to be able to do reach useful conclusions.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Research organizations will initially be the entities to best create individual data management plans and repositories to accommodate variability in local environments. Mutual discovery of these repositories is desirable; common metadata agreements will help discoverability.

And it takes a strong collaborative approach locally. Purdue's approach is a collaboration between University IT (ITaP), Administrative Research Office (OVPR) and the Libraries. Purdue University Research Repository (PURR) is a system for discovery, developed on the HUBzero™ platform—it allows researchers to initiate projects, get help with data management, "publish" data sets, and send data to a preservation environment for long term archiving. Researchers can initiate a project and utilize resources available on the hub (guides, tutorials, videos, etc.), or the Pre-Awards office can notify subject librarians who can provide domain specific expertise on data/metadata standards, discovery, preservation, etc. Assistance in the form of data reference is available both for data management planning on proposals to be submitted, and for implementation of data plans of awarded grants. When it comes time to publish, data collections are curated to ensure metadata for discovery, archiving and preservation (per the OAIS repository model) are attached to the data set(s). This local solution, although likely applicable to other environments, ensures implementation of data management plans.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

As noted in #4, a cost model that evolves out of overhead is key. But in the near term, costs will be difficult to assess and standardize. It might ease the burden of unforeseen costs to data generators and data archives if agencies provided opportunities for supplemental funding awarded through expeditious administrative review of standardized requests.

Additionally, according to a UK JISC funded report issued in2010 ("Keeping Research Data Research" http://www.beagrie.com/krds.php), of the five major staff cost categories associated with data repositories, "activities leading up to and including ingest of the materials into the archive collectively account for 55%" whereas "the process of actually preserving the materials (archive category) accounts for only 15% of total staff costs." Improving ingest would address the first costly part of that equation—the Australian National Data Service's ARROW project "has been very successful in providing tools to enable accessibility and discoverability of research from institutional repositories" and has developed an integration of such tools and systems. Funding to develop tools and protocols for general ingest, discoverability and accessibility could be leveraged with domain specific efforts (e.g., DataONE https://dataone.org/ or Dryad http://datadryad.org/).

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

According to an unpublished report at Purdue University, compliance requirements such as those set out by OMB Circular A110 (section 53) are not known or understood by many researchers. Thus, communications to the research community must provide plain-language descriptions of minimal standards applicable to an award and examples or case studies that highlight common issues of non-compliance.

Few universities have policies which address stewardship of research data, although some have guidelines and/or best practices. Usually data stewardship and access are meant to be addressed as part of Good Lab Practice. Researchers believe they have viable organization, and that they are willing to share when asked. Compliance is usually asserted only under conditions in which there questions about research projects that are not easily resolved by a researcher (e.g., a graduating student switches labs, a co-PI transfers to another university, a corporate entity makes a request for data, or misconduct charges are made)—in other words, after the fact or in a punitive mode. It might be helpful if compliance emphasized the requirement to share as part of the research cycle, not as an afterthought. The NSF "mandate" starts to get at this by requesting "policies for access and sharing," but it might be that protocols or procedures for access and sharing would address both the spirit and the letter of compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

One model might be to bring together stakeholders to identify how to create greater collaboration for cooperation. Per # 6 above, investigating models to increase discoverability, access and ingest to get more research data into information streams could stimulate greater use. One approach might be to leverage the current efforts of libraries to assist researchers with data management planning. Libraries focus on organization, description and access to information and can help develop systems, tool and services to integrate general and domain specific metadata into research data workflows. While it is not clear how large a role that libraries will play in preserving data, it is clear they have a lot in the way of knowledge and experience to contribute. They could play a role in a larger chain of libraries-to-discipline repository pathway model. Further, bringing information profession from the academic world and the corporate world to find commonality might increase accessibility.

Additionally, everyone should be award. As noted by Prof Rudi Eigenmann, co-PI for IT of NEES: "Many efforts are under way to develop data management agreements across communities. Agencies could help by providing a directory of such efforts. Efforts we are aware of include (i) the Workshop on Data Curation and Sharing Cyberinfrastructure for Earthquake Science (http://nees.org/resources/2787/download/NEESDataWorkshopReport.pdf) and follow-on activities and (ii) data-related efforts of the annual NSF Large Facilities Workshop."

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Data volumes should become citable items equivalent to scientific publications. Awareness must be raised in scientific communities. A prominent statement about the importance of citable data volumes as evidence of research achievements for use in tenure committees will help (similar to *Evaluating Computer Scientists and Engineers for Promotion and Tenure*, by D. Patterson, L. Snyder, and J. Ullman)

The development of consistent practices for data citation is essential in an environment in which research data is openly shared. This is critical in allowing secondary analyses of data to be trusted or reproduced, and to be an accepted part of the scientific record. Just as important, mechanisms supporting appropriate citation and attribution of data is important in incentivizing data sharing by data producers. One part of this is the development of standards and best practices regarding the identification of data sets. This need goes beyond the collection of basic descriptive information such as author, title, and version, but also addressing much more complex issues such as the citation of dynamic and longitudinal datasets, and the appropriate levels of granularity for citation.

The second major component in providing effective mechanisms for citation and attribution is the usage of persistent, universally-resolvable identifiers that provide consistent and accurate access to cited data. The work of DataCite (http://datacite.org/), including three U.S. members (California Digital Library, Purdue University Libraries, and the United States Department of Energy Office of Scientific and Technical Information) is addressing many of these issues and beginning to make strides in the development of actual mechanisms for supporting data

citation. This group has created a metadata kernel for citing data sets, and has also developed infrastructure for the assignment of DOIs to data sets. At an international level, DataCite has registered DOIs for over 1 million data sets, facilitating citation by scholars but also facilitating discovery through search engines and scholarly indexes.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

It would likely take such MIAs for many/most/all disciplines to allow easy cross walking between disparate science communities and to accomplish more generalizable interoperability. A common complaint is that researchers don't have the time to integrate new procedures into workflow. It is unlikely there could be an uber-MIA standard; but if there was one (e.g., to at least initially address discoverability), it would likely be effective to develop training programs for graduate students, post-docs, etc. in how to apply and use it.

See also # 3 above on "Preservation, Discoverability, and Access".

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Network for Earthquake Engineering Simulation (NEES) has developed data standards and management plans for a network of 14 diverse laboratories that facilitate earthquake and tsunami research. This was achieved in collaboration of a community-led data advisory committee and the NEES cyberinfrastructure IT development team.

A cross-disciplinary and applied model has worked well in the library science domain. The Online Computer Library Center's (www.oclc.org/) shared cataloging model has facilitated application of standards and metadata for cataloging millions of objects (books, journals, reports, etc.) internationally since 1971. The model is a cooperative approach to applying standards to resource descriptions, where participating libraries each contribute some effort and the results are shared (discovery and access) with all. There may be away to emulate this model for distributed application of standards and metadata for data collections. A protocol such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), which provides machine-to-machine collection of metadata so that services can be built, could underlie and support such a "shared cataloging" model.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

As noted, DataCite members are developing standards for persistence and citation via "DOIs for data." The DataCite Metadata Schema is "a list of core metadata properties chosen for the accurate and consistent identification of data for citation and retrieval purposes." ORCID (Open Researcher & Contributor ID http://orcid.org/) is an emerging international standard which will "solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers." Name authority is crucial for all scholarly contributions, including data.

There are currently avenues by coordination of these efforts could be promoted. CENDI is an interagency working group of senior scientific and technical information (STI) managers which collaborate to address issues related to federal information policy and to help improve science- and technology-based programs, operations and systems (http://www.cendi.gov/). CENDI coordinates related conferences and workshops, and facilitates a range of "interest areas" (i.e., interest group) such as Digital libraries and Information Policy. The Secretariat is headed by an executive director who has vast experience supporting government and industry in managing information as a strategic resource. Perhaps this group could provide recommendations for coordination on digital data standards, national and internationally.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

The continuing work by the NISO/NFAIS Supplemental Journal Article Materials Project (http://www.niso.org/workrooms/supplemental) provides important technical guidance on policies, practices, and standards needed to support linking between publications and the associated data, and is realistic about what is possible within current publishing workflows. However, any discussion that still classifies data as "supplementary" may already be behind the times, as interpretative text and the data it interprets become increasingly intertwined. Data papers are already starting to appear that present the data as the main scholarly output, with supplementary textual documentation and interpretation. In this environment, the main standards need is that Digital Object Identifiers (DOIs) are used and are available in a model that allows the researcher to apply them at the level of the "smallest citable unit." The level of granularity desirable will vary from project to project, but may include assigning DOIs to different protein structures, for example. Universal use of DOIs that are made available in reasonably unlimited quantities, not priced per DOI, is the essential characteristic of an effective system of publication and data links. Although pricing practices are constantly adjusted, the DataCite model of issuing DOIs has an advantage over that of CrossRef in this regard.

Practically speaking, local stakeholders who could have a part in data sharing— researchers, archivists, publishers—primarily work in environments in which they act independent of others, especially when it comes to doing something more or different with data. Librarians and publishers often have a common tie as customer and vendor, but otherwise act independent in any activities related to data. Researchers and publishers work together to make data available in some cases, but it is rarely a primary concern. And yet, because each of these entities has

some role in the curation, sharing and reuse of data it would seem possible that they could collaborate to make data better available. One approach might be to:

- Identify and analyze successes of exemplar stakeholders of publishers, funders, librarians, and researchers, who have developed pilot programs and undertaken digital curation and publishing, such as Dryad.
- Investigate and establish a crosswalk of concepts and terminology across stakeholder domains, mapping conceptual abstractions and fundamental terminology to form a framework that could contribute to a model of collaboration in this area.
- Utilize the framework to initiate further cooperation between stakeholders to cross-link supplemental data to journals and repositories using standard formats, identifiers, protocols, and supporting metadata.

Contributions provided by:

- Jeffrey T Bolin, Professor of Biological Sciences, and Associate Vice President for Research
- Paul Bracke, Assoc Professor of Library Science, and Associate Dean for Digital Programs and Information Access
- D. Scott Brandt, Professor of Library Science, and Associate Dean for Research, Libraries
- Rudolf (Rudi) Eigenmann, Professor of Electrical and Computer Engineering, and Co-PI for IT for the Network for Earthquake Engineering Simulation
- Eugenia S Kim, Visiting Asst Professor of Library Science, and Data Services Specialist
- Charles Watkinson, Director, Purdue University Press

Responses to the 11/4/11 OSTP Data RFI
from the Science and Technology Office
by David B. Lowe, Preservation Librarian
<david.lowe@uconn.edu>
University of Connecticut Libraries, Storrs, CT
1/12/2012

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?	Federal policies and programs that sponsor or otherwise facilitate the **creation and maintenance of discipline-specific repositories for research data** could encourage public access as well as preservation. Progress on this effort would need to start with a proper comprehensive inventory of such repositories, perhaps followed by a certification or at least vetting and recommendation process on behalf of federal funders, ideally involving professional associations and societies. Where the inventory reveals gaps in the spectrum that existing repositories cover, it would then be proper and desirable for federal funding to attempt to foster initiatives to cover these lacunae. The reason for this need is related to a lesson that libraries and archives have learned over the millennia of gathering and organizing information objects of cultural significance: collections contain context. Context is crucial in identifying knowledge entities relative to one another, establishing hierarchies, and assigning priorities that are prerequisites for progress with scientific methods in particular, not to mention with any intellectual endeavor in general. Research assumes such structures as a basis upon which to progress and build further, relying on the credibility and veracity of past work, and our new digital environment should not be an exception. In addition to the context intrinsic to a collection, its niche market tie to its clientele is also critical, in that—out of all the knowledge resources in the world—it can be positioned closer to those who are most familiar with it. This will be especially

	important in the future as we confront migration issues. The best strategies and solutions will involve knowledgeable users who are closest to the content and who can help ensure its viability into the future.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?	Access controls that allow in vetted researchers during an embargo period if they agree to respect intellectual property constraints via attribution and citation could be a solution for stakeholders' concerns about their data in a repository as mentioned in the response to question #1. From talking with researchers as part of formulating our institution's response to the NSF Data Management Plan requirement and also as part of an eScience Institute sponsored by the Association of Research Libraries (ARL), I understand that the most common model that has developed in the academic community over the years has been one of willingness to share data when asked. As we transition to better infrastructure for preserving and also sharing via open access, the part of that established model that could be lost would be the fact that the researcher and the requester are aware of each other, at least at some minimal level of identity, in a relationship of professional trust. Establishing a certain reasonable embargo period, matched to community norms and during which this controlled access could take place, would restore some of that identity clearance, which could take place on a case-by-case basis for individuals, or could be open to established researcher groups, which in turn might have their own certification processes that their communities can trust. No one questions that federally funded research should be made public eventually, except in a relatively few cases of privacy or security, so the problems to solve revolve around the timeframe of active projects and the 3-5 year window that follows. As I write this, the news features stories about the suppression of some of the specifics of federally funded avian flu research, which is just one case related to national security. The point here is that with a robust system of access controls, researchers who legitimately need to know these details could get what they need, while those who lack proper credentials would be denied access.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?	Although inherent differences between disciplines do pose a problem for those seeking to establish equitable data management expectations in the grant funding context, there are some clear watershed areas for distinguishing between groups. One of the most important litmus tests involves research data that needs to contain personally identifiable information (PII) and also data that has sensitive implications across a broad spectrum of security issues. Fortunately, these two areas of PII and security tend to be governed by other rules that can take precedence over federal grant funding guidelines. A second area of concern would be the "haves vs. have nots" in terms of adequate repositories, which my response to question #1 above attempts to address. Until and unless there is an appropriate place for a researcher's data, it does not seem fair to ask her to meet the same requirements for deposit as those who already have adequate places to park those files.
(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?	Long-term usefulness cannot be immediately known or quantified, but the historical lesson that libraries teach us is that information kept just-in-case does in fact tend to come in handy within a sufficiently inclusive timeframe. It would be short-sighted to jettison reasonably retainable data now just because we make some capricious determination that it will be of no use down the road. Tossing information out guarantees that it can be of no help in the future. Also, crosswalking data sources to make them searchable for cross-disciplinary purposes will greatly enhance their potential usefulness as a benefit for all.
(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?	The top priority in contributing to the successful implementation of data management plans is the establishment of an adequate repository infrastructure, especially the core metadata ecosystem that makes ingest, management, discovery, access, sharing, and preservation all feasible.
(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?	Cost-shared efforts for archiving the data produced in grant projects deserve to be weighted more heavily than a 1:1 dollar value. Related service costs should be given higher value and consideration than those traditionally featured in that grant proposal budget column, at least in these early stages when we are attempting to establish adequate workflows. To be more explicit, a project that follows its discipline's established metadata schema has less preservation work to do than a project for which metadata schema development is lacking. Any schema development done within a project, then, deserves to be incentivized.
(7) What approaches could agencies take to	By standardizing repositories and automating their functions, management issues like metrics, verification, and compliance checking would become vastly easier. We need

measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?	to raise expectations in these areas and put the mechanisms in the right places to accomplish these goals.
(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?	After funding proper repositories, next logical steps would be data mining and presentation projects. The pent-up wealth of information could give rise to new fields and specializations that dig into the fabric of the information assembled and cull from there patterns that in turn spawn a demand for eyes and hands that can present the new findings in visually stimulating and meaningful ways, not to mention then applying the knowledge then revealed to the real world to make our lives, our cities, and our societies better.
(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?	Access control mechanisms could enable tracking that conveys full credit and attribution. Components of such controls would include better universal identifiers for people, institutions, publications, and parts thereof. There are significant researcher privacy concerns here, but certainly many would be open to an opt-in identification model if it also made their citation work easier through automation. For the rest, there is no perfect solution to plagiarism and theft, but at least it may be easier to deny access to known past offenders if adequate controls are in place.
Standards for Interoperability, Reuse and Repurposing	
(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.	Any standards that the respective communities develop are the right ones. It is always the hands-on users who should make that determination. This is not to say that we cannot do a better job of aligning variants within a discipline or of making cross-disciplinary standards more compatible with each other, but the specialists should always decide about the particular data points captured as a "business rule," as code developers and analysts would say. Alignment and compatibility is something that metadata librarians would be able to help with and should be involved in.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?	I would point to MARC for bibliographic information in libraries, EAD for collection finding aids in archives, DDI for social sciences data sets, and FGDC for GIS data as effective standards that have created efficiencies and opportunities for sharing. The main characteristic of their development processes is that they all achieved community acceptance above a certain threshold, which in turn made the efficiency pieces happen.
(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?	Federal agencies could promote coordination by creating funding opportunities for metadata development. Targeting disciplines that lack proper common metadata schema, agencies could offer to fund a conference to discuss community needs, with deliverables that would include draft data points toward a schema. It would not be difficult to find metadata librarians, analysts, and information architects to polish that draft into a serviceable metadata approach, and these professionals could also keep an eye out for cross-disciplinary functionality.
(13) What policies, practices, and standards are needed to support linking between publications and associated data?	The key piece for linking support lies in persistent identifier solutions. Such identifiers assume other crucial infrastructure is in place, such as proper stable repositories, so these are the most important priorities for the time being in this area of endeavor, as discussed in my responses throughout above. There is a huge role for the professional organizations to play in establishing community norms around metadata schema, discipline-specific repositories, embargoes, and access controls.

Society for Conservation Biology
A global community of conservation professionals

January 12, 2012

Ted Wackler
Deputy Chief of Staff
Office of Science and Technology Policy
Attn: Open Government
725 17th Street, NW.
Washington, DC 20502

Submitted via e-mail to digitaldata@ostp.gov

Dear Mr. Wackler,

The Society for Conservation Biology (SCB), a global community of conservation professionals which publishes Conservation Biology, among other journals, submits these comments in response to the request by the Office of Science and Technology Policy (OSTP) for input on the Administration's interest in enhancing public access to digital data generated in federally funded research. In the following comments, we borrow substantially from a draft prepared by our sister societies in the Ornithological Council, a consortium of twelve scientific ornithological societies in the Western Hemisphere and from comments submitted individually by our President, Paul Beier.

Conservation Biology is rich in data that are underutilized because they are not accessible. Decades of data are disappearing rapidly and irretrievably because the scientists who collected the data had no opportunity to archive it in a physical or electronic form. Whether on paper or in some kind of electronic medium, datasets collected over the past century could contribute greatly to our knowledge of conservation biology.

Our organization strongly supports the concept of archiving and sharing these data. Sister societies have investigated and discussed the possibility of developing an archive for the types of data generated in different forms of biological research but found that the cost is prohibitive and that it might not be realistic to expect that scientists will voluntarily undertake the somewhat burdensome effort of learning metadata standards and routinely labeling their data for deposit into an archive.

As a preliminary and key issue, we stress the need to allow researchers to have exclusive access to and use of their data for a time period sufficient to allow them to complete their publications. This time period must be flexible; in our field, long-term studies can stretch over decades. The "reward system" for scientists in both academia and in federal agencies stresses publications. The number and quality of publications is a large factor in determining promotion and tenure and strongly affects the researcher's success in

obtaining grant funding. We assume that OSTP is fully aware of the fact that the misappropriation of a researcher's data could have substantial negative impacts on the researcher's career and will take care to assure that any public access policy includes ample protections for the researcher.

As a second key issue, we would like to address something that seems to be outside the scope of the OSTP request and existing agency data management requirements, probably because it would be impossible to impose these requirements retroactively. We would like to stress that if resources are available, the government should commit those resources to help "stabilize" those data, convert them to a digital format, and submit them to appropriate data repositories. The data collected a decade ago or a century ago are, in our field, at least as valuable as the data collected today, if not more so, as these baselines are necessary to assess change. The attics full of paper, note cards, field notes; the offices full of punch cards, floppy disks, and magnetic tape – all need proper storage to guard against physical loss and all should be digitized and contributed to publicly accessible repositories. We cite the example of the North American Bird Phenology Program created by the Patuxent Wildlife Research Refuge of the U.S. Geological Survey. Using volunteers and a high-speed scanner, this remarkable program preserved six million hand-written note cards recording bird migration observations, dating back to 1881. The scanned records were then uploaded to the internet to make it possible for volunteers to enter the data into a database. The USGS and the other partners of the National Phenology Network provide analytical tools, guidance documents, and other resources. More recently, the U.S. Bird Banding Lab was able to stabilize decades of hand-written records by scanning and it is hoped that funds will be made available to make these critical data available to researchers by digitizing the data and making them available on a public access website. To date, researchers and others have been able to access these data only by making a request to Banding Lab staff who would then retrieve the physical records for copying and mailing. The records were at extreme risk of physical deterioration or loss, having been stored in a variety of facilities that were subject to rodent infestation, fire, dampness, and flooding.

Therefore, we strongly encourage OSTP to work with the Office of Management and Budget and Congress, as appropriate, to provide funding and direction to the agencies to stabilize existing physical data records, to digitize those records, and make them available on publicly accessible databases. These processes should not be limited to agency-held data but should be opened to private researchers as well.

