March 2012

Response to the National Institutes of Health (NIH) RFI on Input into Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics.

Denise Troll Covey  
*Carnegie Mellon University, troll@andrew.cmu.edu*

Gloriana St. Clair  
*Carnegie Mellon University, gstclair@andrew.cmu.edu*

Follow this and additional works at: [http://repository.cmu.edu/lib_science](http://repository.cmu.edu/lib_science)  
Part of the [Public Policy Commons](http://repository.cmu.edu/lib_science)

This Article is brought to you for free and open access by the University Libraries at Research Showcase @ CMU. It has been accepted for inclusion in University Libraries by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
NATIONAL INSTITUTES OF HEALTH (NIH) – Request for Information: Input into Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics

On behalf of Carnegie Mellon University and the roughly 4,000 faculty and staff we represent, we write to thank you for issuing this Request for Information (RFI) and to share our perspective on policies regarding the management, integration, and analysis of biomedical datasets. 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

Carnegie Mellon’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.² Our community creates large volumes of federally funded research data. Though we do not have a medical school, we have a Biomedical Engineering Department, a Bone Tissue Engineering Center, and a Center for Bioimage Informatics. We also participate in the National Resource for Biomedical Supercomputing (NRBSC), with funding from the NIH. We strongly support broad sharing of biomedical and other datasets gathered in federally funded research projects because open data increase productivity, innovation, and commercialization.³ Developing open data policies and standards that facilitate sharing is in the national interest and warrants careful examination. We address here several of the areas identified in the NIH RFI, specifically standards development, secondary/future use of data, incentives for data sharing, and support needs.

¹ See http://www.cmu.edu/about/index.shtml.
³ 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.
Standards Development

The critical issue is the lack of standards and best practices in many areas relevant to sharing data. The lack of standards makes it difficult if not impossible to share data effectively and to preserve it over time. Data dissemination and preservation are crucial to reducing redundancy, verifying results, and increasing the integrity and productivity of science. The recent report by the Committee for Economic Development provides empirical evidence of the impact of increased openness on follow-on research, innovation, commercialization, and economic growth. This section provides an overview of areas in need of standards and best practices and suggestions for how the NIH could advance developments.

Data types and formats. Disciplinary differences in data types and formats must be addressed through standards and best practices. Even within the NIH, disciplinary differences exist in data formatting, documentation, and curation. The agency should facilitate the development and dissemination of standards and best practices for sharing and preservation of digital data by:

- Maintaining open access copies of relevant standards and best practices for data management (including metadata).
- Requiring NIH-funded research communities that do not yet have relevant standards and best practices to develop them within a specified time frame. NIH can identify disciplines that are poorly prepared to comply with data sharing policies and encourage them to collaborate with experienced and trusted partners, e.g., university libraries and research communities well-versed in data sharing.
- Participating in and funding standards development activities. (See Support Needs below.)

Data citation. Standards for data management and re-use should include standards for acknowledgement of data used in publications and citation of data products. Ideally, the descriptive metadata bundled with the dataset will convey the licensing terms and include a list of those to be attributed. (See Incentives for Data Sharing below.) However, the attribution of credit for datasets is a relatively new field of endeavor. Many groups are working to develop best practices for data citation in the sciences and humanities. Strict guidelines for data citation cannot yet be provided, but federal agencies requiring data management and sharing can provide ongoing guidance for data citation, keeping close watch on new developments in the field and keeping researchers informed. In addition, federal agencies requiring data sharing

---

5 If providing open access copies is not feasible, NIH should at minimum provide a list of relevant standards and best practices with links to where researchers can get the documents.
6 Two current development activities relevant to data citation should be monitored closely. The National Information Standards Organization (NISO) is developing a recommended practice for use of the International Standard Name Identifier (ISNI) to identify institutions. (See http://www.niso.org/publications/isq/2011/v23no3/gatenby.) The ORCID (Open Researcher and Contributor ID) project is developing unique identifiers for individual researchers to resolve name ambiguity problems in scholarly communication. (See http://www.orcid.org.)
should fund research into best practices and systems for data citation to accelerate the
development of guidelines for researchers in different disciplines.

