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