[Assigned ID #]

[Assigned Entry date]

Name/Email

David W. Robinson, Ph.D.
Executive Vice Provost

Affiliation/Organization

Oregon Health & Science University

City, State

Portland, OR

Comment 1:

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

Free, timely, and re-usable access to the digital data resulting from federally funded scientific research will improve science and human health by promoting transparency, efficiency, and reproducibility. To realize this value, federal policy must speak to several key issues. Policy and practice should create a technical infrastructure that addresses discovery, usability, attribution, and long-term preservation. These specifications must include the archiving of data in publically accessible repositories, using standard record formats, and promoting best practices for interoperability and reuse, such as Semantic Web standards and Linked Open Data. Funding agencies should support awardee data management and compliancy efforts by integrating these expenditures into grant structures and through interagency standards that offer transparent and practical workflows. Finally, award and incentive systems must evolve to recognize the value of data management and sharing to the scientific enterprise.

Markets will emerge to support the management and usability of the data; and, the economy will benefit from derivative products and services. The U.S. can look to the Human Genome Project (HGP) as a strong proof of concept. The HPG has led to groundbreaking discoveries and therapies. For example, our pioneering faculty member Dr. Brian Druker's development of the cancer drug Gleevec is intrinsically linked to the research sharing and advances the HGP fostered. Initially, nearly four billion dollars was invested in the HGP. Since its inception, an entire industry has developed to support genomic research and R&D. The ROI is dramatic; in 2010, the industry produced $67 billion in U.S. economic output, $20 billion in personal income for U.S. citizens, and 310 thousand jobs.[1]

Increasingly, governments, funding agencies, and research institutions are recognizing the scientific, societal, and economic benefits of open data. Since 1999, the NIH has required that crystallography data be submitted to the Protein Data Bank (PDB) upon journal manuscript publication. On December 12, 2011, the European Commission launched an open data strategy

for Europe. Announcing the policy, the Vice President of the European Commission stated that openness is the best strategy for gleaning value from data.[2]

Comment 2:

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

Recognition of data sets as a citable, authored sets of information will help protect the interests of individual stakeholders. This provides a framework for recognition that is based on the currently accepted model of citation within publications. Additionally, when appropriate, a phased approach to access can be taken to protect the IP interests of stakeholders. Time-limited embargo periods could be utilized to manage both access and re-use rights. For example, rights holders may choose to restrict commercial re-use for a specific time period. However, embargo periods should not impede depositing of data in a repository.

Comment 3:

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

In aiming to build policy that acknowledges differences among scientific disciplines, it will be important to leverage the knowledge of scientific communities and experts in the organization of knowledge, such as libraries and information science researchers who are accustomed to providing guidance and resources for disparate kinds of data. Existing discipline/data specific repositories should also be consulted to ensure applicability.

While data management needs differ by discipline, there are qualities and practices that underpin all data types and should inform inter-disciplinary requirements. For instance, there exist upper ontologies that represent the types of things that exist. Classification of data elements can be tied to such upper ontologies via reuse of these upper ontologies. One example is the Basic Formal Ontology as the upper level ontology for all Open Biomedical Ontologies (OBO), which enables representation of things as diverse as a mammary gland, a PCR machine, and mitosis. Similarly, while each discipline's data may require specialized formats, queries and applications, if the federal agencies promote open and extensible standards, the different needs will be met.

Additionally, it must be recognized that data is utilized in ways disconnected from the creator's original research focus. Data from disparate disciplines are combined and analyzed to advance scientific inquiry and support markets. Interoperability standards will benefit these new applications.

Comment 4:

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

There are data management requirements common to all disciplines and data types; and, there are discipline- and data-type-specific needs and expectations. Agencies should build cost and responsibility frameworks that account for both the shared and unique needs. Mutual requirements can be met via interagency collaboration, standardization, and cost sharing. Unique requirements can be met via discipline and agency specific support and innovation. Such a framework has the potential to control costs and maximize benefits by limiting duplicate efforts, distributing responsibility—for both shared and unique needs—and, encouraging public-private partnerships.

EUDAT, a European based data infrastructure initiative, is currently pursing this strategy, and its work should influence U.S. agencies. EUDAT states, "Although research communities from different disciplines have different ambitions, particularly with respect to data organization and content, they also share basic service requirements. This commonality makes it possible to establish generic…services designed to support multiple communities, as part of a Collaborative Data Infrastructure."[3] The figure below illustrates this framework of collaboration and distribution.