We would also like to address certain of the questions asked by OSTP, as follows:

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Response: The key issue here is funding. Developing and maintaining these systems is costly. The intricacy involved in creating any one metadata standard is substantial. Interoperability is a daunting challenge. In our discipline, for instance, DataOne <www.dataone.org> is intended to "ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE will transcend domain boundaries and make biological data available from the genome to the ecosystem; make environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; provide secure and long-term preservation and access; and engage scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualizations." The five-year NSF grant alone amounts to $15,257,190 from the Office of Cyber Infrastructure and it is supplemented by support from the NSF Computer and Information Science and Engineering Directorate (CISE) Pathways Computational Sustainability, the NSF INTEROP Programs, NASA, the Leon Levy Foundation, the Moore Foundation and (until its recent demise), the National Biological Information Infrastructure of the U.S. Geological Survey.

The complexity of these systems requires that they be done right; if not, the end result is a system that hampers, rather than facilitates public access. The federal government must be willing to commit the resources to enable excellence or the undertaking is not worthwhile. We would have an expensive warehouse where nothing can be found, much less retrieved.

We would draw your attention to an article addressing these issues in Science Magazine. The citation, abstract and some of the recommendations follow:

11 FEBRUARY 2011 VOL 331 SCIENCE www.sciencemag.org

PERSPECTIVE
Challenges and Opportunities
of Open Data in Ecology
O. J. Reichman,* Matthew B. Jones, Mark P. Schildhauer
Ecology is a synthetic discipline benefiting from open access to data from the earth, life, and social
sciences. Technological challenges exist, however, due to the dispersed and heterogeneous nature
of these data. Standardization of methods and development of robust metadata can increase data access
but are not sufficient. Reproducibility of analyses is also important, and executable workflows are
addressing this issue by capturing data provenance. Sociological challenges, including inadequate rewards
for sharing data, must also be resolved. The establishment of well-curated, federated data repositories
will provide a means to preserve data while promoting attribution and acknowledgement of its use.

Some fields such as astronomy and oceanography have a history of sharing data, perhaps because these fields rely on large, shared infrastructure. Other disciplines, such as genomics, also have shared repositories, largely due to the homogeneity of their data. Traditionally, ecologists have had few incentives for sharing information. Research involved gathering and analyzing one's own data and publishing the distilled results in peer-reviewed journals. In addition, sharing data was not viewed as a valuable scholarly endeavor or as an essential part of doing science. Recent advances in ecological synthesis, however, are rapidly changing these attitudes to data sharing. Researchers might still be disinclined to share their data until they have fully completed analyzing and reporting on their observations and results. The concern is that if data are made openly available in the interim they may be used by other investigators, effectively scooping the data originators. Properly curated data alleviates this concern, as the use of data without permission or attribution would be condemned by colleagues and funding sources. Proper curation requires time and money and is inadequately supported in research funding.

Establishment of a reward system should further motivate investigators to share their data. For example, if data sets are publishable and citable (e.g., Ecological Archives and Dryad), they will become more respected and valued as an important part of research and scholarship (20). The most effective means to alter the reward system is to make data sharing an expectation of funding and publications and reward those who meet these expectations. The National Science Foundation in the United States now requires an explicit data management plan in all proposals, which is a step in the right direction. Journals and societies that mandate data publication concurrently with research publications also have proven to be effective (e.g., GenBank).

In addition to support for individual researchers

to prepare and submit their data to public archives, the community needs to identify sustainable models for federated data archives that persist over decadal time scales. Models such as DataONE involve leveraging institutional contributions in a large federation to protect against uneven funding for individual institutions. Nevertheless, even these initiatives will not work without a sustained commitment from funding agencies that is specifically targeted at institutional data repositories and coordinating organizations.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Assure that researchers have the ability to control the release date. Do not require release until the researcher has had adequate time to publish the research utilizing a given data set. Recognize that in some fields, research may extend over decades. For instance, studies of long-lived organisms will typically continue over the full life-cycle of the organism. A researcher will likely publish papers throughout this period, but later papers will often make use of data collected at a much earlier stage of the study. Consult with scientific societies to determine the appropriate maximum duration for the sequestration of a given data set.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Consult with the professional societies. We can provide the data and insight as to the policies and practices that will make it possible for our members to archive and share data without jeopardizing their intellectual property interests. We can also provide information about the ability of our discipline to create and maintain these repositories and the appropriate metadata standards. We can identify gaps in opportunities for data management. In ornithology, the existing repositories, though stellar, simply cannot accommodate many kinds of data collected by ornithologists. We have, as a result of the NSF data management plan, been collecting information about all potential data repositories that may be suitable for this kind of data, and we are still finding significant gaps. At the moment, NSF's data management website simply directs those who are unable to find an appropriate public repository to "Contact the cognizant NSF Program Officer for assistance in this situation." We suspect that if NSF were to attempt to compile a comprehensive list of relevant data repositories, these gaps would be quite evident.

We can also compile and provide data about the range and median grant size in our discipline. This information should be taken into account before imposing another time-consuming grant requirement on researchers. The OSTP notice mentions that the NIH requirement applies only to grants with direct costs exceeding $500,000 in a single year. In our discipline, that threshold would exclude most grants. For instance, the average grant size made by the NSF BIO program in 2011 was $149,238. In 2010, it was $140,064 <http://dellweb.bfa.nsf.gov/awdfr3/default.asp>. Most NSF grants in our discipline come from the Division of Environmental Biology (DEB) or the Division of Integrative and Organismal Systems (IOS). In DEB, the average grant in 2010 was $95,649 and in 2011, it had declined to $85,919. In IOS, the average grant size was $150,000 in 2010 and $$151,181 in 2011. Smaller grants simply do not allow the researcher to hire administrative staffers or other technicians to handle this additional work.

If no additional funding is provided, the data management requirements could constitute an unfunded mandate such as would trigger the provisions of 2 U.S.C. §1501. We recognize that the Administrative Procedure Act exempts matters "relating to agency management or personnel or to public property, loans, grants, benefits or contracts" and that therefore, a formal rulemaking as would trigger the Unfunded Mandates Reform Act (UMRA) would likely not occur. Nonetheless, the agencies have made it a practice to use notice-and-comment procedures outside the Federal Register process for this and other policy matters. These quasi-rulemakings should be regarded, for the purpose of the required UMRA analyses, as the equivalent of a rulemaking. Therefore, any agency that wishes to mandate data management should be required to conduct an "UMRA-like" analysis to assure that the requirements are the least costly, least burdensome, or most cost-effective option that achieves the objectives of the rule, or explain why the agency did not make such a choice (2 U.S.C. §1535).

The scientific community should also be consulted with regard to the release of certain types of data. For instance, we have long been concerned about the potential online, public access release of location information associated with bird banding. Some of the birds banded are, of course, legally protected at the federal or state level. Information about the location of banding could facilitate activity that is prohibited under the Endangered Species Act. Other species, protected only under the less comprehensive prohibitions of the Migratory Bird Treaty Act, are very vulnerable to disturbance during the breeding period. If the public could use the location data associated with bird banding to determine breeding locations, the disturbance resulting from human presence could lead to failed breeding attempts. This outcome would contradict Executive Order 13186.

As noted, some data could be used by unscrupulous persons to kill, capture, or harm individual animals or plants. Many agencies (Arizona's Heritage Database Management System, for instance, and other state databases) have largely solved this problem, however, using two simple measures: (1) The publicly available data consist only of low-resolution maps with locations "fuzzed" by up to a few km. This provides enough preliminary information for a potential user to determine if the data cover the area of interest to the user. (2) Precise location data are provided only to legitimate requestors

who agree to specific terms on use of the data, including agreements not to depict or share precise locations in any way.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

For species occurrence data, the costs are miniscule and the benefits are large. We suggest that OSTP might for now require data sharing only for similar types of low-cost high-benefit data. OSTP and other agencies could use the experience to start to produce reliable estimates of long term costs and benefits that could be used to guide future decisions.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

The Society for Conservation Biology publishes several scientific publications. SCB could work with our publisher to require authors to archive their species location data with appropriately coordinated repositories. However, if only SCB took this step, some authors would submit elsewhere to avoid this extra responsibility. But a broad consortium of professional societies in ecology (SCB, Ecological Society of America, The Wildlife Society) and a handful of dominant publishers (e.g., Wiley-Blackwell, Elsevier, Springer-Verlag) could create a new culture in which data-sharing is viewed as a responsibility of publishing. Our President has appointed a Task Force in SCB to investigate how SCB could start a dialogue with our sister professional societies and the publishers of their journals to start to create this culture. It will take years, and there will be strong resistance from some academic PIs, but this is an achievable long-term goal. Again, it makes sense to start with low-hanging fruit (e.g., species occurrence data); once the new culture of sharing has been in place for a few years, I think it will become obvious which other types of data to share, and how to share them.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

As noted above, grants in our field typically do not permit researchers to hire staff to undertake the work associated with effective metadata labeling and deposit of data. There is no point in warehousing data if it is not done in such a way as to make the data easily retrievable and to assure that subsequent users are able to identify the characteristics of those data so they can determine if they are appropriate for the later use. Without additional funding, data repositories are not likely to be of adequate quality and any resources devoted to them will have been wasted.

This is not a hypothetical concern. The U.S. Geological Survey devoted more than a decade of effort to develop the National Biological Information Infrastructure. It is now

being dismantled; it never began to approach the original goal of providing access to distributed data.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

For some types of data, ensuring compliance will be difficult, but it should be relatively easy for species occurrence data. Federal funders of biodiversity-related research (NSF, USDA, DOD SERDP, EPA) could require the Data Management Plan in each proposal to list the species for which occurrence data will be collected. Funders should convey this information to a repository that is well integrated with others, which would need staff persons to track compliance and report non-compliance to all federal funders.

One more drastic measure is worthy of consideration: The OSTP and OMB could set out procedures for identifying institutions with a pattern of non-compliant PIs and baring such institutions from future federal grants and contracts for a period of time. This would motivate universities and other research institutions to monitor compliance.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

By citing in their proposed and final rulemakings more thoroughly the peer reviewed journals and the data reported and analyzed therein, and by working with Congress to help their committees and the Congressional Research Service to do the same in legislative and investigative and oversight committee reports.

Also by making information that will be available publicly someday available sooner in some cases. For example, in addition to considering Federal purchasing of rights to copyrighted material, OSTP might consider working with expert Federal agencies and the Federal Office of Trademark, Copyright and Patents to determine the extent to which currently patented procedures and devices that could help solve serious societal problems, such as increasing energy efficiency and reducing pollution, or sequestering carbon with bio-char produced in biologically sound and safe ways, are being fully deployed and if not, what level of payment would be appropriate for an eminent domain-style assumption of part of or the remaining years of that patent by the Federal Government. Agencies could review indexes of patents or other descriptions of them with the help of the Patent Office. They could then ask scientific societies to help them evaluate those that might be more useful if provided to the public at an earlier point.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

For a number of years, we have discussed this very question with regard to the potential release of bird banding data. It has been the practice of the Banding Lab to interact with

those who request data and to remind them of the professional standards for attribution and credit. This interaction is possible only because data requests are made by individual contact to a staffer who then transmits the data to the requester. In fact, the Banding Lab website makes no mention of these professional standards. The U.S. Bird Banding Lab Advisory committee could not devise a more robust solution, saying that a web-based public access site should be developed and that In consultation with banders and users of banding data, review and revise the current policy for use of banding data, and require all data users to agree to this policy. The BBL should also encourage the adoption of this policy by ornithological societies and scientific journals as part of their scientific code of ethics."

The reality is that there is no effective mechanism to force users to give appropriate attribution and credit. It may be evident, given the age of the data or the geographical or temporal range of the data that the author did not collect all the data used in the paper. In those cases, editors will likely insist that the author provide attributions. However, there will be many cases where this is no evidence that the data used were collected by other than the author, and in those cases, there is really no adequate solution.

Therefore, the only means to protect a researcher who is still publishing papers based on a given dataset is to allow the researcher to determine the date of release of the data to the public, as described above, subject to standards that are appropriate to that particular discipline.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Our task force may be able to help with this soon but we have no comment on this yet.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

In the taxonomic sciences, extensive effort has gone into the development of a metadata standard known as the Darwin Core. Numerous extensions have been developed that will support the addition of "ancillary" data such as ecological conditions, and weather data. We hope that there will someday be extensions for the behavioral data that is commonly collected in biological and related research.

The use of this common metadata standard and extensions would permit interoperability with any other system that uses the same standards. For instance, the Darwin Core has led to the development of ORNIS, HerpNet, MANIS, and FishNET (birds, herps, mammals, and fishes) and these are integrated with GEOLocate, AmphibiaWeb, Map of Life, Specify, Arctos, DataONE, Encyclopedia of Life, and Animal Diversity Web.

These repositories and the metadata standards were initiated by the community and achieved with federal funding. Other organizations (most also federally funded) then built user tools and applications, such as the Avian Knowledge Network at the Cornell Lab of Ornithology. This project also received significant federal funding.

However, no amount of scientific zeal and energy can achieve this kind of result without significant federal funding. Unless the federal government is willing to continue to devote appreciable sums, the government and the public cannot expect to achieve the goal of providing public assess to data derived from federally funded research.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Response: Science knows no geopolitical boundaries. Scientists have long been working on an international basis to develop metadata standards. The Global Biodiversity Information Facility, established in 2001, already holds 8,594 datasets to which access is free and unrestricted. However, the sole U.S. representative to GBIF is a single employee of the now-terminated National Biological Information Infrastructure. The NBII termination page states with regard to GBIF that "While USGS does anticipate continued collaboration with some of these activities, we have yet to determine at what level this will occur." We are informed that it is likely that USGS will continue to participate at the minimal level (i.e., one FTE) that was the case prior to the termination of the NBII.

The federal agencies must commit to increased participation in these international bodies, and commit the necessary resources for that participation.

If the federal government is unable or unwilling to continue funding this activity at an adequate level, then it should hold in abeyance all but the most compelling and reasonable mandates that scientists submit data to any repository. If there is no assurance that the repositories will persist and will be properly managed, and that there will be a continued development of science-driven metadata standards, then the burden imposed on scientists to label their data and submit to data repositories is not warranted.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Response: The DOI (digital object identifier) for each publication should be included in the metadata associated with each data set and conversely, the location of the data should be provided in each publication.
We, and our data sharing task force, look forward to working with OSTP in the future.

Sincerely,

John Fitzgerald
Policy Director

11200 Rockville Pike, Suite 302
Rockville, Maryland 20852
USA

American Society for Biochemistry and Molecular Biology, Office of Public Affairs

BENJAMIN W. CORB
Director

JULIE M. MCCLURE, PH.D.
Science Fellow

THE NEED FOR LONG-TERM PROTEOMIC DATA STORAGE

SUMMARY

The lack of a reliable and secure repository for raw data is a major problem facing science. While there are various repositories for 'processed' information these have substantial limitations and thus only serve a portion of the need (and the community), and importantly cannot store 'raw' data. Therefore there is an essential need for such an entity. This can be accomplished by providing long term fiscal support for creating an over-arching structure, actually capable of capturing not only raw data but also various forms of processed information that would provide a central storehouse.

SUPPORTING EXAMPLE: PROTEOMICS

OVERVIEW

One of the most significant hallmarks of biomedical research in this century, and perhaps one of the most unexpected, has been the size and extent of data sets that have and continue to be generated by the new technologies associated with genomic, transcriptomic, proteomic and metabolomic research (collectively the bio-omic sciences or, by some definitions, systems biology). The microarray field that underpins transcriptomics led the way but it has been supplanted by the massive outputs of next gen nucleic acid sequencing of vast numbers of human and other genomes. However, proteomic data, mostly generated by high throughput mass spectrometry (MS), will eventually dwarf both of these and when coupled with metabolomic data that will likely be collected with similar technology, is destined to create an almost unimaginable amount of information. At issue, therefore, is how to deal with this onslaught?

Clearly the problems for the individual 'omic sciences are not the same, as the types of data are quite different (excepting proteomics and metabolomics). Germane to this report is the collection and interpretation of MS data. There are several issues and several levels of data and each requires its own consideration. For ease of presentation, MS data, in support of a proteomic (or metabolic) experiment can be classified as 'raw', 'processed' or' interpreted'. The interpreted data are suitable for publication and for inclusion in searchable web-based compendia. These are outputs of search engines, which have interpreted the processed data in the form of peak lists or spectral libraries, and can involve additional software analyses including, but not limited to, quantification and functional assessments. Journals that publish proteomic data have various requirements for what information must be included in research articles and how much of the data from which the

identifications of peptides, proteins and post-translational modifications (PTM) were extracted must accompany the manuscript (during review and/or ultimately appearing in the journal, mainly as supplemental material). The extent to which the validity of these assignments can be assessed is accordingly equally variable.

To address this issue, *Molecular & Cellular Proteomics (MCP)*, starting in 2003 and culminating in 2005 (1), developed and adopted publication guidelines for reporting MS identifications and has subsequently updated them (2). As part of this evolution, in 2010 it announced (3) that it would require the deposition of the raw MS data in a public database as a requirement of publication for all accepted papers containing MS identifications. While not mandating it as a requirement, other journals publishing in this area of research supported this policy. For all practical purposes, Tranche, founded and operated out of the University of Michigan, is the only repository capable of handling this type of data submission. Unfortunately, technical problems, due mainly to inadequate fiscal support and largely manifesting themselves in the past year, have substantially curtailed the usability of Tranche and in March of the past year MCP was forced to make raw data deposition once again voluntary. Although the situation has shown signs of improving in the last six months, there is no sustained support of Tranche that has been identified. Thus, at the moment there is not a suitable and reliable repository for raw MS proteomic data available.

Why deposit raw MS data?

There are a number of reasons for why this policy should be universally adopted. First, the interpretation of MS data depends on software analyses and there is considerable variation in the search engines, how they make their determinations, and how they decide whether a result is reliable. It is important to understand that generally less than 50% of the spectra generated in an experiment are interpreted (and sometimes considerably less than that) and that assignments are given scores that indicate the probability that the identification is right after making certain assumptions about what could be in the sample. This is compounded by errors in the databases searched and in the possibility of matching a correct sequence to an incorrect protein. This is considerably exacerbated when PTMs are involved and localizing the modification sites correctly is clearly the most challenging analysis of all. The most effective way to re-examine an assignment is to have access to the raw data. Related to this, software for processing and interpreting MS data continue to improve, so re-analysis of datasets with newer software is likely to lead to the extraction of more information from previously acquired data. However, this can only performed if the raw data is available. Second, essentially all experiments are designed and executed with a purpose, i.e. there is a biological question being addressed. This means that the data will be analyzed from the orientation of this objective, and other information present in the data set will likely be ignored or simply not identified (i.e. be part of the 50% or greater of the data that was not explained during the data analysis). In addition, quantitative information present in the raw data may not have been examined (only qualitative analysis; i.e. peptide and protein identification is performed for many datasets). In fact, it may not be possible to interrogate a data set at the time it is collected for a specific question or possibility because the requisite findings that underlie it had not been previously determined. In essence, this is a manifestation of the axiom that one “sees only what one

looks for". This is particularly true for PTM analysis, as for most datasets only a very limited number of PTMs are considered during data analysis. As a result, potential large amounts of information are not analyzed and the information contained therein lost if the raw data is not made available. This is enormously wasteful from both an intellectual and financial point of view. Finally, knowledge is a continuum and all data collected adds to it. This is particularly important to the bioinformaticians and other analyzers of processed and interpreted data, who can provide the larger prospective that helps to produce the global understanding of biology and medicine, which is the real goal of the bio-omics. By not reporting the actual data collected or placing it where it can be used by others, it defeats a major part of what experimentation is supposed to be about.

It must fairly be pointed out that not everyone is in favor of raw data deposition. Some individuals, clearly recognizing that large MS data sets have unused or undiscovered potential and not wishing to have this be exploited by others, do not want to share their raw data, rather hoping to find new things in it themselves. Others are concerned that their misinterpretations and mistakes would be made plainly (and painfully?) available for others to point out and thus for all to see. And lastly, some simply don't want to be bothered with the hassle of making the necessary uploads, which can indeed be time consuming. Although in part understandable (from the human nature point of view), none of these reasons are particularly compelling or scientifically and fiscally well justified.

What is needed?

It should be made clear that the shortcomings of Tranche are basically due to lack of support rather than any inherent design flaws. It was created as an academic exercise and was largely supported initially by grant funds. When these were ultimately not renewed, it became difficult to maintain the servers and deal with user problems. Ultimately the principal designers and creators of Tranche left the project and were not appropriately replaced for financial reasons. Although data does still flow in and out of Tranche, the reliability of these activities and consequently the integrity of the data is not at earlier levels (and below the threshold that could be tolerated by *MCP*, leading to its decision not to make raw data storage mandatory until the situation is sufficiently rectified). While an infusion of money would certainly help (and there has been recently a small amount generated by the ProteomeXchange network, supported by a grant from the European Union), it is the consensus of a number of interested parties, which has been expressed at several international workshops and meetings, that either permanent support for Tranche needs to be identified or a new entity needs to be created with a reliable basis of support that would ensure the long term viability of the enterprise. The latter, which could be described as an International Repository for Proteomic Data (IRPD), would require a central facility and mirror sites placed in appropriate locales internationally and be staffed with network administration / IT staff to oversee its operation.