**Trusted data repositories.** Standards and best practices must also be created for trusted
repositories committed to open data and its preservation. The federal government should
establish minimal service criteria to be met by such repositories, for example:

- Support for appropriate open data licenses. (See Incentives for Data Sharing below.)
  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.\(^7\)
- Searchable metadata that includes the licensing terms and, if attribution is required, a list
  of those requiring attribution.\(^8\) (See Incentives for Data Sharing below.)
- 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.\(^9\)

Trusted repositories will have a commitment to long-term maintenance of digital datasets
documented in a service-level agreement, and the financial resources and knowhow to sustain
the operation.\(^10\) To facilitate the development of trusted repositories for open data, federal
agencies with data management requirements should work with university libraries, disciplinary
societies, research consortia, and other stakeholders to distribute the many responsibilities
associated with establishing and maintaining a trusted repository for digital data.\(^11\)

**Secondary / Future Use of Data**

The primary stakeholders in secondary use of federally funded biomedical datasets are (a)
those who fund the research,\(^12\) (b) those who conduct the research, and (c) the human research

---


\(^8\) For a brief explanation see Rubin, Richard E. *Foundations of Library and Information Science*. New


\(^12\) This includes the federal agencies and taxpayers who underwrite the research, and the institutions that
manage the grants, provide laboratory space, and pay researcher salaries.
subjects who provide the data by consenting to be studied. The critical issues are protecting the rights of these three groups. From the funders’ perspective, the critical issues are maximizing return on investment and accountability, interests well served by openness and secondary use of data. From the researchers’ perspective, the critical issues revolve around acknowledging and leveraging their rights to the data. From the research subjects’ perspective, the critical issues revolve around privacy and informed consent. This document addresses researcher concerns in the section on Incentives for Data Sharing. This section addresses human subject concerns.

Institutional Review Boards (IRB) are responsible for protecting human subjects participating in research studies. To Carnegie Mellon’s IRB, the critical issue related to secondary / future use of federally funded biomedical datasets is the protection of the human subjects from which the data were originally collected. Our IRB’s position is documented in the Council on Governmental Relations’ (COGR)\textsuperscript{13} response to the U.S. Department of Health and Human Services’ (HHS) Advance Notice of Proposed Rulemaking (ANPRM), \textit{Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators}.\textsuperscript{14}

The COGR supports a general future use provision in consent forms, but does not consider it a prerequisite for any and all future research use. The COGR does not support the informed consent requirement proposed by HHS for unanticipated future use and analysis of data or biospecimens collected for research. The current regulatory framework for future use of pre-existing data and biospecimens provides human subjects with sufficient protection. Questions of privacy and confidentiality and assessments of risk should remain the responsibility of the IRB and be addressed on a case-by-case basis. Innovative secondary uses – for purposes different from what drove the original data collection – should not be discouraged.

The COGR calls for an IRB to review repositories established for research purposes – specifically their policies and procedures for obtaining and disseminating data – to determine whether informed consent is required based on established criteria. When secondary uses are planned, the IRB should review the proposed use to determine whether informed consent is required. If the proposed use meets the established criteria for waiving informed consent, informed consent will not be required. If data in repositories have no identifiers, informed consent for future research use of these data should not be necessary because the proposed research is not human subjects research.\textsuperscript{15}

\textsuperscript{13} The COGR is an association of 188 research universities and their affiliated medical centers and research institutes that works to ensure that federal agencies understand how the academy operates and how proposed regulations affect the academy. Carnegie Mellon is a member of the COGR.
\textsuperscript{14} The ANPRM is available at http://www.regulations.gov/#!documentDetail;D=HHS-OPHS-2011-0005-1057. The COGR response to the ANPRM is available at http://www.cogr.edu/.
\textsuperscript{15} COGR response to HHS-OPHS-2011-005, p. 21. The COGR opposes expanding the meaning of “human subjects” to include biospecimens without identifiers because it reverses long-standing definitions and creates a significant burden for investigators, delaying if not undermining the conduct of research, without increasing protection or reducing risk to tissue donors.
Incentives for Data Sharing

Researchers in different disciplines have different levels of understanding of the benefits of openness and different pragmatic needs. The NIH needs to understand its research communities, take steps to remove unnecessary barriers, and make appropriate concessions that facilitate both science and openness. The critical issue is negotiating a compromise that effectively meets researcher needs for attribution and advancement without unnecessarily diminishing the return on investment in federally funded research afforded by sharing data.

Researchers must be willing to share their data. Many are not, either because they do not understand the benefits of sharing data, because sharing is burdensome, or because they – and their institutions – have a vested interest in not sharing data. Like other universities, Carnegie Mellon 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. While many datasets are not protected by copyright and are not, in the legal sense, “owned” by the researchers, the researchers’ de facto rights to the data cannot be denied. 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 careers, develop patents, and contribute scholarship is a top priority for researchers across the nation.

