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