The stakeholders in such an IRPD would be of several varieties. First and foremost, the publishers of the main proteomic journals would be expected to be prime users. The American Society for Biochemistry and Molecular Biology, who publishes MCP, would be a strong supporter of such an activity but it can be expected that other publishers would be as well. The Nature Group is on record as actively supporting raw data

deposition. Based on the activities with other 'omic sciences, various private and public funding agencies are likely to be so as well and instrument and software vendors have a vested interest in this process (and have actively participated in workshops and discussion panels addressing this issue). Various government laboratories and agencies have also expressed support in the past. Finally, there are the end users – the scientists who create this data and then ultimately use it for different purposes. There seems to be no lack of support for the concept among any of these groups – only in the process of administration. Such a repository would presumably also become part of the ProteomeXchange consortium, which has international membership, who would be able to provide additional advice and potentially limited financial support.

REFERENCES

1). R. A. Bradshaw, A. L. Burlingame, S. Carr and R. Aebersold (2005) "Protein Identification: The Good, the Bad, and the Ugly" *Mol Cell Proteomics 4: 1221-1222.*

2). R. A. Bradshaw, A. L. Burlingame, S. Carr and R. Aebersold (2006) "Reporting Protein Identification Data: The next Generation of Guidelines" *Mol Cell Proteomics 5: 787-788. doi:10.1074/mcp.E600005-MCP200*

3). R. A. Bradshaw and A. L. Burlingame (2010) "Technological Innovation Revisited" *Mol Cell Proteomics 9: 2335-2336. doi:10.1074/mcp.E110.005447*

Advancing Transfusion and Cellular Therapies Worldwide

Sent via e-mail to digitaldata@ostp.gov

January 12, 2012

John P. Holdren
Director, OSTP
725 17th Street, Room 5228
Washington, DC 20502

Re: Document #2011-28621

Dear Mr. Holdren,

AABB (formerly the American Association of Blood Banks) is pleased to respond to OSTP's November 3, 2011 *Federal Register* notice requesting comments on "Public Access to Digital Data Resulting from Federally Funded Research." AABB appreciates the opportunity to respond to the issues raised in the notice.

AABB is an international, not-for-profit association representing individuals and institutions involved in the field of transfusion medicine and cellular therapies. The association is committed to improving health by developing and delivering standards, accreditation, and educational programs that focus on optimizing patient and donor care and safety. AABB membership consists of nearly 2,000 institutions and 8,000 individuals, including physicians, nurses, scientists, researchers, administrators, medical technologists, and other health-care providers.

AABB owns *TRANSFUSION*, the foremost peer-reviewed publication in the world for new information regarding transfusion medicine. Written by and for members of AABB and other health-care workers, *TRANSFUSION* reports on the latest technical advances, discusses opposing viewpoints regarding controversial issues, and presents key conference proceedings. In addition to blood banking and transfusion medicine topics, *TRANSFUSION* presents submissions concerning tissue transplantation and hematopoietic, cellular, and gene therapies.

Like many other societies, AABB depends on non-dues revenue such as that generated by data collection, analysis, and dissemination activities to support important work that serves not only a specialized (in this case, medical) community, but also society in general.

AABB offers the following responses to the Request for Information appearing in 76 FR 68517.

1. ***What specific federal policies would encourage public access to, and the preservation of, broadly valuable data resulting from federally funded scientific research, to grow the US economy and improve the productivity of the American scientific enterprise?***

8101 Glenbrook Road
Bethesda, MD 20814-2749
301.907.6977 MAIN
301.907.6895 FAX
www.aabb.org

Through its publisher, John Wiley and Sons, *TRANSFUSION* encourages growth in existing and new markets. The journal has a policy for open access to data from federally funded research. That policy has been in place for some time without controversy or challenge, and appears to meet the needs of the journal's constituency. AABB and Wiley have made investments in digital and online technology, and have actively participated in library consortia worldwide to accelerate and broaden access to research data submitted to the journal. There is more access to more content by more users now than ever before.

However, AABB is unaware of any studies showing that free access to the research data will increase research productivity or economic growth. *Access* to the data does not automatically translate to the ability to *use* that same data. The modern research enterprise is complex and requires huge investments. Limited resources are the constraint, not access to the data.

AABB does not accept the premise that because government funds scientific research, the government is entitled to full access to and control of data reported in this research. Managing, analyzing, disseminating, and archiving data are expensive. The government pays only for the conducting of research; it is unfair for it to lay claim to the fruit of labor by others.

Many research funders require research progress reports on all grants. Expanding this information by requiring the addition of a one-paragraph lay summary, and making both freely available, has more potential to enhance public understanding than does providing free access to data. AABB's strong preference would be that the federal government does ***not*** mandate deposit of research data in a freely available archive, regardless of format, process, or timing. Rather, the federal government should strive to provide public access to the information that it already controls and has a right to distribute — for example, research summary reports.

Typically, these reports are produced as part of each federally funded project, and they are provided to the government as a contract deliverable. Thus, there is a report for virtually every project. Each project itself undergoes peer review before being selected for funding, and the research results being reported on are solely those that the government funded. In short, these reports are the federally funded research results. Thus, if the policy is to provide public access to federally funded research data, then these reports are the natural vehicle for doing so. The government already has them, so all it has to do is make them publicly available. Several federal science agencies already do this; no new system is required.

2. ***What specific steps can be taken to protect the intellectual property interests of publishers, scientists, federal agencies, and other stakeholders with respect to any existing or proposed policies for encouraging public access to, and preservation of, digital data resulting from federally funded scientific research?***

Input from stakeholders is key. Partnership with publishers will deliver more to taxpayers at lower cost, with minimum economic burden. Publishers maintain an interest in long-term stewardship and improved public access to the data generated by federally funded research. What should ***not*** be considered is to take

data that have been collected and analyzed by publishers or learned societies (directly or via a mandate placed on grantees) and make the data freely available.

3. ***How could federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?***
4. ***How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?***

Agencies should collaborate closely with publishers, scholarly associations, universities, and other research entities to achieve the full potential of publicly accessible, interoperable databases. Increasingly, investigators are being asked to share, or provide plans regarding how they will share with other researchers, the primary data and other supporting materials created or gathered in the course of their work. As publishers and societies respond to increasing author demand to making research data available we are focusing on: 1) establishing best practice guidelines to make data available and retrievable in a consistent way, 2) collaboration with publicly endorsed community archives to make data and publications interlinkable, and 3) presenting data in more sophisticated formats to increase reuse.

5. ***How can stakeholders (eg, research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?***

Scientific, technical, and medical publishers (including learned societies) make significant amounts of data available as supplementary material to published articles and are already participating in initiatives designed to facilitate the sharing of data. AABB would be willing to work with funders and database/ repository operators to develop recommended practices for assigning Digital Object Identifiers (DOIs) to data sets and supplementary material, so that datasets could be linked to primary research articles.

6. ***How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?***

Federal agencies are not always aware of existing technologies and solutions in the marketplace, resulting in unnecessary spending and a misallocation of taxpayer dollars—particularly when the Government duplicates and competes with products and services provided by the private sector. For example, the National Institutes of Health (NIH) did not proactively seek collaboration with journal publishers as it developed its procedures and policies for the deposit of NIH-funded researchers' manuscripts into its central repository. Consequently, NIH created an unnecessary separate archive and tagging system at considerable expense and with minimal interoperability with existing data repositories.

It is questionable whether the government could become a credible provider of data management services. Given government budget constraints, the government would be unlikely to use taxpayer dollars to

duplicate an existing, well-functioning service. PubMed Central, the repository for mandated NIH grantees, is not a simple archive, but a sophisticated platform requiring millions of dollars of investment. Criteria for funding should address and prevent duplication of, or competition with, products and services offered by the private sector.

7. ***What approaches could agencies take to measure, verify, and improve compliance with federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?***

Again, AABB does not accept the premise of federal data stewardship, especially with the added burden of compliance and verification. Government agencies may fund scientific research, but that does not entitle the government to control of the data reported in the research. Managing, analyzing, disseminating, and archiving data are expensive, value-added activities of publishers and learned societies. The government pays only for the conducting of research; it is unfair for it to lay claim to the fruit of labor by others.

8. ***What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?***

As noted earlier, AABB is unaware of any studies showing that free access to the research data will increase research productivity or economic growth. *Access* to the data does not automatically translate to the ability to *use* that same data. The modern research enterprise is complex and requires huge investments. Limited resources are the constraint, not access to the data.

AABB believes that data that have been collected, analyzed, or otherwise managed by publishers should not be made freely accessible without the publisher's permission. AABB believes that publishers — and learned societies — themselves should determine the business models under which they operate. Peer-reviewed papers containing data are the direct result of investment and value added by societies and/or publishers, not the federal government. Thus, material should ***not*** be made freely available to the public unless the publisher or learned society authorizes the government to do so.

Respectfully,

Laurel V. Munk, MLS

AABB Publications Director
8101 Glenbrook Road
Bethesda, MD 20814
301-215-6595
laurie@aabb.org

Open Public Response to Request for Information: Public Access to Digital Data Resulting from FFSR

1/12/2012

This is a public and open document intended to draft a collective response to the request of information posted by the Science and Technology Policy Office (OSTP), on whether digital data resulting from federally funded research should be required to be made publicly available.

Dear Office of Science and Technology Policy,

Kitware applauds the initiative of the OSTP on seeking public feedback on these matters of high relevance to the scientific community and to the American public. However, please note that this is not an official Kitware response.

In order to contribute to this process, we reached out to our many collaborators and invited them to join us in writing a collective and thoughtful response to the insightful questions of the RFI. The result is the document attached to this submission letter. The names of the contributors and those in favor of this response are found at the end of the document.

Please find below our response to the RFI on "Public Access to Peer-Reviewed Publications from FFSR". NOTE: In the responses below we use the following acronyms:

FFSR: Federally Funded Scientific Research

Preservation, Discoverability, and Access

Question 1: What specific federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Response:
In summary our response advocates:

- Immediate release of acquired data
- Disclosure of a broad estimation of acquisition cost
- Proper open licensing
- Adoption of open standards for data files
- Adoption of extensible standards for metadata

Immediate Release
Federal agencies funding scientific research must establish policies by which the data acquired in federally funded scientific research (FFSR) must be made immediately and fully available in public data repositories. These policies should include provisions for protecting private information in the case of human subjects participating in medical research.

The policies should follow the model of:

- Bermuda Principles
- Pantom Principles (http://pantonprinciples.org/)
- http://sciencecommons.org/projects/publishing/open-access-data-protocol/

In particular on:

- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute, and reorganize the data.

Disclosure of Acquisition Cost
When reviewing proposals for funding opportunities, federal agencies should require that the sections requesting public funds for data acquisition activities provide a clear estimation of the cost of acquiring the data. If funded, researchers should be required to make data available in public repositories within 24 hours after acquisition, and in the metadata used to describe a dataset; researchers should also be required to include a disclosure of the cost of acquisition.

The goal will be to develop a sense of the economic cost of not releasing data. For example, not releasing a dataset that cost $1M to be acquired is a loss for the federal government from the $1M funds provided by taxpayers. This is the direct value lost from the overall economy; the

actual value lost is much larger since it includes the missed opportunities that could have resulted from the exploitation of the data.

The European Commission, for example, recently adopted a policy of open data dissemination http://www.kitware.com/blog/home/post/212. The principle, rooted in the arguments that Yochai Benkler makes in his book "The Wealth of Networks" is that data is more valuable when shared; in economic terms, data is an "anti-rival good". It is a good that becomes more valuable when more people have access to it and use it.

Proper Open Licensing

Current copyright legislation has been strongly focused on protecting the creators of artistic works, and in the process have created an inhospitable environment for the daily sharing of scientific information. The litigious tendencies of many institutions regarding copyrighted materials also results in over-cautious behaviors from potential data users and documents resulting from scientific research activities.

To dispel this environment of uncertainty, it is fundamental to clarify the rights of the public to make use of data acquired as a result of FFSR. The most effective way of achieving this goal is by affixing a clear statement of licensing, indicating what the recipients of the data are legally allowed to do with the data, to every released dataset. Licensing issues are expanded on in both the Science Commons Protocol for Implementing Open Access Data and the Panton Principles for Open Data in Science.

Some of the best examples of proper licenses are:

- The Creative Commons Zero Waiver: http://creativecommons.org/publicdomain/zero/1.0/
- The Open Data Commons licenses: http://opendatacommons.org/licenses/

Federal agencies should identify a set of licenses that ensure the rights of the general public to deal with the data, in particular to copy, distribute, and create derivative works, and in this way ensure that the data get to reach their maximum economic potential to foster the growth of the U.S. economy. It should then require federally funded researchers to make their data publicly-available under those selected licenses. The pool of licenses must be small, two at the most, to prevent confusion and to maximize the ease by which data can be integrated into subsequent research activities.

Adoption of Open Standards

Federal agencies must ensure that data are released in a usable form. The first step in that direction is to require the adoption of open standards for file formats, and forbid the use of proprietary formats that could prevent the general public from having access to the data.

Standard file formats used for digital storage of scientific data are abundant and vary greatly from one domain to the next. Therefore, the scientific community will have to be engaged with

the federal agencies in identifying the proper open standard to be used on each discipline, and to create new standards in the cases where no suitable standard file format exists yet.

For standards to reach their full potential, it is fundamental to have an open source reference implementation of the standard, and to encourage the development of an ecosystem in which commercial applications implement the standard as well. In this way, it becomes possible to maximize the use of the data acquired as a result of FFSR. The standards themselves must be unencumbered by patents and copyrights.

Open Standards for Metadata
In order to make use of FFSR data, the public must first be able to find it. This is typically done by implementing search engines that rely on publicly available metadata that is affixed to the actual FFSR datasets. The effectiveness of the search engines can be improved by the adoption of open standards that define the form and content of these metadata entries.

Just as with the data formats themselves, open metadata standards require an open source reference implementation, combined with an ecosystem where commercial applications implement the same open standard, to be effective. This wide adoption leads to interoperability, ease of communications, and data exchange.

These standards may have to be defined by different groups in different disciplines. For example, the genomics community will have different needs and interests than the astronomy community, the nano-sciences community, etc.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Response:
In addition to the stakeholders listed in this question, it is critical to note that the general public (the American taxpayer) is the primary stakeholder to be considered here. Given that in the context of FFSR, the public's tax dollars are paying for the scientific research being undertaken, and thus the public's interest is the first one that should be considered when making trade-offs between available options.

In order to have a productive discussion on intellectual property, it is important to first deconstruct the term "intellectual property" and clarify its meaning in the context of current U.S. laws. We do this in **Appendix A** and conclude that **copyright** is the only concept of intellectual property that is relevant for the purpose of this RFI.

Under U.S. copyright laws, the only aspect of scientific data that is subject to copyright protection is the creation of organized collections of data. Beyond that unique exceptional case, scientific data are not copyrightable, given that scientific data are factual and must never contain material resulting from the creative labor of artistic work. No scientific endeavor should include

data that are the result of the “creative work” of the researcher. Such a practice would be unethical in the context of scientific research. Scientific data must be the result of systematic measurement of real world parameters, or the outcome of computational models that operate on such real world measurements as inputs. In either case, such data do not fit the nature of “creative work” for which U.S. copyright laws provide protection.

Researchers may have applied creative works in the process of designing the experiments and methodologies that lead to the data acquisition. However, the actual data acquired must be factual, and therefore free of creative content, if it is to be considered worthy of the scientific process.

Regarding the copyright for organized collections of data, it is required by U.S. copyright laws that the data organization be non-trivial. For example, the simple ordering of temperature data acquired through time is not worthy of copyright protection. A novel and non-obvious approach to organizing data in such a way that it can be exploited for analysis, or that it reveals patterns and trends never seen before, is more aligned with the kind of creative work that copyright is intended to protect.

That being said, U.S. copyright laws are rooted in the economic bargain by which the government grant creators provide the exclusive right of exploitation of their creations for a limited time, as a way to provide an incentive for the production of such creative works.

In the context of FFSR, such an economic copyright incentive is not needed at all, because the federal government has already provided the funding for researchers to engage in the gathering and organization of the data in the first place. Therefore, the economic incentive has already been provided in the very concrete form of public funds awarded to federally funded researchers. Hence, the economic problem of provisioning “public goods” has already been solved proactively by paying up front for the scientific research using the monetary contributions of American taxpayers. Therefore, attention should turn to making sure that the American taxpayers get unfettered access to the data resulting from FFSR, which they have already paid for.

Scientists who gather data in FFSR do so as part of their job duties, and therefore under U.S. copyright laws they are performing “work for hire”. This means that their employers are the copyright holders of any creative aspect of that data gathering (as pointed above, that only includes the organization of data collections). Given that the scientists’ employers received funds from the federal government, it should be expected that they will be subject to the same demands of the Federal Acquisition Regulations (FAR) as other contractors of the federal government. In particular, with respect to the licensing of data acquired as part of federal contracts.

In the past, it has been a common practice for publishers to demand from researchers the transfer of copyrights related to the materials encompassed in a published scientific article, as a requirement for the publication of such article. No monetary compensation is given by publishers to researchers in exchange for that transfer. The policies of federal agencies should establish that the copyright of FFSR data collections should no longer be transferred to publishers, given that

publishers do not provide researchers, their employing institutions, or the federal government with any monetary compensation for such transfer of value.

To maximize the value of data to the public, federal agencies should require researchers to make FFSR data publicly available immediately upon acquisition by using open licenses that clearly state the rights of the general public when dealing with the data.

(3) How could federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Response:
Working groups should be established for different disciplines, involving representatives of leading research institutions for each discipline.

Working groups should define differences on how the data are represented, indexed, stored and exchanged, but should **not** have the latitude to restrict the free dissemination of information in any way. All the policies should consistently have a common requirement for immediate and full release of data, unconstrained by any embargo periods or licensing restrictions. Credit for the acquisition of data could be ensured by data publications (eg http://datacite.org) that can be cited by further works.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Response:
The working groups in the different disciplines (from Question 3) should establish guidelines on practices for dissemination and storage of different types of data. For example, in genomics, it may be reasonable to store the secondary sequence information but not the primary sequence (given their great difference in data size). Analogously, the guidelines may require primary sequences to be stored only for 2 years, while the secondary sequences should be stored for 10 years.

In astronomy, it may be required that certain types of images be stored for different periods of time. Some images may be required to be stored with different compression ratios, and therefore correlate their storage cost with the potential expected benefit for future studies. In this cost-benefit evaluation, the original cost of acquiring the data should be taken into account. For example, a project that invested $50M in acquiring data should not attempt to make savings of a few hundred dollars in storage.

Economists must be involved in the working groups with the mission of providing guidelines for storage and dissemination, as that this is a problem in which the trade-off for the benefit of society-at-large must be continually evaluated.

The recommendations of these working groups should be reviewed and updated regularly in order to keep up with the constant advances in storage technology and the rapid decrease in the cost of storage. The federal government should stimulate the development of storage technology, either by creating large storage decentralized facilities, creating consortia to manage data storage services, involving the public in facilitating distributed (and redundant) storage systems based on peer-to-peer network technology that has already proven to handle large amounts of data.

All these guidelines should be prepared following open and transparent procedures in order to prevent proprietary standards and vendor lock-in situations that would prevent the policies from maximizing the utility of FFSR to the general public.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Response:
They can join the working groups established in their respective disciplines of interest that will define practices for data management, including consortia combining universities, commercial companies, and government agencies.

As standards and agreements are developed, working groups can help implement and test such plans in pilot projects. It will be of great help if federal agencies provide seed funding for these pilot projects.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Response:
A. Specific funding streams should be created for researchers and institutions that dedicate themselves to hosting and distributing data. Today, there are very few (if any) funding opportunities for institutions that provide data storage services to their scientific communities, despite the fact that such services are of immense value for fostering the progress of their fields.

B. The rewards and merit systems of federal funding agencies must be adjusted to give proper incentives to researchers (and their institutions) who dedicate themselves to facilitate the storage and free dissemination of scientific data. These activities must be valued when researchers and their institutions pursue further funding. Today, only peer-reviewed publications are counted as part of the merit system of researchers when they apply for further funding opportunities. Therefore, researchers have no incentive to engage in public data sharing, and instead have self-interest in retaining data with the hope that it can help them produce more peer-reviewed publications that will contribute to fostering their careers.

C. Standard funding streams (such as R01 grants) must include provisions to fund the initial storage and dissemination of data acquired during a research project. This should be enough to cover the period of performance and two years after the end of the project. After that period, data should be moved to dedicated storage services. This practice will replace the current

approach of having data storage and processing as an "afterthought," which leads to inadequate data management, therefore data loses and underutilization of data. See a blog on"Software Forethought" by Kitware CEO Will Schroeder: http://www.kitware.com/blog/home/post/196.

D. Federal agencies should track researchers' compliance with releasing data resulting from previous funding when considering new proposals from those same researchers.

E. The Data Sharing plans in grant proposals should be evaluated based on specific provisions for storage and dissemination of the data to be acquired. Review panels should include reviewers with expertise on data storage and web-based distribution services.

(7) What approaches could agencies take to measure, verify, and improve compliance with federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Response:
A. Define standard annotations that include information about the funding stream (e.g. grant number, researcher identification, dates of funding) that supported the acquisition of the FFSR datasets.

B. Require awardees to tag their data releases with the type of annotations defined in (A) when they post the FFSR datasets to public repositories.

C. Fund the creation of a distributed indexing system that allows many institutions to consistently index the annotations (A), and helps the public search those indexes to efficiently locate and gain access to the data. These systems must be decentralized and be open for other organizations and individuals to introduce innovative searching and indexing mechanisms.

