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.