Education can address researcher lack of understanding of the benefits of sharing data. Standards, best practices, and infrastructure support can reduce the burden of sharing data. (See Support Needs below.) Protecting researcher interests can be addressed with data citation standards and appropriate licenses and embargoes on public access. Compliance can be facilitated by reducing the burden of data sharing, funding the activity, and holding researchers accountable.

Research shows that among both academic- and industry-based scientists, as the competitive value of the requested information increases, the likelihood of sharing the information decreases. 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 NIH should acknowledge that different disciplines operate under different constraints and work with its research communities to specify the conditions and timeframe within which data must be made open. Depending on the discipline, this may be before or after peer-reviewed publication of research findings.

---

16 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).

Licenses. Ideally, licenses for open data should be human- and machine-readable. Appropriate licenses for open data include:

- 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.

Additional licenses might need to be developed. An appropriate license will preserve rights and provide incentives for researchers to make their data publicly accessible.

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 of the scientists and federal agency.

Embargoes. While the ideal is prompt public access to data, in some disciplines the goals of growing the economy and increasing the productivity of science might be achieved more effectively by granting researchers control of the data for some finite time, after which the data become open and competitors can use them 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 use – something akin to an embargo on public access to scholarly publications. The point at which the dataset will become open should be specified in the administrative metadata.

Compliance. Reducing the burden of data management and sharing will facilitate compliance with open data policies. Federal agencies with an open data requirement should take steps to clarify, streamline, and support the process, for example, by:

---

18 For details, see Open Data Commons Licenses FAQ, available at [http://opendatacommons.org/faq/licenses/](http://opendatacommons.org/faq/licenses/).

19 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](http://blogs.talis.com/n2/cc).


21 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.
• Providing a common definition of “data” across federal agencies to help clarify rights and responsibilities regarding ownership, access to and retention of data.22
• Working with research communities to develop and disseminate standards, best practices, and appropriate licenses and embargo periods.
• Providing guidance and resources to help research communities address constraints and contractual obligations (e.g., privacy, confidentiality) regarding managing and sharing data.
• Maintaining a list of trusted repositories for various types of data.23

Perhaps the most effective strategies for facilitating compliance will be making data management and sharing quid pro quo for future funding, and allocating grant funds that can only be spent on data management. (See Support Needs below.)

Support Needs

The critical issues are developing the infrastructure to support data sharing and reducing the burden of data sharing on researchers and institutions. If the infrastructure is inadequate or the burden too great, the data management requirement will not be met and the societal and economic benefits of data sharing will not be reaped.

There is much work to be done to address issues surrounding data management and preservation. Federal agencies requiring data sharing should fund research and development of needed standards and tools for data curation and citation. In so far as they represent the interests of their constituent communities, federal agencies 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.24 DataCite supports data archiving that permits verification and repurposing of the data and works to establish easier access to data.25

Federal agencies with data management requirements should also provide funding exclusively for data management and preservation, both as part of existing grant programs and as a

22 Response of the Council on Governmental Relations (COGR) to the Office of Science and Technology Policy’s FRI on public access to digital data resulting from federally funded research. Available at: http://www.cogr.edu/.
23 This list could be generated from registry records such as DataCite. See http://www.datacite.org/repolist.
24 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/.
separate set of programs for retroactive preservation projects. The requirements for the disposition of these funds should allow for support of the data management infrastructure beyond the grant period, 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.

Given that important collaborative work and data sharing can happen with datasets of all sizes, researchers must be required to include meaningful data management plans in their grant proposals for awards of all sizes – not just large awards. Data management costs must be included in detailed budgets and budget justifications. These costs must not be used explicitly or implicitly to penalize grant proposals based on the amount of money required for data management.

In closing, we encourage the NIH to consider carefully the new burden that data management places on researchers and institutions and the complex context in which this burden is faced – limited human and financial resources, inadequate infrastructure and understanding, disciplinary differences, and competing rights. To the extent that the benefits of open data exceed the costs (both personal and financial), openness should be pursued. The pace of requirements should be commensurate with the pace of developments needed to support compliance.

Thank you for the opportunity to provide comments on this important initiative.

Sincerely,

Gloriana St. Clair
Dean of University Libraries
gstclair@andrew.cmu.edu

Denise Troll Covey
Principal Librarian for Special Projects
troll@andrew.cmu.edu