D. Provide a public Dashboard where the record of data releases for every funded researcher will be displayed publicly. The information should be provided in such a way that it can easily be harvested and data-mined by any other institution for the purpose of generating statistics and comparative studies. Public, open and transparent reporting of compliance with data release policies is the most effective way to ensure that researchers adopt data dissemination practices as a regular and standard activity.

E. Award institutions and researchers who excel at data dissemination. For example, a federal agency could provide honorary awards to the researchers each year who excel at sharing data.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Response:
A. Clear Licensing
Identify licensing practices that provide clarity on the types of activities that users can perform with the data. It is particularly important for companies and start-ups to invest in the utilization

of the data for fostering businesses and creating jobs. Business must be able to trust that they will not end up in litigation for having used data that was generated as a result of FFSR.

Work by multiple scholars have covered this topic, see for example: Stodden, Victoria, "Enabling Reproducible Research: Open Licensing for Scientific Innovation" (March 3, 2009). International Journal of Communications Law and Policy, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1362040 .

A.1 Creative Commons Public Domain

One of the best licenses to be considered for scientific data is the **CC0** license: http://creativecommons.org/publicdomain/zero/1.0/

This license has been defined as the closest we can get to put resources in the public domain. The **CC0** license lowers the bar of requirements and controls what the potential rights holders can impose on the recipients (the downloaders and users) of the data.

Our recommendation is to adopt the **CC0** license as the default standard of data sharing in order to ensure that American taxpayers get the maximum return on investment on the resources that they have put in the scientific research enterprise. The **CC0** license removes the majority of obstacles that can be imposed to the free dissemination of scientific information.

For a licensing discussion, see this podcast: http://inscight.org/2012/01/08/episode-22-public-access-to-federally-funded-research/

A.2 Open Data Commons

Another good set of data licenses is the one defined by the Open Data Commons: http://opendatacommons.org/licenses/ . Among this set, the recommended license is the Public Domain Dedication License: http://opendatacommons.org/licenses/pddl/, which simply states that the data is in the public domain. Placing data in the public domain makes sense because scientific data must not contain any glimpse of creative work. Instead, scientific data must be factual, and facts are not copyrightable.

A.3 Compliance

Once a set of acceptable licenses are defined for data, funding agencies should require that researchers and institutions use such licenses when delivering data for dissemination, or for storage in external repositories. All such licenses must allow for redistribution, reorganization, and repackaging of the data.

It is reasonable to demand attribution of data sources. Attributions will cascade when data has been passed through multiple stages of processing from one institution to another. In order to prevent the attribution process from becoming a heavy burden, federal agencies should adopt the policy that attribution must be done by citing the URI (Uniform Resource Identifier) of the

datasets used. This form of attribution has the properties of being machine readable, searchable, indexable, unique, and compact.

B. Pilot Educational Projects
Create streams of funding for pilot projects that will demonstrate how to systematically access public data repositories and generate concise representations of the data. The goal of the pilot projects will not be to innovate by themselves, but to educate the larger public on how to harvest data. Empowered with skill, citizens and institutions will have a lower barrier of entry into the practice of taking advantage of public datasets.

In parallel, funding agencies should spur educational programs for researchers to provide training on the management of data and data collections. Libraries, archives and repositories will be the organizations with the proper background to compose such training programs.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Response:
A. Tagging Data with Attribution MetaData
A commonly defined set of metadata annotations will facilitate tagging data with identifiers that point to the funding source, researcher name, research lab, institution, and other key attribution information.

When considering articles for publication, publication venues should require researchers to disclose if they used data from third parties, and if so, to provide the proper attribution using the standard annotation identifiers corresponding to that third-party data source.

As with the rest of the scientific publishing practices, this will be a combination of an honor system, with a light-weight verification system for publishers and funding agencies. The whole becomes effective if it is done in an open and transparent manner, given that any other third party, and in particular any other researcher who suspects that the data she/he disseminated has been used without proper attribution, could raise concerns and trigger corrective measures.

B. Promoting the Creation of Self-Regulating Governance Bodies
The problem of proper attribution to the providers of FFSR data is equivalent to the socioeconomic problem of governing the use of common pools of resources (CPRs). As described by Elinor Ostrom, 2009 Nobel Laureate in Economics, such governance models are successful when they have the following characteristics, among others:

- Collective-choice arrangements that allow most resource appropriators to participate in the decision-making process. For the purpose of discussing attribution in this RFI, a "resource appropriator" will be any person or institution who takes FFSR produced data and uses it to further their own mission and goals.
- Effective monitoring by individuals who are part of or accountable to the appropriators. In the case of this RFI, both monitors and appropriators are the researchers who produced and used data.

- A scale of **graduated sanctions** for resource appropriators who violate community rules. This system makes possible to actually apply sanctions when needed, given that the first scale of them will mostly be used as a "call to order," so that researchers who inadvertently broke rules have a chance to fix their omissions without dramatic consequences. At the same time, those who dismiss the "calls to order" can be progressively exposed to increasingly serious sanctions.
- Mechanisms of conflict resolution that are cheap and easy to access.
- Self-determination of the community recognized by higher-level authorities.

The Funding agencies should foster, but not control, the creation of researchers' managed **Data Attribution Tribunals**, perhaps with a less dramatic name, in the image of the "Water Tribunals" that have been used for centuries to successfully manage common water resources. This is one of the practical examples of Governance of Common Pools of Resources from which Ostrom deduced the governance principles listed above. Note that these tribunals are not government organizations; on the contrary, they are community groups composed by the same researchers who have a stake in the process of generation, dissemination, and attribution of scientific data. They would operate on the same honor system and volunteer bases that the current peer-review process operates, but whilefully public and transparent.

Given the sensibilities of researchers, a name less dramatic than "Tribunal" will certainly be more conducive to engage them in the process. For example, **"Open Data Attribution Arbitration Group"** could be a better name.

Reference: Elinor Ostrom, "Governing the commons: the evolution of institutions for collective action", Cambridge University Press, 1990

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

Response:
Digital data standards must be open, patent-free, and must require an open source reference implementation. Research communities will each have different standards to suit particular needs, which is perfectly acceptable as long as they are all open.

In some domains, there are organizations or working groups formed from community members to establish data formats, standard descriptions, or common interfaces with open implementation. For instance, the International Neuroinformatics Coordinating Facility (INCF, www.incf.org) has international working groups on standards for data sharing in neuroimaging and electrophysiology. An efficient use of funds would be to promote the established standards and join existing working groups on metadata standards.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

Response:
Effective standard definitions are the result of:

A. Involving the users of the standard in the definition process. This may require funding to initiate representative working groups for establishing the standards, or continue the work of existing groups.

B. Ensuring the full openness of the standard by requiring patent disclosures and royalty-free patent licensing from any institution participating in the definition of the standard.

C. Developing free and open source reference implementations of the standard at the same time that the standard is being defined. This ensures the practical applicability of the standard being defined, and also greatly promotes wide adoption of the standard.

D. Promoting an ecosystem in which commercial applications are encouraged to provide implementations of the standard without having incentives to create proprietary variations of it.

(12) How could federal agencies promote effective coordination on digital data standards with other nations and international communities?

Response:
A. Ensuring that internationalization (of language and "locale") is made an integral part of the standards.

B. Starting with simple standards that can progressively be improved, instead of spending a lot of time on top-down design, committees, and long-term procedural approaches to the definition of the standard. In other words, following the Agile methodologies that have proved to be successful in open source communities.

C. Working with existing international organizations that have already defined standards in different disciplines. See for example, INCF.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Response:
Adopt the standard use of

- Unique Resource Identifiers (URI)
- Digital Object Identifiers (DOI)

These two mechanisms have been used for several years to refer to digital resources in the literature.

An URI points directly to the location of the data in an unambiguous way.

The interest for DOIs is that, in many cases, researchers and institutions want their data to be addressed through another level of indirection to enable the moving of data from one hosting service to another. Services such as DOI (http://www.doi.org/) enable that level of indirection in a standard way.

Signatures

Name	Title	Institution
Luis Ibanez	Technical Leader	Kitware Inc.
Wesley Turner	Technical Leader	Kitware Inc.
Berk Geveci	Technical Leader	Kitware Inc.
Amitha Perera	Technical Leader	Kitware Inc.
Marcus Hanwell	R&D Engineer	Kitware Inc.
Matthew McCormick	R&D Engineer	Kitware Inc.
David Stoup	R&D Engineer	Kitware Inc.
Andy Bauer	R&D Engineer	Kitware Inc.
Paul Tunison	R&D Engineer	Kitware Inc.
Zack Galbreath	R&D Engineer	Kitware Inc.
Jean-Christophe Fillion-Robin	R&D Engineer	Kitware Inc.
Katie Osterdahl	Communications Specialist	Kitware Inc.
Katie Sharkey	Communications Specialist	Kitware Inc.
Steve Jordan	Graphic Designer	Kitware Inc.
Arno Klein	Asst. Prof. Clinical Neurobiology	Columbia University
Jean-Baptiste Poline	Researcher	UC Berkeley and CEA France
Cameron Smith	Computational Scientist	Rensselaer Polytechnic Institute
Ziv Yaniv	Principle Investigator	Children's National Medical Center
Raphael Ritz	Scientific Officer	INCF

Appendix A - Intellectual Property in Scientific Data.

The term of "intellectual property" is commonly used as an aggregate of the concepts of

- Copyright
- Patents
- Trademarks
- Trade secrets

In order to understand how these concepts apply to the challenge of maximizing access to the results of scientific research funded by the federal government, it is important to analyze the concepts independently.

Copyright is a government-awarded monopoly given to the creators of works of art. This monopoly awards creators the exclusive right to (1) reproduce the work, (2) prepare derivative works of it, (3) distribute copies of it, (4) perform it publicly and (5) display it publicly. The duration of copyright is: (a) the lifetime of the authors plus 70 year, (b) 95 years for works created by a corporation, or (c) 120 years for unpublished works created by a corporation. The goal of copyright is to provide an incentive to the creators of works of art by giving them exclusive rights on the exploitation of the works for a limited time

In the context of dissemination of scientific data, the economic bargain of copyright bears very low or no relevance, given that researchers (those who acquire and process the data) do not get paid when publishing that data. Instead, they get funded proactively for performing the research that leads to gathering information that is later published. Therefore, a very concrete economic incentive has already been provided and delivered to the researcher in the form of funding that American taxpayers have invested in the acquisition of the data.

As opposed to a novelist, whose income if purely based on the sale of copies of her/his book, the salary of a researcher is based on their performing the duties of scientific research. Granted, publishing datasets is part of such duties, but it is not equivalent to the creative activity of writing works of art (such as novels, music, or poems). Given that, in the context of FFSR, researchers are already paid by the public beforehand and so there is no need for the economic incentive of copyrights to address any "market failure" on the production of public goods (in the economic sense of non-rival and non-excludable goods), as is the case for novels, poems, and music. On the contrary, once the FFSR data has been acquired, every day that passes without this data being publicly shared is a day in which economic waste takes place and the economy at large performs less efficiently. It is also a day in which American taxpayers do not get anything back from the funds that they provided to the research enterprise.

Additionally, the nature of scientific research requires that the content of scientific datasets must be measurements of facts and should be devoid of any "creative elaborations". In other words, the more "scientific" a dataset is, the less "creative artistic content" it should have in it; therefore, the less it deserves the protection that copyright is intended to provide to creative works of

authorship. The creativity of the researchers lies in the definition of the acquisition protocols, the experimental design, and in the specific apparatus or software used during the data acquisition, which sometimes are made especially for a specific dataset. The dataset itself, on the other hand, shall not include any creative content. A high quality scientific dataset must be a concise collection of facts, measurements, and computations on those measurements. Datasets with high levels of "creative content" are by definition not scientific datasets, and should not be produced as the outcome of federally funded research, or any other process that aspires to be called "scientific".

Patents are government-awarded monopolies on the commercial exploitation of an invention. This 20-year long monopoly is awarded to the inventors in exchange for the public disclosure of the invention, and its eventual delivery (at the expiration of the patent term) to the Public Domain. Given that public disclosure is a requirement of the patent economic bargain, for awarded patents there is no concern about including information in articles intended for publication. The full information about the invention should already be publicly available at the U.S. Patent Office at the time that the patent is awarded to the inventors. Data is not "patentable subject matter" given that it is not the result of a creative process and is not useful, non-obvious, or novel. Datasets collected in the course of scientific endeavors are expected to be a collection of factual data, and therefore, they are as far as they can get from the type of "creative" work that patents are intended to protect.

Trademarks are symbols, designs, and terms that identify a product, service or company in the public marketplace. They are intended to prevent confusion in the marketplace, to protect the reputation of the producers of goods and providers of services, and to reduce the transaction cost that consumers have to invest in finding good and services that satisfy their needs. In the context of dissemination of scientific data, trademarks play a minimal role given that datasets are not supposed to be mechanisms of marketing goods and services. It is actually contrary to ethical standards in the scientific research field to use dataset publication as a venue for promoting goods and services in the context of commerce.

Trade Secrets refer to information that organizations keep confidential. For a piece of information to be considered a trade secret, it must have some value and derive part of its value from the mere fact of being secret. Trade secrets are managed via contracts, typically established between organizations in the form of non-disclosure agreements and between organizations and their employees in the form of confidentiality clauses that are incorporated in employment contracts. It is the responsibility of the institution to take affirmative steps to prevent its confidential information from becoming public.

In the event that a piece of confidential information is leaked publicly, there is no legal protection that can prevent the further dissemination of such information, except from forbidding an intruder to make use of data that was acquired illegally (e.g. by trespassing into private property). Therefore, in the context of dissemination of scientific data, trade secrets are only relevant as a context in which institutions should establish policies and verification mechanisms that prevent confidential information from being included in any dataset that is submitted for public release. It is the responsibility of the institution and its employees to protect such confidential information. Once data is published, the institution has relinquished its claim for such data to be considered a trade secrets.

Provost

Northwestern University
633 Clark Street
Evanston, Illinois 60208-1101

Phone 847-491-5117
Fax 847-467-1630
nu-provost@northwestern.edu

NORTHWESTERN
UNIVERSITY

January 12, 2012

The Honorable John P. Holdren
Assistant to the President for Science and Technology and
Director, Office of Science and Technology
New Executive Office Building
725 – 17th Street, NW
Washington, DC 20502

Comments in response to Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting From Federally Funded Research
Federal Register Doc No 2011-28621
http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28621.htm

Dear Dr. Holdren:

We are grateful to the Office of Science and Technology Policy for the opportunity to submit comments about providing public access to research data. Rather than a point-by-point response to the questions in the Request for Information, we offer these general comments about the challenges and benefits of data sharing, and issues the federal government should consider in seeking to implement new requirements.

Northwestern University is a private research institution with 16,377 students and approximately 3,000 full time faculty. In academic year 2010-11, Northwestern researchers attracted total awards and grants of approximately $511.7 million. Northwestern's libraries hold more than 5 million volumes, 4.6 million microforms, and provide access to 110,341 current periodicals and serials. In addition, the library system boasts more than 700 databases and 6,000 electronic journals. 56% of the libraries' $14 million collections budget is devoted to these e-resources.

Northwestern is recognized both nationally and internationally for the quality of its educational programs at all levels. *U.S. News & World Report* consistently ranks the University's undergraduate programs among the best in the country.

Among graduate programs, the Kellogg School of Management regularly ranks among the top five business schools in the country for both its traditional curriculum and its executive master's program. *U.S. News & World Report* rankings placed Northwestern's School of Law 11th, and the Feinberg School of Medicine in the top 20.

Sharing and Public Accessibility

The absence of a policy requiring investigators to take specific action to share and preserve research data has resulted in management practices that are idiosyncratic and incomplete, which all but guarantees future data loss. Many Northwestern researchers already share their data with colleagues,

but exchanges may be informal or involve temporary sharing mechanisms (email attachments, temporary FTP servers, etc.) that do not take reuse or long-term preservation into consideration. While specific approaches will and should vary across disciplines, a policy that clearly articulates the definition and goals of providing public access to research data will support gradual development of standards and repository systems to enable responsible stewardship.

Current publication and preservation methods also often fail to identify clearly and consistently the data, data creators, and other provenance necessary to provide attribution in future work, or to address problems such as patent disputes. Researchers have an inadequate understanding of effective data management and curation practices. Many labs do not have the means to store permanently and provide internet access broadly to very large datasets, which may be petabytes in size. Therefore, a policy must be sensitive both to the significant technical challenges and the financial impact of sharing requirements. Centralized repositories for research data may prove to be a cost effective alternative that may alleviate investigators from the financial and technical burden of providing secure, reliable access to published results.

Perhaps most importantly, a policy should make clear what a public access requirement is designed to support. Reproducibility of research, independent verification of findings, and more rapid adoption of previous research to new investigations are all good reasons to mandate data sharing. Some data may not be sharable, either temporarily or permanently, for reasons of national security, privacy, or pending legal action, but these restricted data will still benefit from preservation services and application of standards to describe adequately and store safely research data. A clear statement of intent will help researchers determine whether all raw data, only significant findings, or only data directly linked to a publication are affected by a policy. It may be that an expansion of practice, if not policy, to encourage investigators to share negative results and other types of data not usually shared will also advance discovery.

A coordinated data sharing program must also clarify investigators' obligations to keep data safe, clearly define minimum acceptable practice for effective data management and curation, and tie compliance to ongoing funding. Careful consideration should be given to the design and development of tools that simplify metadata creation. Researchers who may be willing to share data will be very resistant to using awkward, poorly designed tools that disrupt active research or being forced to re-enter the same information repeatedly.

Copyright and Ownership

The legal status of research data must also be carefully considered. Facts do not satisfy the threshold for originality, and are therefore not eligible for protection under United States copyright law. 'Research data' is a broad term encompassing many different types of content, from the massive raw output of sensing instruments to text markup to painstakingly curated survey data and everything in between. While published research articles, survey instruments, software, and other research products will qualify for copyright protection, the data themself may not, so a different set of legal instruments may be needed to express the rights associated with data. Copyright transfer and licensing agreements, or the Creative Commons licenses that operate under a presumption of copyrightability, will not be sufficient to document the expectations of researchers. Open science and open data initiatives such as the Science Commons, specifically its database protocol project http://sciencecommons.org/resources/faq/database-protocol/, and the Panton Principles http://pantonprinciples.org/ provide a good discussion of data IP issues and examples of appropriate legal instruments. As with published research articles, a federal policy to promote public access to

data should not permit publishers to compel researchers to permanently restrict access to and use of their data as a condition of publication.

Funding and Implementation

Universities, their research administrators, libraries, and technology specialists are in a good position to advise investigators as they develop data management plans, and to help identify appropriate metadata standards, data description and normalization tools, and storage solutions. However, the costs of building these data management and preservation systems will be massive, and cannot be fully borne by individual universities.

The federal government has also struggled to maintain funding for large data storage projects, as demonstrated by the near closure of the Sequence Read Archive (SRA) in 2011. However, the SRA and other NCBI databanks, as well as those at the Food and Drug Administration and the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) are available to anyone worldwide and are well known throughout the research community.

Centralized or multi-organization approaches have the potential to be more cost-effective, create larger linked stakeholder communities both to advocate for continued funding and to develop standards, and increase success of data normalization efforts through provision of shared technology platforms. A hybrid approach of local infrastructure for active research phases, and centralized or multi-organization solutions for broad public sharing or long-term preservation is likely inevitable. These initiatives should emerge in parallel so that critical information about projects, software, algorithms, and other meta-information are consistently captured beginning early in the data lifecycle, and travel with the data to reduce barriers to sharing. If the government cannot provide centralized storage for research data, grant funding to researchers and their institutions must be increased to support expansion of local or disciplinary capacity, or to pay incremental costs associated with a single project.

Standards for versioning, selecting, describing and citing/attributing data must evolve in conjunction with the researcher communities who use them. Although far from comprehensive, here are a few comments and examples of current standards and development activities:

Citation, attribution and linking

If data are received directly from another researcher, new publications arising from these data must mention this in a methods section, and all data sets should be properly cited in methods and references sections (depending on norms for discipline and the specific journal's format). Failure to properly cite datasets should have the same consequences as other instances of plagiarism: retraction of manuscripts. The data must be cited consistently; see the DataCite project http://datacite.org/ for an example of a promising data set registry and identifier minting service. Implementing a data sharing and citation system is also a ripe opportunity for linked open data (LOD) and RDF to take the forefront. If each dataset has a unique identifier, it can be linked through RDF triple format (e.g. "Paper [paper identifier] has related dataset Y [dataset identifier]"), further enforcing consistency, but also significantly improving machine readability.

Standards for interoperability and reuse

The FDA is evaluating similar standards to MIAME (Minimum Information About Microarray Experiments) for ChIP-Seq and RNA-Seq data descriptions. The Gene Ontology project is an initiative with the aim of standardizing the representation of gene and gene product attributes across

species and databases. The project provides a controlled vocabulary of terms http://www.geneontology.org/GO.downloads.ontology.shtml for describing gene product characteristics and gene product annotation data http://www.geneontology.org/GO.downloads.annotations.shtml from GO Consortium members, as well as tools to access and process http://www.geneontology.org/GO.tools.shtml this data. This is a good example of a standardization initiative in an area where everyone was formerly using different terms for the same objects. This type of success requires cooperation among leaders in the field in question and a workflow that produces an accepted standard. Although not examples of standards, these papers, whose contributing authors include Rex Chisholm, Dean of Research for Northwestern's Feinberg School of Medicine, are examples of consortial projects dealing with identifying and re-using data:

1. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, et al. Towards BioDBcore: a community-defined information specification for biological databases. Database (Oxford). 2011;2011:baq027. PMID: 21205783. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3017395/

2. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, et al. Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res. 2011;39(Database issue):D7-10. PMID: 21097465. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013734/

3. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79re1. PMID: 21508311.

4. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4:13. PMID: 21269473. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3038887/

Defining standards for data exchange is difficult, but a bare bones framework of required minimum fields for describing a dataset will be useful. Likewise, using tools like JHOVE2 or DROID to identify, validate, and extract features from data sets will greatly enhance compliance with description requirements by reducing the number of fields for which data must be manually supplied.

Thank you for this opportunity to comment.

Sincerely,

Daniel Linzer
Provost

Thu 1/12/2012 4:27 PM

Response to OSTP RFI on Public Access to Digital Data Resulting From Federally Funded Scientific Research

Access to primary research data is important for the advancement of the scientific enterprise. It facilitates the validation of existing observations and provides the raw materials to build on those observations. These benefits, however, must be balanced against the burden to researchers of providing such access. The time and money required to provide access to research data are time and money lost to research.

The National Science Foundation recently enacted a blanket policy to require sharing of all data generated using their funds. But it is impractical and not essential to require researchers to share every piece of data they acquire. Sharing policies will have to be more specific about the types of data that will be useful to share and whether preliminary data sets are included. The NIH policy takes steps in this direction. In addition, funding agencies will have to provide the mechanisms for sharing large data sets. Individual scientists cannot be expected to provide these mechanisms. Sharing policies will also be ineffective unless they are enforced.

In biomedical research, there are numerous instances in which the members of the research community have determined that a particular type of data would be useful and necessary to share. These include gene sequences, protein structural data, and gene and protein expression profiles. In these cases, the community united to standardize the structure of the data and its associated metadata, and federal agencies created centralized repositories (or funded their creation) to facilitate deposition, promote discoverability, and ensure the longevity of the data. Funding agencies will have to maintain an ongoing dialog with the research community to decide what other types of data will benefit from standardized repositories, and they will have to fund the creation of those repositories.

In addition to standardized data and metadata structures, two other elements are essential for the success of any repository: customer service, to make it as easy as possible for researchers to deposit data in the correct format; and curation, to ensure that data are properly formatted and tagged, and to monitor any crowd-sourced tagging through Wiki applications. Proper tagging of data with accession numbers and/or digital object identifiers will help to ensure the longevity of links to those data.

Access to research data is important, but data without context are not useful to third parties. Why were the data obtained (to answer what question)? How were they obtained? What was the interpretation of the data by the person who obtained them? These questions closely mirror the introduction, methods, and results/discussion structure of primary research articles, and the most obvious way to provide context to data is through a scientific publication. Thus, the primary burden of enforcing deposition of data into repositories has fallen on journals.

Biomedical research journals require that certain types of data such as nucleotide sequences, protein structural information, or protein/gene expression profiles be deposited in repositories hosted and curated by (or at least funded by) funding agencies. This enforcement process, which evolved with the development of these repositories, functions for specific types of data underlying published research articles.

In rare cases, a publisher may develop its own repository to fill a gap in providing access to a type of data that is prevalent in a particular journal. One of the journals published by The Rockefeller University Press, *The Journal of Cell Biology (JCB)*, publishes a large amount of microscopy image data. In the absence of a standardized, international repository for this type of data, the journal developed the JCB DataViewer – a browser-based application for viewing original, multidimensional, microscopy image data.

The imaging community is coming to some consensus about the data and metadata structures necessary for sharing and archiving microscopy image data, but most of these data reside on desktop computers in proprietary file formats (PFFs) that cannot be shared. The JCB DataViewer uses an interpreter to convert those PFFs into a standardized format and display them over the internet. It can also host the complete data sets from high-content imaging screens. Deposition of image data by authors into the JCB DataViewer is encouraged but remains voluntary.

The JCB DataViewer is currently used for original image data supporting articles published only in the *JCB*. We hope that it will serve as a prototype for the development of a larger repository for images published in any journal. But it is not reasonable or sustainable for an individual publisher to undertake such an expansion. This must be done by national or international funding agencies.

Funding agencies have relied on journals for enforcement, and, indeed journals are in a strong position to place requirements on authors before publishing a paper. However, journal publishing is competitive, and journal editors may be reluctant to afflict potential authors with additional demands that they may consider burdensome for fear they will submit their papers to another journal with less stringent requirements. Most journals will establish an access policy to a particular type of data only once a standard for sharing that data type has been set and an expectation of compliance has been established in the research community. But even then, there will be variability in the stringency of enforcement.

For newer standards (for example, high-content image screens), there will be great variability in requirements by journals until an expectation of compliance has been established. Given these variabilities, the funding agencies should monitor published data for compliance with sharing policies, and they should not rely solely on the journals. Enforcement of sharing polices for data that have not resulted in a publication will fall completely on the funding agencies. They will have to decide what types of data to monitor (they can use editorial policies of biomedical research journals as examples) as part of the grant application/renewal process, and they will have to create a monitoring step in that process.

It will be vital to provide context to unpublished data by ensuring that sufficient metadata are associated with the data for a third party to understand their origins (and to recognize that they are unpublished, and thus the methodology has not been vetted through peer review). Funding agencies will also have to develop policies about the timing of data release to the public. For data underlying a published research article, it is easy to set such a policy – the date of publication. For unpublished data, sufficient time will have to be provided to license data that may have commercial value. Funding agencies will have to monitor licensing terms to ensure that reuse by non-profit institutions is allowed.

Blanket policies regarding sharing of primary research data sound impressive and progressive, but they are neither practical nor enforceable. Funding agencies need to be specific about the types of data that they expect to be made public. Relevant scientists must be engaged to develop these standards, and funding agencies must provide the mechanisms for applying them.

Mike Rossner, Ph.D
Executive Director
The Rockefeller University Press

These comments are the opinion of the author and do not necessarily reflect the position of The Rockefeller University.

AMERICAN COLLEGE OF RHEUMATOLOGY
EDUCATION • TREATMENT • RESEARCH

Specialists in Arthritis Care & Research

2200 Lake Boulevard NE • Atlanta, GA 30319
Phone: (404) 633-3777 • Fax: (404) 633-1870
www.rheumatology.org • info@rheumatology.org

January 12, 2012

National Science and Technology Council
Task Force on Public Access to Scholarly Publications
c/o Office of Science and Technology Policy
Attn: Open Government Recommendations
725 17th Street
Washington, DC 20502

Re: ***Public Access to Digital Data Resulting From Federally Funded Research; Request for Information [FR Docket No. 2011-32947]***

Dear Task Force Members:

The American College of Rheumatology, representing over 8500 rheumatologists and health professionals, welcomes the opportunity to comment on the Administration's approach to public access components of the scientific research enterprise.

Rheumatologists treat patients with arthritis and other rheumatic and musculoskeletal diseases. These conditions can be painful, debilitating, life threatening and costly. Biomedical research plays a pivotal role in advancing diagnostics, treatments, and prevention strategies for patients with chronic diseases. Advancements in arthritis-related research have helped to prevent disabilities, allowing patients to continue working or return to work and contribute to their communities and the economy.

The ACR believes that federal policy to promote access to digital scientific data produced in the federal and federally-funded realms should produce a climate of equitable access while protecting intellectual property rights. This climate provides a dynamic and healthy environment for basic and applied research that will enable the United States to continue as a leader in discovery and innovation.

We contend that federal partnership with publishers of scholarly articles on scientific research will deliver results to taxpayers with minimal economic burden to taxpayers. Publishers have a strong interest in long-term stewardship and improved public access to the results of federally funded research. These publishers have long been stewards of the literature and increasingly the data related to the associated research, which is usually delivered as supplementary material.

It is important to note that federal agencies are not always aware of existing technologies and solutions in the marketplace, which can result in unnecessary spending and allocation of taxpayer dollars, particularly when the government duplicates and competes with products and services provided by the private sector. For example, the NIH did not proactively seek collaboration with journal publishers as it developed its procedures and policies for the deposit of NIH-funded researchers' manuscripts into its central repository. Consequently, NIH created an unnecessary separate archive and tagging system at considerable expense and with minimal interoperability with existing data repositories.

The ACR believes that federal agencies should collaborate closely with publishers, scholarly associations, universities, and other research entities to achieve the full potential of publicly accessible, interoperable databases. Increasingly, investigators are being asked to share, or provide plans regarding how they will share with other researchers, the primary data and other supporting materials created or gathered in the course of their work. As publishers respond to increasing author demand to making research data available they are currently focusing on: (1) establishing cross-publisher best practices to make data available and retrievable in a persistent way; (2) collaboration with publicly endorsed community archives to make data and publications interlinkable; (3) presenting data in more sophisticated formats to increase reuse.

STM publishers, including learned societies, make significant amounts of this material available as supplementary material to published articles and are already participating in a number of initiatives designed to facilitate the sharing of data. Publishers should be willing to work with funders, as well as database and repository operators, to develop recommended practices for assigning Digital Object Identifiers to data sets and supplementary material, so that datasets can be linked to primary research articles.

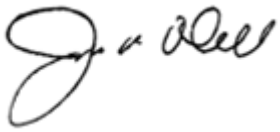
We contend that mandating a single approach to public access could actually stifle innovation in today's rapidly changing environment, by decreasing the amount that the private sector is able to invest and by reducing their incentive to try new approaches to managing and preserving data. In laying appropriate policy foundations, agencies should consider all components of a comprehensive agency data policy, such as preservation and access guidelines; assignment of responsibilities; information about specialized data policies; provisions for cooperation, coordination, and partnerships; and means for updates and revisions.

The federal government should devote its efforts to providing public access to the information that it already controls and has a right to distribute, such as research summary reports. As the federal government considers complex issues surrounding data access and dissemination, agencies can take action now to deliver substantive public access to research outputs. Many research funders require research progress reports on all grants. The government, through its funding agencies, supports the research enterprise that generates outputs such as experimental data, technical reports, grant reports, and conference papers. Consequently, government has an important interest in ensuring that research data and technical reports are accessible to the public whose taxes funded their production.

The ACR believes that this approach would adhere to President Obama's pledge in his Transparency and Open Government memorandum to "take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use." Such policy measures would maintain copyright protection for private-sector investments, consistent with other priorities of the President for strengthening economic growth and job creation, cooperation with the private sector, innovation, and protection of intellectual property rights.

The ACR appreciates the task force's review of recommendations for ensuring long-term stewardship and broad public access to unclassified digital data that result from federally-funded scientific research. We stand ready to assist you further on these issues that affect the conduct of scientific research related to rheumatology and the broader rheumatology community, including the health and quality of life of our patients. If we can be of assistance to you in any way, please contact Adam Cooper, ACR director of government affairs, at acooper@rheumatology.org or (404) 633-3777.

Sincerely,



James R. O'Dell, MD
President
American College of Rheumatology

Thu 1/12/2012 4:41 PM

Comment on RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Mary Ochs / mao4@cornell.edu

President / United States Agricultural Information Network

Ithaca, New York

I am writing on behalf of the United States Agricultural Information Network (USAIN) to respectfully respond to the Request for Information for recommendations related to public access to digital data resulting from federally funded scientific research.

USAIN (usain.org) is an organization of over 150 agricultural information professionals that provides a forum for discussion of agricultural issues, takes a leadership role in the formation of a national information policy as related to agriculture, makes recommendations to the National Agricultural Library (NAL) on agricultural information matters, and promotes collaboration and communication among its members. USAIN has testified before Congress, played an advisory role in the National Agricultural Text Digitizing Project, written a national agricultural literature preservation plan, served on blue ribbon panels to review NAL services, and participated in the selection process for new NAL Directors. Our members are skilled librarians and information specialists with knowledge of the modern theories, principles, practices, techniques, and policy issues pertinent to the current practice of librarianship and information science. Many of our members work at Land Grant institutions with extensive federally-funded research programs and are experienced in acquiring, organizing, and preserving scientific and agricultural data. The USAIN Executive Council is privileged to provide the following input related to this important topic of public access to information.

Comment 1. What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Access to digital data is critical for scientists, students, innovators, entrepreneurs, and other interested citizens. This vital part of the process of the discovery of knowledge ultimately leads to the creation of new products, job opportunities, economic growth, research achievements, and the strengthening of

our society's knowledge base. Digital data created at the public's expense, but left unavailable or referenced only via subscription-based journal articles may not be equitably available to all who might benefit from its content**.** A system requiring a data management plan for data output of federally-funded research would necessitate a level of data planning. Ideally it would be beneficial to require researchers to also provide data to a subject-based or institutional repository so the data would be more readily available.

On one hand it is exciting to see the growing interest in digital data dissemination, but it is bittersweet given the recent budget woes and the mandated termination of the National Biological Information Infrastructure (NBII). The main Web site, www.nbii.gov, will be taken offline on January 15, 2012, along with all of its associated node sites. "January 15, 2012, will see the end of a long-term project to empower users of biological resources data and information. The National Biological Information Infrastructure, or NBII, was begun in 1994 within what was then the National Biological Service (NBS) of the Department of the Interior. Its purpose and mission were to ensure that scientists, resource managers, decision makers, and concerned citizens could go to a single place on the Web and find biological resources data and information from vetted sources—whether in government, academia, non-governmental organizations, or the private sector." See the announcement in USGA @ccess http://www.usgs.gov/core_science_systems/Access/p1111-1.html.

Comment 2. What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

Nothing in a public-access system for data should threaten the protection of intellectual property. In fact, greater access to research information and data will ensure greater visibility and recognition of an author's intellectual achievements. For sensitive or proprietary data or data that data owners don't want to make public for whatever reason, the system can provide access controls, such as password protection.

Comment 3. How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Just as there are different journal articles and citation styles in various disciplines, there are different expectations for sharing discoveries and reporting results. There may be differences in the disciplines

but stakeholders and research communities should be encouraged to establish standards that enable sharing and interoperability across disciplines. This will aid in the discoverability of data and data sets by libraries and research portals. Agencies should allow for and encourage subject-based repositories and build on the models of successful discipline-based data models such as ISCPR, GenBank, Dryad, and others.

Comment 4. How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

There is no magic guarantee that when a data set is created, there is an immediate recognition of the benefits of the long-term stewardship and dissemination. The same could be said for books and journals. At least with journals there are metrics for journal use and citation patterns. Similar metrics may emerge over time for data. Stakeholders, data creators and librarians/data managers can play a collaborative role in determining the standards for preservation. It is understood that large amounts of raw data may not be useful or understandable, so standards must be set to describe the data in its most usable form.

Current models of stewardship and dissemination that merit consideration include the approach provided by the NIH-mandated deposit of peer-reviewed research articles in PubMed Central. Their deposit requirement and sufficient program funding have made this repository successful. A comparable repository for USDA-funded research could be managed by the National Agricultural Library (NAL) as an expansion of the existing NAL Digital Collections (NALDC). The advantages of a centralized repository include better control of the deposit process, author compliance, and consistent metadata applications. Funding agencies managing a smaller grant portfolio may have a more difficult time supporting a separate repository, so centralization would benefit these agencies. Centralization also minimizes issues of interoperability, consistency and redundancy. Many universities maintain an institutional repository and could help facilitate required deposits within the institutional site or a centralized repository. Even with clearly articulated standards, achieving full interoperability across many repositories may be a challenging goal. Although the examples above relate to the management of articles, a similar system could be created for various types of data.

Comment 5. How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Stakeholders can best contribute to the implementation of the data management plans by being involved in the creation and use of the plans. Professional societies can play a role by providing leadership in defining data structures and types pertinent to the scholarship of their disciplines. Data creators may need encouragement to realize how their datasets may be of value to others. Funding agencies and institutions should promote and reward exemplary projects and best practices.

Comment 6. How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Allow costs of data storage, management, preservation, etc to be included in grant applications and funding. Funding support for national libraries and data repositories should also be provided.

Short-term funding would get projects such as iPlant off the ground, but limited funding for the beginning of projects would not be sufficient to sustain a project long-term. "iPlant is a community of researchers, educators, and students working to enrich all plant sciences through the development of cyberinfrastructure - the physical computing resources, collaborative environment, virtual machine resources, and interoperable analysis software and data services– that are essential components of modern biology" http://www.iplantcollaborative.org/about. It would be critical to have ongoing funding towards digital data management and accessibility included in the budget of national libraries (NAL, LOC) and agencies such as USDA.

Comment 7. What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The national libraries, which are mandated with providing stewardship of printed scholarship and given suitable resources, should play a similar role in managing data and electronic information.

Comment 8. What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

No comment.

Comment 9. What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

There should be policies and development of persistent identifiers that would allow the tracking of provenance, ensure data integrity, and contribute to successful citing of data and attribution to the authors/creators of the data. Creative Commons licensing may provide additional support for this effort.

Comment 10. What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

The national libraries have the requisite skills, experience and mandate to define and implement the standards that must be put in place to create an interoperable data repository system. The minimum metadata elements for describing bibliographic information are currently well-defined by the Dublin Core metadata standard. These elements can be readily derived from publisher data and incorporated as part of the deposit. Adherence to this standard, as well as the OAI-PMH standard for metadata harvesting, will facilitate the sharing of data from multiple repositories and lead to discovery by the public. Metadata standards are critical for describing publications and data within a repository, but institutions are also faced with the added challenge of increasing access to those resources. Resources must be highly discoverable and understood within a larger context of scientific data and research. For that to happen, several things must occur: 1) the advanced support of author disambiguation initiatives, such as ORCID, which "aims to solve the author/contributor name ambiguity problem in scholarly communications;" 2) a general mandate requiring federally funded authors to identify their funding source when submitting publications to a repository; and 3) the development and support of Semantic Web technologies that allow for the re-purposing, reuse, and analysis of publication and other data. By design, Semantic Web technologies are machine-readable; continuing to encourage the development and accessibility of these technologies would allow for flexible re-purposing of data, regardless of the model - centralized, decentralized, or mixed-model - chosen by Federal agencies.

Comment 11. What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The National Information Standards Organization (NISO) provides an example of a successfully improving procedures and streamlining standard development processes. This has resulted in a reduction of time spent releasing consensus documents and sped up the process for launching new initiatives. Library of Congress' Cataloging in Publication Program (CIP) offers a detailed explanation of the process on their website, including pre- and post-publication. National Institute of Standards and Technology (NIST) has also been instrumental in producing effective standards for the U.S. Their process involves working with cooperative programs and partnering with 1,600 manufacturing specialists and staff at locations around the country. Details regarding these organizations and their role in standards development can be found at their respective websites. (http://www.niso.org/; http://www.loc.gov/publish/cip/; http://www.nist.gov/)

Comment 12. How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Federal agencies could work closely with an organization such as the World Wide Web Consortium (W3C), an international community in which member organizations, staff and the public collaborate on the development of Web standards. Another possibility is The InterNational Committee for Information Technology Standards (INCITS), a forum for information technology developers, producers and users for the creation and maintenance of IT standards. A third option is International Organization for Standardization (ISO), the world's largest developer and publisher of International Standards. ISO is a network of the national standards institutes of 163 countries.

Comment 13. What policies, practices, and standards are needed to support linking between publications and associated data?

There should be policies and development of persistent identifiers that would allow the tracking of provenance, ensure data integrity, and contribute to successful citing of data and attribution to the authors/creators of the data.

Thu 1/12/2012 5:21 PM

response to RFI

Response to Request for Information "Public Access to Digital Data Resulting From Federally Funded Research", November 2011

January 12, 2012

Clifford Lynch
Executive Director
Coalition for Networked Information

Cliff@cni.org

I am pleased to have the opportunity to submit comments to this request for information on "Public Access to Digital Data Resulting From Federally Funded Research" on behalf of the Coalition for Networked Information (CNI). CNI is a membership organization consisting of some 200 organizations, primarily but far from exclusively universities, who share a common commitment to advancing the intelligent use of information technology and digital content in support of scholarship. You can find more information on CNI at www.cni.org.

I want to be clear that while these comments are certainly informed by discussions with CNI's member organizations, they should not be viewed as representing the position of any specific member of CNI.

There are a tremendous number of questions in the request for information, and I cannot comment on all of them here; I also know that you will be getting many other well informed and thoughtful responses, including some that I have already seen in draft. But I want to begin my comments with an overarching strategic point: we are relatively early in a great transition in scholarship. We have now explicitly recognized the large scale emergence of information technology enabled and data

intensive scholarly practice, and have begun to make systematic accomodation of this transition in our funding, policy, infrastructure and scholarly communication mechanisms. We need to carefully monitor what is actually happening as this transition moves forward. We need to examine the effects and outcomes of policy interventions on a continuing basis, and to be prepared to adjust these policies as experience dictates. We still have a great deal to learn about what data is of greatest lasting value, and what data is most likely to be reused in ways that produce new and important scholarship. This is a process, not a one time event. Our ability to manage this transition will be greatly improved by the availability of funding to help underwrite data collection, research and analysis, and also by policies that promote transparency and the public availability of data that can support research, analysis and evaluation. The level of investment needed here is miniscule relative to the scale of the scholarly and research enterprises.

What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Funding agency requirements for data management are an excellent start, and should be extended from NIH and NSF to all federal funders. Funders also need to continue to reinforce policies that require investigators to share data publically; this would include guidance not just to investigators, but also, importantly, to proposal reviewers. As an overall strategy, most data curation needs to be institutionalized: responsibility needs to transition away from the investigator and to institutions (either universities or disciplinary repositories) early in the data lifecycle, with these institutions being the primary long-term contacts for data access.

Both one-time (startup) and even more importantly sustained federal investment in disciplinary data interchange standards, and in data repositories (both institutional and disciplinary) and related infrastructure (such as disciplinary and cross-disciplinary discovery tools for datasets placed in repositories) are essential parts of the federal contribution.

A particularly problematic set of issues - legal, technical, policy, and ethical -- that may well be ripe for federal leadership are those inhibiting access and reuse of data that involves human subjects and personally identifiable information. Here we see disconnects and conflicts between long-standing but continually evolving practices and policies designed to protect human health, privacy and dignity, and the new opportunities for large scale computational reuse, recombination and analysis. These conflicts are becoming major barriers to progress, notably but far from exclusively in the health sciences. Note that these issues arise on both national and international levels.

What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

I am not a lawyer, but I think that the way this question is framed helps to illuminate one of the barriers to greater data sharing and reuse. Research data, as I understand it, basically isn't subject to copyright in the United States; while access controls and contracts can clearly be used to limit the ways in which it is shared and used in specific cases, and may be appropriate tools for supporting short term exclusive use embargos or similar arrangements, I don't think that there are traditional intellectual property issues here - data, and particularly data from federally funded research, should be regarded as part of a knowledge commons that belongs to all of us. Clarifying and codifying this, for all of the players, including researchers themselves, is extremely important, as is differentiating legal rights and obligations from moral or ethical ones (such as the moral obligation to acknowledge sources of data that are reused in subsequent research).

How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

It is clear that disciplinary differences must be recognized, honored and accommodated. And obviously different kinds of data require different policy frameworks (see my earlier comments on human subjects, for example). Yet we should also recognize that some disciplinary differences are largely disciplinary traditions, and some of these traditions are, in my view, ripe for re-assessment. And disciplinary practices and traditions that are inherently inconsistent with ideas about an open knowledge commons cannot be excused simply on the basis that they are disciplinary practices and traditions. Further: as we move into an era where interdisciplinary and multidisciplinary research is increasingly commonplace and increasingly necessary, greater consistency (or at least interoperability) across disciplinary practices will be more and more desirable - particularly in terms of describing and managing data resources and facilitating reuse of such resources.

How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

I think that this is a very poorly understood area and one that merits ongoing examination and research. There are classes of data that can be re-created, and here one can look at the cost of preserving versus the cost of re-creating. There are also ethical issues involved in data from experiments involving human beings, and, at least arguably, animals. Most observational data, once gone, cannot be replaced. We also have great problems quantifying the likely benefits of preserving various kinds of data. Some of the best thinking currently revolves around thinking in terms of ten or twenty year re-evaluation cycles for data stewardship, where at the end of a given cycle stewardship might transition from one responsible party to another in an orderly way. Cultural memory organizations have considerable expertise to contribute in managing this process.

How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Greater clarity about the extent to which data stewardship costs can be included in

grant budgets would be very helpful. We also badly need funding mechanisms to address the existing base of data that has already been created.

But I want to stress that this is not simply a matter of funding mechanisms - I think that federal funding agencies need to be clearer, as a matter of fundamental policy, that they share in the ongoing fiduciary responsibility for stewardship of data that is created as a result of their research funding.

What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

There are several approaches that should be applied in parallel. Investigators can be encouraged to ensure good data stewardship by asking about data specifically as part of the reported results of previous federal research funding supplied with new grant applications. To the extent that operational responsibility for stewardship and access are shifted to institutional actors (institutional and disciplinary repositories) and away from individual investigators, it should be much easier for major funding agencies to measure and verify compliance on a programmatic basis.

What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Clearly, discovery systems that can look broadly across the repository infrastructure are a key investment here (the search work going on as part of the NSF-funded DataONE datanet grant looks to be a very promising contribution). Beyond this, one could imagine SBIR-type programs to target small business investment to exploit available research data. Speaking personally, I would love to see some sort of prize competition making awards annually for the most creative and highest impact reuse of publically available data in commercial, educational, and research categories.

What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported? Work here is well underway in initiatives like DataCite. The federal funding agencies can help a great deal by asking investigators about datasets that they have created and made publically accessible, and about the impact that the availability of these datasets have had. They can also help by encouraging research proposals that make creative reuse of existing datasets, and by encouraging review panels to consider whether proposals are making effective use of existing data resources.

Standards

I want to first recognize that these are essential in effective data sharing, reuse, and stewardship, and that lack of appropriate standards is a real barrier. Standards evolve, and there is a continuing need to develop or update standards to reflect new scholarly practices. This is a high-leverage area that I think has suffered from chronic lack of funding - there's no clear source to fund the development of standards that are needed to support most data intensive scholarship, or to maintain those standards, or to finance the necessary software development or maintenance/upgrades to make the standards part of the community's tools and workflows. (One rare and noteworthy exception to this has been the INTEROP program that has been part of NSF's Office of Cyberinfrastructure for the last few years.) There's also a lot of experience in standards development, but yet all too often specific scholarly communities establishing standardsf for the first time seem to be re-inventing the wheel. Standards development and support are an area where I believe modest investments by funders would have high payoffs.

COMMENTS on Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research

January 12, 2012

Submitted by Rebecca Kennison, Director, Center for Digital Research and Scholarship, Columbia University (e-mail: rkennison@columbia.edu)

On behalf of Columbia University Libraries/Information Services

New York, New York

Comment 1

The key to encouraging public access to and preservation of digital data is the creation of open data repositories that adhere to common standards for the identification, description, and storage of data, coupled with a policy framework for funding agencies that requires the deposit of data from funded research in such repositories wherever possible, as well as a clear framework for communicating the usage rights for that data.

Given the variability of agency funding (the termination of the National Biological Information Infrastructure [NBII] Program this month is a case in point), the wisest policy is to encourage the growth of existing repositories and the development of new ones that will be managed by individual academic institutions, consortia, and/or scholarly societies in partnership with government, rather than by any individual government agency alone. The National Digital Information Infrastructure and Preservation Program (NDIIPP), which has now become the National Digital Stewardship Alliance (NDSA), offers an excellent example of how agency funding can be used to enhance existing infrastructure investments and encourage inter-institutional (and international) collaboration on data repositories, while the new International Standards Organization standard ISO/DIS 16363, based on the Trusted Repository Audit and Certification (TRAC) checklist (http://public.ccsds.org/publications/archive/652x0m1.pdf), will provide a benchmark against which data repositories will be able to measure their performance. The establishment of baseline metadata requirements for interoperability will also be a key area where agencies can provide leadership, working closely with discipline-specific groups such as professional and scholarly societies, information technology specialists, librarians, and research administrators to ensure that the data in these repositories are stored and described in ways that enhance their discoverability, as well as providing for machine-readability, which will be essential to support the creation of portals that aggregate data from multiple repositories.

Building data repositories and aggregation services alone, however, is not a sufficient step: funding agencies such as the National Institutes of Health (NIH) and the National Science

Foundation (NSF) must move to adopt policies that require researchers to, wherever possible, make their research data available in such repositories to enable public access and reuse. That some agencies have instituted data management plan (DMP) requirements for funding applications has been an important first step in that direction, but further compliance mechanisms will be needed. Those mechanisms should be closely tied to existing workflows for grant management so that important stakeholders, including sponsored projects administrators, repository managers, granting agencies, and the researchers themselves, are minimally burdened by these new requirements, thereby reducing both administrative costs and obstacles to compliance.

These mechanisms will also need to be tied to shifts in the workflows for publication and dissemination of research more broadly speaking. The inability to reproduce and verify research results that serve as the basis for scholarly articles, for example, is a major limiting factor in the efficiency of scientific research, particularly with regard to technology transfer. Since access to research data (and the software that are used to analyze them) is absolutely necessary to ensure reproducibility, there will need to be identification and description standards built into the compliance process that ensure that data are clearly associated with the publications that cite them and the code used to process them. The work of DataCite (http://datacite.org/) offers a promising model in this regard because it makes use of widely-accepted standards such as Digital Object Identifiers (DOIs) to facilitate the discovery, reuse, and impact tracking of data. The work the NIH has done to integrate compliance for published articles based on NIH-funded research into standard publication and grant workflows offers a model for similarly handling data compliance tracking.

Agencies overseas, including the United Kingdom's Joint Information Systems Committee (JISC), the Open Knowledge Foundation, and the Australian National Data Service, also have ample experience in these areas, having undertaken major initiatives in the areas of open data in the past decade. In addition to offering useful models upon which we in the United States can build, their work has begun to demonstrate the wide-ranging and significant economic benefits of open data. A recently commissioned Australian report on public sector information (Houghton 2011: http://ands.org.au/resource/houghton-cost-benefit-study.pdf), for example, found that the benefits for the Australian Bureau of Statistics of moving to Creative Commons licensing for its data would outweigh the costs (including foregone revenues) by 5.3 to 1 and those for the Office of Spatial Data Management and Geoscience by an amazing 15 to 1. A 2008 ACIL Tasman report (http://www.anzlic.org.au/Publications/Industy/251.aspx) suggested that increased public access to spatial data alone brought about a 0.6% to 1.2% boost to Australian GDP and a comparable boost to real wages, as well as a decrease in the trade deficit and an increase in household consumption. Using an econometric methodology, Houghton and Sheehan (2006: http://www.cfses.com/documents/wp23.pdf) determined that a mere 5% increase in access to United States government–funded research results would have produced an additional $2 billion in economic benefits in 2003 alone. Based on 2010 values, that benefit would have been over $7.5 billion, which, using the GDP-to-total-employment ratio, would translate into an additional 700,000 jobs, and, cross-referencing with data from the Bureau of Labor Statistics, a 0.5% decrease in unemployment.

Clearly, then, this is an opportunity that should not be lost: both the path to take and the benefits that will accrue are right in front of us.

Comment 2

The primary challenge here is ensuring that the intellectual property status of data is clearly communicated. For example, works directly produced by the Federal government and its various agencies are part of the public domain as a matter of course, but without their being labeled as such, we have observed that specific uses of the data remain unknown to commons users, thereby thwarting the original intent. Likewise, in the United States, data themselves are not subject to copyright, nor should they be, as any change in their current legal status would impede the technology transfer process and lessen its attendant economic benefits. However, users are frequently unclear as to the disposition of data, a situation that is further complicated by the differing legal frameworks for data in other countries and regions, particularly where data are copyrightable.

Policy in this area, then, should focus on encouraging the clear labeling of data rather than on interferences to the United States' stated position on the noncopyrightable status of data so that all stakeholders — researchers, funding agencies, repository administrators, members of the public, and so on — are aware of the use conditions of a given dataset. That labeling should be done in a human- and a machine-readable format, such as that developed by the Creative Commons (http://creativecommons.org/), with an awareness of the global context in which data live today, as well as of the ever-greater speed with which novel uses for them emerge. It is also vital that government agencies encourage the use of such labeling at all steps of the data lifecycle, since raw data may go through many transformations before they find their way into publications and other end-uses, but the ability to trace those data end-to-end will be an essential part of the verification process.

Comment 3

Federal agencies can expend energy and resources effectively in directing the management of data by embracing and incentivizing cross-institutional partnerships that distribute responsibility and value equally and fairly. The government should certainly assume a role in directing the message about the importance of research data as a first-class scholarly resource, and this would be well expressed in direct and ongoing support for these partnerships — anchored by research institutions with specific domain expertise in concert with the representative scholarly societies, whose membership have the disciplinary focus necessary to identify and address issues of data heterogeneity in the areas of size, sensitivity, format, and long-term value. The library research community, by virtue of its station in cross-institutional information service provision, has been particularly active in assessing the intrinsic issues produced by varying disciplinary data management needs. Federal agency action should therefore be properly informed by

existing studies, including the NSF-backed Data Conservancy project (http://dataconservancy.org/objectives) and the Institute of Museum and Library Services (IMLS)–funded Data Curation Profiles work (http://datacurationprofiles.org/).

Perhaps even more instructive than the differences across disparate disciplinary traditions, however, are the similarities apparent in digital research data that make certain functions common regardless of discipline. The work to be done by federal agencies in promoting and incentivizing best-practice solutions for data archiving and preservation are two such critical vectors. These areas, in particular, may require discrete, focused attention and fostering by federal agencies, as researchers continue to direct limited resources on dissemination goals and problems of access, when they address data management issues at all. A successful plan to address data archiving and preservation will tackle questions of governance, adoption or development of standards and conventions among disciplinary communities, and necessary new investments in technological infrastructure that make data management possible.

Comment 4

At this stage, differentiating between the relative costs and benefits of different data types is premature; the similarities between data types are more significant, and cost–benefit data are limited. What we do know from studies in the United Kingdom and elsewhere is that the costs tend to be localized and front-ended, while benefits accrue broadly and over an extended timespan. Some 42% of costs are associated with pre-archival processing and ingest, while storage and preservation account for 23% and access accounts for 35% (Fry et al. 2008: http://ie-repository.jisc.ac.uk/279/2/JISC_data_sharing_finalreport.pdf). Benefits, on the other hand, tend to be wide-ranging and long-term, beginning with the immediate savings produced by avoiding duplicative research and reducing the costs of access to data, followed by the near-term increases in efficiency, and, ultimately, the spillover effects — that is, the broader social and economic gains that result from improved access and that increase over time. Data gathered by the Census Bureau, to take just one US example, are of use to many other government agencies, not to mention a wide range of commercial and non-profit enterprises, and their value is greatly enhanced by their longitudinal nature.

Only by integrating the costs of long-term stewardship and dissemination of data into the granting process will it be possible to gather enough information to allow for a proper consideration of the relative costs of various data types, which will be a prerequisite for an evaluation of the full benefits as well. What is certain, however, is that the overall economic benefits will outweigh the costs. Indeed, as detailed above (see Comment 1), economic studies repeatedly indicate that while the rate of return on investment may vary, the benefits of making data more broadly accessible routinely exceed the costs, even when foregone revenues are taken into account.

Of course, formally integrating these costs into the granting process, while an absolutely vital first step, will not be the end of the story. In addition to setting the stage for further evaluation of the costs and benefits of different data types, agencies must pay attention to the ongoing, unanticipated costs of data stewardship and create mechanisms for meeting those emergent needs that cannot be integrated into and accounted for in the existing grant funding workflows.

One type of data that may be worthy of special consideration from the start is so-called "big data": that is, datasets that are on the order of tera- and petabytes rather than mega- or gigabytes. Due to their enormous size, these data are exceedingly difficult to disseminate via widespread computing networks, even the high-capacity Internet2. In some instances, they are only disseminated on storage media that can be shipped around the country or overseas; in extreme cases, they cannot be disseminated at all, and researchers must visit them directly in order to gain access to them. By adopting policies that encourage the development of functional access solutions for big data, such as migrating large datasets to NoSQL datastores that allow for efficient querying and expanding existing shared computing network infrastructure, funding agencies could have a major impact on the efficiency of scientific research, as well as opening the door to innovative small businesses that cannot afford to run their own high performance computing (HPC) centers.

Comment 5

The better question here is what have proactive stakeholders already been doing to contribute to the implementation of DMPs. Here at Columbia University Libraries and Information Services, for example, we have been working with the Office of Research to coordinate education and outreach efforts on the NIH and NSF DMP requirements, establishing workflows and developing resources to serve the data-archiving needs of Columbia-based researchers, and collaborating directly and indirectly with leading science data groups on campus such as the Center for International Earth Science Information Network (CIESIN) and Integrated Earth Data Applications (IEDA) to ensure that we understand their data preservation needs, collaborations that have led to our partnership with the Socioeconomic Data and Applications Center (SEDAC) on a long-term archive for their data (http://sedac.ciesin.columbia.edu/lta/index.html).

On a national level, there are projects such as those already referenced above: the IMLS-funded Data Curation Profiles project; the NSF-funded DataNet projects, including the Data Conservancy; and the National Information Standards Organization's Institutional Identifier (NISO I^2) project (http://www.niso.org/workrooms/i2). On an international level, we see the emergence of new standards, such as the Open Researcher and Contributor ID (ORCID: http://orcid.org/) and DataCite, that are being developed and supported by a wide range of stakeholders, including researchers, publishers, funding agencies, and research institutions.

The key here is to document the best practices and standards that are emerging and to foster existing systems that already serve research communities. Federal agencies can play an important supporting role in this area in a variety of ways, from funding projects to develop models for data services, to convening meetings to facilitate the codification and dissemination of best practices, to requiring adherence to those best practices and tracking that adherence as part of the grant compliance auditing process.

Comment 6

The first step here is to formally build the funding and tracking of research data costs into the DMP requirements that already exist. As part of that policy shift, agencies will also have to provide greater guidance to applicants to ensure that they are accounting for those costs as best they can and coordinating with the appropriate stakeholders, particularly data repository managers. At many larger research institutions, libraries and other information services providers are already offering what guidance they can on these issues, so there are models that can be built upon; however, guidance coming directly from the funding agencies will carry greater weight with researchers, and it will allow for the development of clear standards that will result in greater uniformity in the resultant cost and benefit data.

Beyond the absolutely vital need to acknowledge the real costs of data management formally by building those costs into research proposal expectations, it is important to recognize that it is not possible to completely account for those costs in advance. In the long term, it is virtually certain that unexpected events will occur that require significant new investment to ensure the ongoing integrity of data. Data migration is a particular challenge that can result in major one-time infrastructure costs: witness the expense of digitizing print media. However, as print digitization projects have made clear, there can also be unexpected benefits to such format migration. The key, then, is for funding agencies to be alert to emerging challenges and opportunities and provide data repositories with the resources they need to meet them.

Comment 7

Automation will be essential to minimizing the burden of compliance. Such automation requires that compliance be integrated into existing workflows (for granting, research, and publication/dissemination) and that it be based on clearly communicated standards. This means that funding agencies will need to improve the instructions they provide to grant writers as they craft their DMPs; it also means that researchers and other stakeholders will agree on basic standards for the identification and description of data, though those baseline standards should remain minimal, with specific disciplines having room to establish their own, more granular standards to meet discipline-specific data and metadata needs.

Comment 8

There are three primary areas in which wider availability of research data will be of short- and long-term benefit to the economy:

- Improving education (both K–12 and postsecondary), to ensure that there is a ready supply of highly skilled individuals ready to enter the STEM workforce;
- Increasing the speed of scientific innovation, to provide for the creation of new technologies and improvements to existing ones;
- Encouraging the growth of small business, by lowering the barriers to entry for high-tech industries and increasing overall competitiveness.

Individual government agencies will have an important role to play in encouraging the benefits in each of these areas, particularly by creating data-aggregating portals that provide a unified point of access to disparately archived data in order to best serve specific stakeholders; as Houghton's models show, the key to increasing the benefits of open data is to provide for the maximum access to that data. Even an increase of 1% can translate into millions of dollars a year. Therefore, efforts like those already undertaken by the Small Business Administration (SBA) to create application programming interfaces (APIs) for other kinds of data should be taken as models for all agencies. Given the SBA's collaboration with NASA's Small Business Innovation Research/Small Business Technology Transfer (SBIR/STTR) group, there are clearly already frameworks in place that would allow for the rapid development and dissemination of tools to maximize the positive economic impact of open data.

The Data.gov portal is another important platform for emerging and new markets. Because of its role as a clearinghouse for open data, it is especially useful for attracting interest from application developers, and thus for generating new technologies and new businesses that can provide additional services to the public. Thus, integrating open research data into Data.gov would offer greater potential for spillover benefits, particularly from unforeseen uses of research data, including but certainly not limited to data mash-ups, and innovative (and perhaps serendipitous) discovery resulting in new products.

Comment 9

There are at least two such mechanism-types that could be brought to bear on the attribution of data for secondary research: (1) the improvement and disciplinary standardization around persistent identifiers for data, institution, and researcher; and (2) the incentivizing of the practice of data citation in a way that raises it to the scholarly standard of publication citation already commonly well-observed. To the former, the issue of identifier persistence is one that has been taken up by several parallel groups (ORCID for researchers, NISO I^2 for institutions, and DataCite for the data themselves). Through identifier persistence, explicit stakeholder credit may be tied directly to unique data assets, which strengthens not only the potential of explicit attribution but also for the potential to trace reuse through citation, a significant value proposition to researchers looking for the

rationale behind data sharing. Rather than look for new mechanisms to begin developing, therefore, a powerful way to support existing momentum would be direction for federal agencies to provide incentives around their use — much in the model of the NIH in its program to ensure funded research publications in the PubMed Central repository are also cataloged within the freely accessible PubMed database. Further, an early driver of potential movement could come from federal agencies to catalyze research around publicly accessible datasets, with specific caveats about the methods of attribution to be adhered to. Such programs would reward early adopters of best practices in data management while contributing to the hoped-for sea change in source acknowledgement in secondary research.

Comments 10 and 11

International standards efforts, such as the recently approved standard ISO 16363 (Space Data and Information Transfer Systems — Audit and Certification of Trustworthy Digital Repositories) and the development of the proposed standard ISO 16919 (Space Data and Information Transfer Systems — Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories), offer one approach to improving the likelihood of success of standards development processes. Both efforts were conducted under the guidance of the Consultative Committee for Space Data Systems (CCSDS) by a diverse, international team of experts in archiving and data stewardship. Team members developed their recommendations independent of influence from government agencies or proprietary interests, which enabled the team to address key standards issues without being compromised by other influences. The development of these standards was encouraged and supported by the institutions that employed the team members and by the funders of those institutions. Such support and encouragement enabled the experts to collaborate as a cohesive team over extended periods of time, ensuring that the process had continuity in terms of the stakeholders and the expertise brought to bear.

Many of the standardization efforts around attribution and credit for data reuse (noted in the response to the previous question) will apply here as well in the context of interoperability, and our suggestions regarding the support of those standards should be carried over. Further, there are many established disciplinary data and metadata standards and communities of which of the RFI reviewers will be well aware — the Flexible Image Transport System (FITS) for astronomical data, and the Federal Geographic Data Committee (FGDC) and the Open Geospatial Consortium (OGC) for geospatial data, are several that have special relevance and utility to the Columbia University community. The multiplicity of these standards prohibits a comprehensive response here, although we do join in the endorsement for a centralized index of such standards, believing that such a resource could foster adherence to community practice and reduce barriers to interoperability.

Perhaps more significantly, however, we invoke once more the call to involve the scholarly and professional societies in a direct way in the identification and development of these domain-specific digital data standards and of the data repositories themselves. As both liaisons among and representatives for their constituencies, societies are equipped to deal with the inevitable idiosyncrasies of the data in their domain. Empowering these organizations (again, through incentives articulated centrally through individual agencies) thus strengthens their positions as arbiters of authority and respects the individual established contexts, initiatives, and standards.

Comment 12

Given that such initiatives to coordinate on data standards are presently underway through several international bodies, including the ISO and the International Council for Science (ICSU) and its Committee on Data for Science and Technology (CODATA), we can advise in the first place for federal support for the activities of these groups and others of a disciplinary orientation as they represent national interest in international consensus. Further, the direct involvement of the National Institute of Standards and Technology (NIST) in matters of standards-making with our global counterparts would be a natural activity, and we can advise for further federal agency involvement in related activities. This can be understood to mean a push for proactive engagement of national agencies with initiatives of nations, some such as the United Kingdom whose work on data standards is tracking very closely with activity in the United States. The relevant National Research Council boards may also play a constructive role in this arena, e.g., the Board on International Scientific Organizations (BISO) and the Board on Research Data and Information (BRDI).

Comment 13

For the appropriate association between research article and dataset to be made, it is not enough that digital research data be attached as supplementary material to published research articles. It is, in fact, the research article that is the supporting documentation of primary research data, although the infrastructure around the publication and archiving of written material is much more mature than that around datasets. The research article itself may come to be supplemented by secondary research against the original dataset or only a subset of it, but in any robust future scenario, the dataset needs to occupy a position of importance above that of supplementary material. The work advocated in the responses to the previous questions — particularly in the support of archiving and preservation environments for datasets, as well as for community coalescence around persistent identifier schema — are important parts of the publication/data association infrastructure. Look to the German academic institution-based Publishing Network for Geoscientific and Environmental Data (PANGAEA, a DataCite member: http://www.pangaea.de) and its formal affiliations with Elsevier for an example of how linkages and partnerships may begin to be realized in practice (http://bit.ly/9SSkHQ). The work of the international data

repository called Dryad in facilitating data and publication links in additional to data reuse is also particularly instructive (http://datadryad.org/).

12-Jan-2012

TO: Science and Technology Policy Office

FR: Edward Van Gemert, Interim Director, University of Wisconsin-Madison Libraries and Bruce Maas, CIO and Vice Provost for Information Technology, University of Wisconsin-Madison

Response: Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research
RFI Document Citation: 76 FR 68517
RFI Document Number: 2011-28621

The University of Wisconsin at Madison's General Library System and the Office of the CIO, in consultation with its Research Data Services staff, strongly endorse OSTP's interest in preserving and providing public access to digital data from federally-funded research. Responding to OSTP's specific questions:

(1) Blanket preservation policies should apply to digital data arising from Federal grants. These policies should define as clearly as possible, taking disciplinary differences and research workflows into account, which data are of sufficient importance, quality, and reusability to warrant the cost of preservation. Policies should authorize either specific preservation retention schedules or periodic reassessment of preserved datasets so that obsolete ones may be discarded. Access policies, which must be considered separately from preservation policies, should require public access except in clear cases of human-subjects confidentiality, national security, or similar. Institutional Review Boards may be suitable arbiters of human-subjects questions surrounding access to data, but clear Federal guidance will help them considerably.

Granting agencies requiring data management plans should strive for consistency in terms of the data plan requirements, with each plan addressing data preservation, data security, and access. To the extent possible, such consistency will encourage easier compliance resulting in improve access to a greater amount of material over time. These requirements should be integrated in the grant submission guidelines, clearly outlining the purpose and elements of the data plan. At the time of the award, grant recipients should have a documented and clear understanding of their responsibilities with respect to data retention including retention schedules, which data are to be retained (e.g. raw data, summaries, etc.), access rules, and so forth. An additional suggestion pertaining to IRB policies should be considered: as part of IRB policies, study participant consent forms should provide information indicating that certain data they provide could be used in other contexts.

(2) It is vital to remember, and for Federal policies to state clearly, that many datasets do not meet the originality standard for copyright. For such data as do have copyright or patent encumbrances, however, and to accommodate most disciplinary cultures, Federal policies should allow delayed (but not indefinitely-delayed) public access to data. Deposition into suitable data archives should be as immediate as possible, as this best protects dataset viability, but Federal policy should permit embargoed access until after publication, after patent application, and/or for a discrete length of time after grant end. Federal agencies should insist that data be licensed for reuse, commercial and non-, via licenses such as the Open Data Commons Public Domain Dedication and License (opendatacommons.org/licenses/pddl).

(3) The National Science Foundation's implementation of its data-management-plan policy is an excellent example: the NSF's broad policy guidance has been interpreted and expanded upon by each directorate in disciplinarily-appropriate fashion. Federal standards agencies may also wish to endorse suitable data and metadata standards that arise from research and library communities and informatics initiatives.

(4) This question is extraordinarily complex and difficult, and of course discipline-dependent. One relatively simple answer would be to track dataset reuse, and publications based on given datasets, plotting these data against cost data to decide about continued preservation. We also hope that federal agencies will continue to play an active role in funding research pertaining to long-term sustainable data standards and formats given their potential to reduce the costs of storage, facilitate discovery, and improve upon the interoperability of research data sets from heterogeneous sources.

(5) We believe that roles and responsibilities around data preservation and access are very much in flux, and that this very uncertainty is contributing to valuable research and innovation in both the public and private sectors. We therefore suggest that Federal policy mandate *ends, not precise means to those ends*, whenever possible. Here at the University of Wisconsin-Madison, researchers, librarians, the School of Library and Information Studies, and IT professionals are working together to raise consciousness of data-management issues and provide expert consultation and training in responsible data stewardship. We believe that our year-old Research Data Services (researchdata.wisc.edu/), while not a comprehensive solution to the broad panoply of data-management challenges, is a promising example for other stakeholders.

(6) The research and library communities frequently lament that research grants are of finite duration, while preservation responsibilities last indefinitely. Moreover, some researchers perceive preservation costs as subtracting from the pool of research funds available, and may oppose data preservation policies on that basis. Federal policy should therefore consider strategies for ensuring that preservation is

considered during the earliest stages of grant development. We encourage Federal agencies to:

- regularly and consistently fund national-level disciplinary data centers, existing and new;
- provide portable funding sources to "endow" preservation of and access to specific datasets;
- provide separate budget lines and guidelines to fund preservation and access, rather than lumping them in with overhead costs; and,
- authorize or mandate the engagement of data management professionals as part of the grant submission process. Clarify what is meant by "incremental" costs for data management and specific types of costs agencies are willing to fund (e.g., costs for storage, backup, consultations, metadata development, etc.). It is presumed that funding applied to data management services would enable institutions to grow their cyberinfrastructure and expertise, which in turn, would enhance a given institution's ability to assist PIs in their efforts to be responsible stewards of data generated in federally funded research.

(7) Federal data-management policies should insist that persistent, Web-compatible identifiers (such as DOIs, ARKs, PURLs, and handles) be provided to grant agencies for applicable datasets, much as the NIH Public Access Policy now insists upon PMCIDs/NIHMSIDs in grant reports and subsequent grant applications. Data archives should provide identifiers for embargoed datasets, and be willing to certify to Federal agencies that the dataset is indeed present in the archive. Grant agencies should develop policies that clearly articulate preferred repositories which will aid said agencies with respect to auditing and other compliance issues.

(8) Data registries help connect data creators with data users. Quite a few state and local governments have successfully stimulated dataset-based innovation by holding developer contests, as well.

(9) Dataset and author identifier-assignment and citation standards are under construction, notably the ORCID (orcid.org) and DataCite (datacite.org) efforts. Funding these standards, and insisting they be employed in communication with Federal agencies around grants, will help assure appropriate attribution and credit.

(10) Almost any digital-data standards will be helpful! Presently, many disciplines utterly lack such standards; others have developed them, but not managed to implement them discipline-wide owing, in part, to lack of incentive or funding for researchers to use them. Federal attention to developing, promulgating, and insisting upon use of standards should be a clear priority! Discipline-specific standards developed in cooperation with (or by) researchers in those disciplines are more likely to gain adherents, thus building momentum around a given standard's adoption from funding agencies, publishers, and professional societies. In turn, widespread adoption of standards will clearly enhance our collective ability to

provide for the preservation, discovery, and reuse of research data within and across disciplines.

(11) Standards sometimes arise from a widely-acknowledged need to share data, as happened with the International Virtual Observatory Alliance (ivoa.net); they also spring naturally from the establishment of discipline-dominant data repositories such as the Interuniversity Consortium for Political and Social Research (icpsr.umich.edu) and Long Term Ecological Research Networks (http://www.lternet.edu/). Should Federal policy jumpstart broader data sharing as well as more disciplinary-data repositories, standards development is likely to follow naturally. That said, Federal policy can help by providing funds for standards development, and one or more registries of relevant standards.

(12) International standards coordination is the natural role of Federal standards bodies such as NISO and ANSI, as well as the Library of Congress.

(13) As mentioned in our response to question 7, persistent dataset identifiers are a necessary prerequisite to citation. We do not believe Federal policy need endorse one identifier scheme over another; a list of acceptable identifier types will do. Citation of datasets from published papers is a somewhat harder problem, governed as it is by style guides firmly mired in the 20th century. We suggest instead that Federal policy require a set phrase with a list of dataset identifiers for papers published from Federally-funded research and datasets, much as is often done now for acknowledgement of Federal grants in published papers.

Dr. Melissa Haendel, Ph.D.
Assistant Professor
haendel@ohsu.edu
Oregon Health & Science University
Portland, OR

On behalf of the Resource Discovery Group, a consortium of researchers from eagle-i (https://www.eagle-i.net/), Vivo (http://www.vivoweb.org/), the Neuroscience Information Framework (NIF; http://neuinfo.org/), Biositemaps, and the CTSAs, whom are interested in promoting research resource representation and discovery in the scientific enterprise.

Preservation, Discoverability, and Access

> *(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?*

Federal agencies should create a technical standard that enables discovery, usability, attribution, and long-term preservation of digital data. These specifications need at a minimum to include the archiving of data in publically accessible repositories, using standard record and metadata formats, and promoting best practices for interoperability and reuse, such as Semantic Web standards and Linked Open Data. Once the technical standards are there, policy can be established that requires data to be made available in a compliant manner as a deliverable of all federally funded grants and contracts, not only for those over $500,000. Grants with a data-sharing component should have a required budget line item for data sharing and archive.

A critical aspect of this policy will be to define “digital data” in the context of the policy. Funding agencies can support researcher efforts to meet the policy requirements by integrating semantic reference to these digital data into grant application and reporting structures. With appropriate tactical issues worked out, funding agencies could partner with publishers to require (and verify) data sharing before research results can be published. Finally, award and incentive systems (including institutional APT committees) must recognize the value of quality data management and sharing to the scientific enterprise.

It is estimated that it costs $24,100 and from 1.5 to 3 years to develop a transgenic mouse from scratch (eagle-i, unpublished economic analysis). What if that mouse were available to the research community at or during its development? This could expedite both public and private research endeavors. One of the issues is that “this” mouse is neither shared nor represented in a standardized manner such that it can be found for general reuse. It is not until a curator at a specialized database sees it in a publication, that it becomes part of the public record of available resources- sometimes years after it was developed. The point here is that the metadata about research resources themselves is digital data, and standardized representation and sharing of research resources should be included in any digital data policy. It should be noted that a lack of data annotation and sharing may not be for lack of desire to do so. Funding agencies, libraries, and research offices should offer training and helpdesk facilities to educate researchers in best practices for data annotation and sharing.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

One issue is that currently it is largely only publications and patents that are attributed. The scope of attributions needs to expand. Researchers need the ability to access the components within the publication (e.g. a knockout mouse, viral vector, database, datasets, etc.) This will protect the interests of individual stakeholders and they will feel more inclined to share these important and relevant outcomes of the scientific enterprise. Specifically, data sets can be citable, authored sets of information that can be referenced in the context of publications, grant reports, etc. While mechanisms are underway to support such efforts (Bioresource Research Impact Factor, Beyond-the-pdf, nanopublication), it will not be until funding agencies, employers, and publishers consider such citations in the context of evaluating a candidate proposal or manuscript that they will be adopted.

However, federally funded research produces data that is generated using taxpayer money and it belongs to the people. The person who generated it has no intellectual property interests on the *data.* What they do with it is a different matter, and they should be given some amount of time to do something with it (publish, patent, market, etc.)- 9 months or a year, perhaps.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

In developing policy that accommodates differences between scientific disciplines, libraries and information science researchers that are accustomed to providing guidance and resources for disparate kinds of data should be consulted. While data differs in different disciplines, there are qualities common to all data types, and these should inform inter-disciplinary requirements. For instance, there exist upper ontologies that represent the types of things that exist. Classification of data elements can be tied to such upper ontologies via reuse of these upper ontologies. One example is the Basic Formal Ontology as the upper level ontology for all Open Biomedical Ontologies (OBO; http://www.obofoundry.org), which enables representation of a catheter, a zebrafish liver, diabetes, and regulation of cell adhesion. These entities may not on the surface appear to have anything in common, but use of a common upper ontology can facilitate data integration about all of them (for example, in the context of designing an experiment). However, it is equally important to consult the end-user who is attempting to query across disciplines to ensure data consistency of representation. To this end, existing discipline/data specific repositories should also be consulted to ensure applicability. Furthermore, to support innovative reuse of digital data, it is important to recognize that these uses are not usually the original creator's intent. Data from disparate disciplines, projects and sources can be combined for synthetic and synergistic scientific inquiry - this in itself will also support new markets. Interoperability standards will benefit these new applications. Therefore, each discipline may require specialized data formats, queries and applications, but federal agencies can promote open and extensible standards to meet cross-disciplinary needs.

Another facet of this that must be considered is the extent to which there exist different data sensitivity issues in different fields. For example, publication about uranium enrichment metadata may require different consideration than data on the Arabidopsis genome.

> *(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?*

The most important aspect of what federal agencies can do with respect to garnering an understanding of the cost benefit analysis of stewardship and dissemination is to promote scientific inquiry that depends on public digital data. It is currently difficult to obtain funding for such projects, and as such, it remains somewhat of an idealistic rationale that publication and availability of data will be good for the research enterprise. In fact, we know that it is difficult to reuse others' data without standards, and it is often more cost-effective and time saving to create one's own data. If we are to tip the scales and actually save time and money, it will be because there exist standards and requirements to promote data reuse. Such requirements can be met via interagency collaboration, standardization, and cost sharing. In doing so, there is the potential to control costs and maximize benefits by limiting duplicate efforts, distributing responsibility, and by educating researchers. Furthermore, with respect to research resources, there is a clear indication that reuse of such entities saves time and money. If standards were promoted to enable their identification and relevance, and researchers incentivized via funding streams to leverage preexisting resources, this could lead to a very solid understanding of cost-benefit to sharing digital data for these particular data types.

> *(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?*

Participation by the many stakeholders must be regulated by technical and legal standards to ensure and promote free public access, discovery, re-use, and preservation. The expertise and methodologies of these stakeholders should be leveraged collaboratively both in the development of policy and in its execution. Such collaboration is required for success, and can drive best practices, innovation, market creation, and compliance.

The present repositories of research communities, publishers, and institutions can be utilized and developed (e.g. Pangea, TreeBase, eagle-i, NIF, Biositemaps). Existing partnerships between publishers and repositories, such as Dryad, can be grown. Organizations like DataCite and BioCoreDB work to improve the discoverability and utility of data. However, none of these systems alone will be successful until they are integrated into the research workflow. It has to become easy to submit data to such repositories in the context of publishing manuscripts or submitting grant reports. These repositories must also supply the submitters with some form of unique identifier. These identifiers can be used to track submissions and, eventually, resource usage.

Universities, research institutions, and libraries will need to play a key role in building infrastructure to support their researchers' compliancy and education, as with NIH public access policy, and guiding archival and discovery standards. They must also include data sharing and stewardship as a component of performance evaluation, where applicable.

Libraries are also well positioned to enable these infrastructures to be compliant with the Semantic Web and population of Linked Open Data from these data sources.

> *(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?*

The scientific and economic reward of sharing digital data will not be realized until the cost of data management and preservation are built in as part of any research program (whether it be in the context of a grant, laboratory management, a library, etc). Funding agencies should require researchers and institutions to document the cost of data management and publication within their proposals and reports. As this would be a new policy, better guidelines for what types of data management and preservation are satisfactory should be developed. Included in these guidelines would be requirements for sharing metadata about research resources. These guidelines should also promote the collaboration and/or inclusion of information specialists or libraries in supporting this aspect of the research. Furthermore, agencies should consider funding information scientists and libraries to perform more research on making specific data types conform to standards and archived for maximum query potential. In summary, information scientists now are needed more than ever to be a part of the research endeavor rather than solely involved in after-the-fact archival activities.

> *(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?*

Interagency standards that offer practical workflows and mandate deposit in publically accessible repositories will improve compliance and facilitate verification. These standards would require that: Digital data and research resource metadata are deposited in publically accessible databases in conjunction with manuscript acceptance and final grant reports, and that standardization of data format and minimum metadata are applied and *verified*. Different levels of compliance would need to be defined. At the very minimum, data must be understandable and reproducible based on free text descriptions and "readme" type files. Higher levels of compliance, e.g. structured metadata to enabled querying and reasoning across datasets, would be optional.

Many publishers already require certain data sharing standards and yet authors do not always comply in spirit or in letter. In support of these standards, it is recommended that several submission workflows be supported, including third-party deposit. One very important aspect of this is to involve the publishers, in particular with the assignment of persistent, unique, and linked identifiers. Currently, a manuscript may be published wherein the subject is a unique gene (or some other common data element), and yet these elements are never uniquely identified. There **must** be a partnership between researchers, publishers, reviewers, and funding agencies to ensure that such entities are properly referenced. Only then will their reference be linked to the research landscape and enable maximal inference and discoverability. Perhaps even more importantly, only then will the research be reproducible. This is especially relevant in the context of research resources, where without reference to a specific resource ID (for example, an antibody ID) one will never be able to reproduce the experiment let alone find the resource. For verification, publication of digital data on the Semantic Web can further enable systematic review of the data.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Federal agencies could establish incentives to stimulate the use of research data through preferential review of research proposals producing and leveraging public data in addition to requirements for archiving data. For example, NIH grant guidelines could be modified to include leveraging public data as part of a grant's Approach or Environment scores. Application showcases (e.g., http://www.data.gov/developers/showcase) or contests can also raise the profile of public data and capture the attention of the media on data standards and public availability of data. Small-scale venture capital solicitations patterned on http://www.kickstarter.com and data marketplaces such as http://www.crunchbase.com offer models for value-added services on top of data where relatively small investments of capital could produce significant results in the private sector.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

As with manuscript publication, secondary results should cite primary data. Standardized unique data identifiers will enable identification and linking resource (be it data, research resources, etc.) to relevant documents, data, persons, and grants. The use of controlled author and institutional identifiers (e.g. ORCID registry, http://orcid.org) will be critical to support disambiguated and resolvable attribution. Furthermore, use of a common metadata standard to tag various kinds of data with appropriate attribution in a standardized way will ensure proper attribution. It is not always enough to know whom the data came from, but also the version, from where, and how is it related to other documents, data, experiments and grants. Simply stating the author and the year is not sufficient to understand the methodology or process in which the data was reused. These additional metadata could promote a standard for provenance, quality and trust of scientific data.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

First and foremost, a minimum attribution standard for any kind of content should be created. Anything that is reportable as linked to grant funding activity should meet this standard. Following that, it will be important to develop metadata standards that facilitate machine reading and Semantic Web linking of information. Such metadata standards can be high level, as per the upper ontologies mentioned above. Basically, what kind of resource is it? Who, where and when is it attributed to? What is it linked to? Following this, each discipline will have further requirements and standards to better inform reuse in those fields. However, a simple adherence and strategy for including the aforementioned metadata will support and inspire more extensive data annotation. In the context of research resources, we are working on a metadata standard

to this end. Publishers and granting agencies can adopt this metadata standard and provide guidance to contributors in support of meeting this new standard.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

There are successful standards developments in many domains. The W3C standards process has successfully produced HTML, XML, RDF and other languages. Key to the success has been its openness and community participation. Successful standards development relies on the contributions of a diverse population of experts, including scientists, information professionals, and technologists. It has to be field tested- if it is not useful or doesn't work for end users then it will not be adopted. If scientists themselves begin to reap the benefits of standardization, they will no longer feel the burden of having to comply. For example, if they can search all completed grants for specific research resources that may be advantageous to their work, and then find some that they reuse, they will not feel such a large obligation when it is their turn to provide the metadata necessary to make their own research resources available.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

A technical infrastructure that utilizes international standards for interoperability and re-use, such as Semantic Web Standards and Linked Data, should be adopted. Agencies can and should leverage the work of organizations focused on international data sharing and utility, such as CODATA, the Global Biodiversity Information Facility, the Open Archives Initiative, and the Digital Curation Center. It would also be worthwhile for federal agencies to participate in and support international efforts to connect data collections and build collaborative data infrastructures that aim to deliver cross-disciplinary data services. Similarly, adoption of other international efforts to standardize metadata, for example, coordination between VIVO and EuroCRIS, the European organization for international research information, will facilitate data integration internationally. Promoting such coordination as part of existing granting mechanisms or via new ones to promote international collaboration will be beneficial. Mechanisms could include specific RFAs for projects that coordinate internationally and hosting international workshops to bring these groups together.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

To facilitate linking between publications and data, the use of persistent, unique identifiers for data, research resources, publications, authors, and institutions is required. These identifiers should be unique Internationalized Resource Identifier (IRI)—the standard for identifiers on the World Wide Web, so that the data (or the metadata about the resource) can be made directly available trough the web. Unique identifiers enable visible links between entities, as well as re-use and the development of new services. For example, browsing a publication could include integrated data displays. In support of this functionality, we need standards for citing datasets and models, along the lines of what SageCite is working towards. Further, linkouts between publications, datasets and resources are needed; similar to the way they work today for genes. Clicking on links to research resources could take you to a place where you can obtain

the resource. Retrospective curation, at least for major datasets and publications (e.g. TCGA, Wellcome Trust, etc.) should be considered.

Disambiguation services such as the Virtual International Authority File (VIAF, http://www.oclc.org/research/activities/viaf/) offer a promising path forward for improving data quality. While VIAF focuses on organizations and people, other much lighter weight efforts could be established using open tools such as Google Refine (http://code.google.com/p/google-refine/) to support disambiguation web services from data repositories that could be integrated into desktop systems, websites, and publication submissions tools. Services that enable linking to data and linking both data and publications to known identifiers or terminology at the time of submission of a new publication could push much of the linking upstream to where incentives for documenting work are the highest.

Enabling such capabilities will require a new age of semantic awareness on part of the researcher, the reviewers and the publishers of manuscripts and data. Enhancing current research training to include modern information management strategies will be key, and funding agencies should support integration of information management into their research workflow.

Adrian Pohl
Thu 1/12/2012 5:07 PM
Response to the RFI on Digital Data

Dear people at the OSTP,

below are my answers to your questions on Digital Data. Again, I am responding as an individual working in an institution which provides information (research tools as well as licensed content) to academic libraries. Also, I am coordinating the Open Knowledge Foundation's "Working Group on Open Bibliographic Data" but cannot and do not speak for this group.

All the best
Adrian

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Response:
In summary our response advocates:
Immediate release
Disclosure of broad estimation of acquisition cost Proper open licensing Adoption of open standards for data files Adoption of extensible standards for metadata

Immediate Release
Federal agencies funding scientific research must establish policies by which the data acquired in federally funded scientific research
(FFSR) must be made immediately and fully available in public data repositories while ensuring subjects privacy.

The policies should follow the model of the Bermuda Principles (<http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml>).
In particular on:
- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute and reorganize the data.

Disclosure of Acquisition Cost
When reviewing proposals for funding opportunities, federal agencies should require that the sections requesting public funds for data acquisition activities provide a clear estimation of the cost of acquiring the data. If funded, researchers should be required to make data available in public repositories immediately after acquisition, and in the metadata used to describe a dataset, researchers should also be required to include the cost of acquisition.

The goal will be to develop a sense of the economic cost of not releasing data. For example, not releasing a dataset that cost $1M to be acquired is a loss for the federal government of the $1M funds provided by

taxpayers. This is the direct value lost from the overall economy; the actual value lost is much larger since it should include the missed opportunities that could have resulted from the exploitation of the data.

The European Commission, for example, recently adopted a policy of open data dissemination (<http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=EN&guiLanguage=en>).
The principle, rooted in the arguments that Yochai Benkler makes in his book "The Wealth of Networks" is that data is more valuable when shared; in economic terms, data is an "anti-rival good". It is a good that becomes more valuable, when more people have access to it and use it.

Proper Open Licensing
Current copyright legislation has been strongly focused on protecting the creators of artistic works, and in the process have created an inhospitable environment for the daily sharing of scientific information. The litigious behavior that many institutions have developed around copyrighted materials, results also in a reaction of over cautious behaviors on the part of the potential users of data and documents resulting from scientific research activities.

To dispel this environment of uncertainty, it is fundamental to clarify the rights of the public to make use of data acquired as a result of FFSR. The most effective way of achieving this goal is to affix to every released dataset, a clear statement of licensing indicating what the recipients of the data are legally allowed to do with the data. Licensing issues are expanded on in the Panton Principles for Open Data in Science (<http://pantonprinciples.org/>).

Federal agencies should identify a set of licenses that ensure the rights of the general public to deal with the data, in particular to copy, distribute, and create derivative works, and in this way ensure that the data get to reach their maximum economic potential to foster the growth of the U.S. economy.

Adoption of Open Standards
Federal agencies must ensure that data are released in a usable form.
The first step in that direction is to require the adoption of open standards for file formats, and forbid the use of proprietary formats that could prevent the general public from having access to the data.

Standards file format used for digital storage of scientific data are abundant and vary greatly from one domain to the next. Therefore, the scientific community will have to be engaged with the federal agencies in identifying the proper open standard to be used on each discipline, and to create new standards in the cases where no suitable standard file format exists yet.

For standards to reach their full potential, it is fundamental to have an open source reference implementation of the standard, and to encourage the development of a ecosystem in which commercial applications implement the standard as well. In this way, it becomes possible to maximize the use of the data acquired as a result of FFSR.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

In addition to the stakeholders listed in this question, it is critical to note that the general public is one of the primary (if not the primary) stakeholders to be considered here. Given that in the context of federally funded scientific research, it is the public's tax dollars that are paying for the scientific research being

undertaken, and thus the public's interest is the first one that should be considered when making trade-offs between available options.

Scientists who gathered data in federally funded scientific research did so as part of their job duties, and therefore under U.S. copyright laws they were performing "work for hire." This means that their employers are the copyright holders of any creative aspect of that data gathering (as pointed above, that only include the organization of data collections). Given that the scientists' employers received funds from the federal government, it should be expected that they will be subject to the same demands of the Federal Acquisition Regulations (FAR) as other contractors of the federal government. In particular with respect to the licensing of data acquired as part of federal contracts. The data should be published using an open license
(<http://opendefinition.org/licenses/#Data>) and at best follow the Panton Principles.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Working groups should be established for different disciplines, involving representatives of leading research institutions for each discipline.

Working groups should define differences with how the data are represented, indexed, stored and exchanged, but should not have the latitude to restrict in any way the free dissemination of information. All the policies should consistently have as a common factor the requirement for immediate and full release of data, unconstrained by any embargo periods or licensing restrictions. Credit for the acquisition of data could be ensured by data publications (eg
http://datacite.org) that can be cited by further works.

In this process, it is vital to invest in and commit to the emergence of standards that enable interoperability of, and thus reuse of, digital data. Linked Open Data standards for publishing (meta)data on the web build on central features of the Internet and the World Wide Web. As long as those data are in a tower of babel of formats, incoherent names, and might move about every day, they will be a slippery surface on which to build value and create jobs. Federal policy could call for a standard method for providing names and descriptions both for digital data and for the entities represented in digital data using URIs for identifying datasets and RDF and vocabularies like the DataCite metadata core for their description.

Standards also make it far easier to provide credit back to scientists who make data available, as well as increasing the odds that a user gets enough value from data to decide to give credit back. Embracing a standard identifier system for data posters will make it easier to link back unambiguously to a researcher as well as to make it easier for grant review committees and universities to receive a full picture of a scientist's impact, not just their publication list.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

The working groups in the different disciplines (from Question 3) should establish guidelines on practices for dissemination and storage for different types of data. For example, in genomics, it may be reasonable to store the secondary sequence information but not the primary sequence (given their great difference in data size).
Analogously, the guidelines may require primary sequences to be stored only for 2 years, while the secondary sequences should be stored for
10 years.

In astronomy it may be required that certain types of images be stored for different periods of time. Some images may be required to be stored with different compression ratios, and therefore correlate their storage cost with the potential expected benefit for future studies. In this cost-benefit evaluation, the original cost of acquiring the data should be taken into account. For example, a project that invested $50M in acquiring data should not attempt to make savings of a few hundred dollars in storage.

Economists must be involved in the working groups charted with the mission of providing guidelines for storage and dissemination, given that this is a problem in which the trade-off for the benefit of society at large must be continually evaluated.

The policies of federal agencies should be affected by the constant advances in storage technology and the rapid decrease in the cost of storage. The federal government should stimulate the development of storage technology, either by creating large storage decentralized facilities, creating consortia to manage data storage services, involving the public in facilitating distributed (and redundant) storage systems based on peer-to-peer technology that has already proven to handle large amounts of data.

All these guidelines should be prepared following open and transparent procedures in order to prevent proprietary standards and vendor lock-in situations that would prevent the policies from maximizing the utility of federally funded scientific research to the general public.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

They can join the working groups established in their respective disciplines of interest that will define practices for data management, including consortia combining universities, commercial companies and government agencies.

As standards and agreements are developed, working groups can help implement and test such plans in pilot projects. It will be of great help if federal agencies provide seed funding for these pilot projects.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Preserving and making digital data accessible is closely related to the issue of preserving and making scientific publications accessible.
If libraries and other non-profit organizations take over these tasks from the current commercial publishers as suggested in my answers to the RFI on scientific literature, there will be more than enough funds available from the current publisher profits to allow libraries to store and make digital data publicly accessible.

Once data and literature are stored in a database where both are linked semantically, innovators have a bounty of opportunities to provide commercial services and develop new applications and drugs/therapies to then generate a profit from.
In the current system, this information is restricted to a small set of academics, with innovators largely barred from access.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Once all data and literature are available to innovators, market forces should be allowed to take over without any additional policy interference, as the government is already funding the establishment of this resource.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

The DataCite initiative has been working for some years on providing answers to this question.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

I have also heard of Minimum Information for Biological and Biomedical Investigations (<http://www.mibbi.org/>) and Minimum information about a bioactive entity (MIABE) (<dx.doi.org/10.1038/nrd3503>).

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

A. Ensuring that Internationalization (of language and "locale") is made an integral part of the standards.

B. Starting with simple standards that can progressively be improved, instead of spending a lot of time in top-down design, committees and long-term procedural approaches to the definition of the standard. In other words, following the Agile methodologies that have proved to be successful in open source communities.

C. Working with existing international organization that have already defined standards in different disciplines, e.g. DataCite.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

At best, the Linked Open Data approach should be used. Publications should get a HTTP-URI as identifier and also their different parts.
Datasets should be assigned a HTTP-URI, e.g. a DOI like used by the DataCite project. OAI-ORE is a well-known standard for representing a complex publication containing datasets etc. in RDF.

Thu 1/12/2012 9:14 AM

Response to Request for Information: Public Access to Digital Data Resulting from Federal Funded Scientific Research

Dr. Karen Cole, Director, and library staff
January 12, 2011
kcole@kumc.edu
Archie Dykes Library of the Health Sciences, University of Kansas Medical Center
Kansas City, Kansas

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Federal policy should require that data and calculations resulting from federally funded research be deposited in open, curated digital archives and released to the public. We suggest that policies follow the model of the Bermuda Principles.

In particular:

- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute and reorganize the data.

Rights of use for data should be clearly stated using common, successful mechanisms such as The Open Data Commons licenses: http://opendatacommons.org/licenses/ and The Creative Commons Zero Waiver: http://creativecommons.org/publicdomain/zero/1.0/

Moreover, the data and calculations should be accompanied by provenance and descriptive metadata. The metadata should also reference resultant publications.

Publicly funded data restricted behind a publisher's paywall should not be an option. This type of restrictive practice prevents scientific discovery and progress.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

It is important to note that data sharing and archiving is already carried out in many different ways without intellectual property or patent conflicts. Nevertheless, there is general agreement

by publishers, researchers, and legal jurisdictions that data cannot and should not be copyrighted. Federal law should recommend or require that all data, calculations, and analysis of data be waived of copyright or licensed to allow reuse and modification. Creative Commons Zero, Science Commons and Open Data Commons License provide examples of such approaches.

Patent rights, including IP on materials/reagants, and privacy rights are different issues, and need not be waived along with a waiver of copyright (Dryad, http://wiki.datadryad.org/wiki/Terms_of_Reuse).

Publishers, e.g. journal publishers, should be allowed the choice of offering authors the ability to set embargoes on the release of data in accordance with embargoes on manuscript publication. Data repositories should establish mechanisms and workflows to support this.

Publishers should be allowed to expect that data repositories offer secure access to the data and calculations for editors and reviewers during the manuscript review process. Publishers and authors should be allowed to expect that data repositories suppress information about related manuscripts until the article has been published.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Embargo periods may differ depending on the discipline and the rate of discovery within the discipline. Less "volatile" disciplines should be allowed longer embargo periods if necessary. Embargo periods may be necessary to protect intellectual property, pending patents, the researcher or institution's interests, the funder's interests, or the public good. These exceptions will likely be more prevalent in some disciplines than in others.

Support for submission integration between journals and data repositories should be flexible enough to accommodate publishing, editing, and review workflows.

Metadata profiles must be flexible enough for different disciplines while still being interoperable. Dryad's implementation of Dublin Core Metadata Initiative Abstract Model (http://dublincore.org/documents/abstract-model/) is one example.

File formats for data and calculations will be different among disciplines. Federal agencies should require open, non-proprietary file formats whenever possible since they are more likely to be readable in the future. For example, a plain text file has a longer life than a propriety word processing format, and a file of comma or tab-delimited values has a longer life than a proprietary spreadsheet file format. ASCII text should be preferred over color, images, and other embedded objects which are difficult to migrate.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Each agency, each research project funded by an agency, working groups within the various disciplines, and the institutions responsible for long-term stewardship of data should work together to develop needs for retention, preservation, and long-term stewardship for their cases.

Agencies should make available funding and support for institutions providing long-term stewardship and dissemination.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Stakeholders can contribute to data management plan implementation by providing education, software, curation, metadata, storage, replication, and distribution.

Publishers, research communities, libraries, repositories, and institutions can contribute by working together to develop mutually beneficial workflows and submission integration between journals and data repositories.

We are a university library who has a close working relationship with our biomedical research community and with our university's information technology department. We support and contribute to software development in support of research (e.g. DSpace and BibApp). We create and manage author metadata, works metadata and content. We provide services to publicize research. We educate researchers about open access and copyright.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Funding mechanisms should provide for storage and dissemination of data. They should also provide resources and incentives for long-term preservation of data. Agencies must make available funding and support for institutions providing long-term stewardship and dissemination. This funding must occur at the agency level and not be parsed out in individual grants

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Agencies must provide researchers with unambiguous methods for linking funding agencies, data, and resulting works.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Adopt clear licensing requirements for data that are easy for researchers to comply with.

Provide tools that reduce the burden of licensing, ease compliance, reduce duplication, and open the data for use and re-use.

Fund educational initiatives that promote the use of data and computational thinking within primary, secondary, and higher learning institutions.

We also refer you to Semantic Web: Revolutionizing Knowledge Discovery In the Life Sciences (Baker and Cheung, 2007. ISBN-13 9780387484365) which provides insightful, modern, and practical analysis of types of innovation possible and necessary when the right data, tools, and standards are available.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Linked Data and Semantic Web standards such as OWL and RDF.

Established discipline-specific or community-developed data standards. MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

We also refer you to Semantic Web: Revolutionizing Knowledge Discovery In the Life Sciences (Baker and Cheung, 2007. ISBN-13 9780387484365) which provides insightful, modern, and practical analysis of types of innovation possible and necessary when the right data, tools, and standards are available.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Adopt simple minimal standards to allow for inter-operability and search that have buy-in and participation from a wide range of stakeholders. Optional discipline based standards could then evolve within those scholar communities.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Unique and persistent identifiers for publications and associated data are critical for linking the two. Identifiers should conform to modern namespaced URN/URI schemes. Examples of current successful schemes include DOIs, Handles, and PURLs. These schemes and standards are currently widely by publishers, content repositories, and digital libraries. The schemes have proven their utility in modern semantic web and linked data applications.

Access to identifiers and registries should be publicly available so that identifiers may easily be dereferenced and resolved to content endpoints via linking standards such as OpenURL.

Publications and data should be as atomistic as is economically feasible so that clear references and inferences can be made by human as well as by machine. For example, an OWL Reasoner and a human alike should be able to traverse the path from an assertion made within a publication to some element within a dataset or step within a calculation. In general, policies and practice should follow and build upon modern Linked Data practice.

In addition, please identify any other items the Working Group might consider for Federal policies related to public access to peer-reviewed scholarly publications resulting from federally supported research. Please attach any documents that support your comments to the questions.

Karen Cole
Director of Dykes Library
University of Kansas Medical Center

From: Baker, Gavin
Sent: Thursday, January 12, 2012 6:44 PM
To: publicaccess@ostp.gov; digitaldata@ostp.gov
Subject: Openness maximizes the benefits of science

I am writing in response to the Office of Science and Technology Policy's recent requests for information on public access to peer-reviewed scholarly publications and digital data resulting from federally-funded research (76 FR 68517 and 76 FR 68518) on behalf of the National Science and Technology Council.

I encourage the Council to issue strong recommendations to maximize public access to data and publications resulting from federally-funded research. Openness maximizes the benefits of research by increasing its scientific and economic impact while upholding scientific integrity.

To ensure that taxpayers derive the most benefit from the research they support, the Council should recommend that federal agencies require grantees to make their publications freely accessible to the public at no cost no later than six months after publication.

In addition, the Council should recommend that federal agencies require grantees to submit data management plans describing how they will manage, share, and provide public access to their data, if at all. The Council should also recommend that agencies establish expectations that grantees will provide public access to their data to the greatest extent possible, with narrow and specific exemptions (such as to protect human subjects and national security).

By issuing these recommendations and encouraging agencies to promptly implement them, the Council will fulfill its responsibility to advance federal science and ensure the best use of taxpayer dollars.

Sincerely,

Gavin R. Baker
Graduate student
School of Library and Information Studies Florida State University

Ecological Society of America
1990 M St, NW, Suite 700
Washington, DC 20036

January 12, 2012

Re: FR Doc. 2011-32947

Dear Madam or Sir:

The Ecological Society of America (ESA), the professional society of 10,000 ecological scientists, appreciates the opportunity to provide input to the Office of Science and Technology Policy's Request for Information on public access to digital data resulting from federally funded scientific research. As a publisher of peer-reviewed ecological journals for over 90 years, ESA policies and capabilities support data sharing and archiving and we welcome the development of agency policies and standards that will preserve and enhance access to digital data resulting from federal support.

ESA provides means for data publication and citation through *Ecological Archives*, which publishes data papers, supplements and digital appendices for our journals. This archive allows authors to make available supporting materials such as methodological details, data tables, photographs and supplemental discussion. ESA requires data archiving for papers published in our journal *Ecological Monographs* while data archiving is currently encouraged but voluntary for our other journals. The Society's interest in the development of data sharing policies is reflected in a series of workshops ESA organized. Sponsored by the National Science Foundation, the five workshops explored common data sharing policies among scientific societies, the needs for data registries and repositories and obstacles and incentives to share data.

The points below address our key ideas in regard to the development of federal agency policies and standards for public access to digital data:

• As noted by the National Science Board's recent report on digital data policies, "a single data sharing and management policy will not apply to all research communities." Stakeholders should have input in developing policies and standards within the various research communities.

• Bearing in mind the various needs of different research communities, discipline-specific, consistent standards should be developed among agencies to minimize the burden on researchers who receive funding from multiple federal sources.

• Separate funding for data archiving and curation should be included in grants and agreements to supplement rather than compete with existing research funds.

• Federal agencies should examine existing standards already in use in the private sector to determine if any may be applicable.

• Funders should consider permitting short embargo periods for required sharing of data linked to publications, in order to allow researchers to complete multiple analyses and publications based on a single dataset.

• To ensure appropriate attribution and credit, data should be published in such a way that they are traceable to authors, to related publications and to funders. For example, digital object identifiers could be used for authors and funding sources could be included in metadata.

Thank you for the opportunity to contribute to the deliberations of the National Science and Technology Council's Interagency Working Group on Digital Data.

Sincerely,

Katherine S. McCarter
Executive Director and Publisher