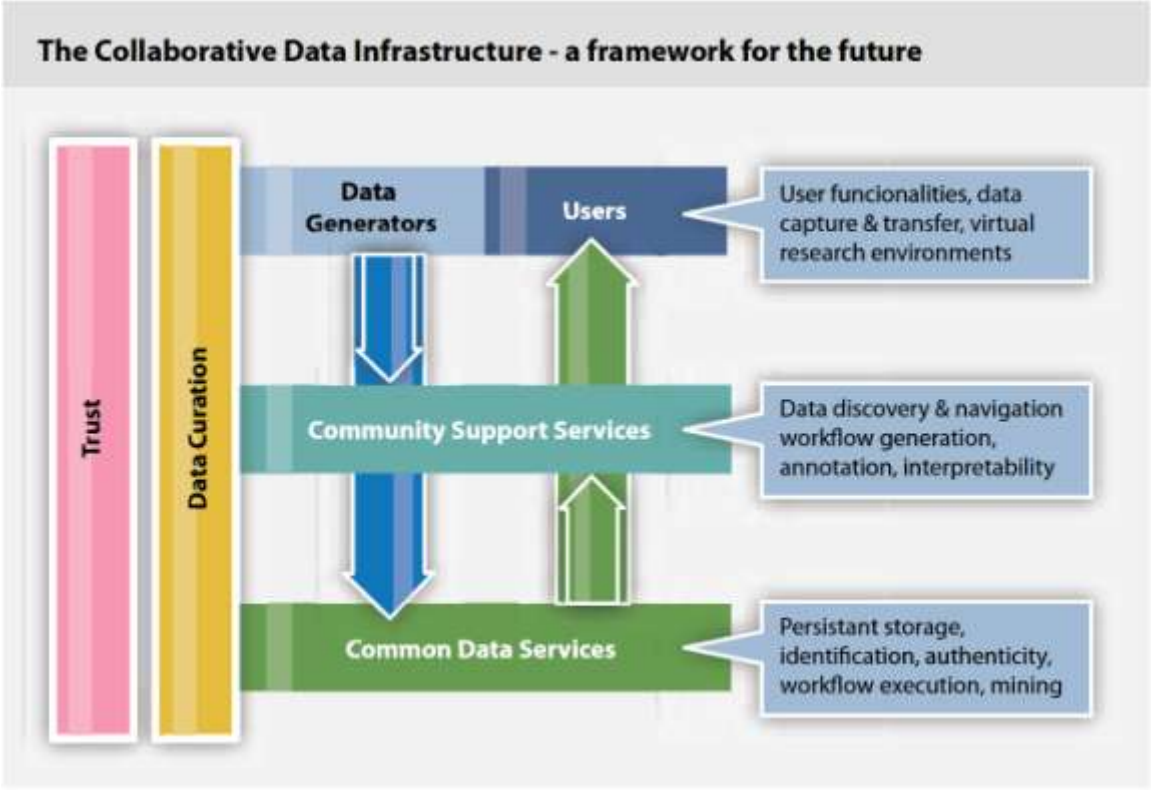

Image courtesy of International Science Grid this Week: http://www.isgtw.org/

Comment 5:

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

There is room and opportunity for many contributors; however, it is critical that participation is regulated by technical and legal standards (e.g. use rights) that ensure and promote free public access, discovery, re-use, and preservation. The expertise, technologies, and infrastructures of stakeholders should be leveraged both in the development of policy frameworks and their execution. Such collaboration will drive best practices, innovation, market creation, and compliance. The present repositories of research communities, publishers, and institutions can be utilized and developed (e.g. Pangea, TreeBase). Existing partnerships between publishers and repositories, such as Dryad, can be grown. Organizations like DataCite work to improve the discoverability and utility of data. Universities, research institutions, and libraries have been and should continue to be key contributors, building infrastructures to support their researchers' compliancy, as with NIH public access policy, and guiding archival and discovery standards. Libraries are also well positioned to enable these infrastructures to be compliant with the Semantic Web and population of Linked Open Data from these data sources.

Comment 6:

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

Maximum scientific and economic rewards will not be realized if the cost of data management and preservation are an after-thought.[4] Researchers and institutions should be required to document the real cost of data management and publication within their proposals and reports. The cost and effort associated with doing so should be accommodated for in the total budget. Funding mechanisms, requests for proposals, and agency budgets must also address the real costs of long-term preservation, the latter being independent of grant-specific costs. In this regard, leveraging and supporting the services and expertise of institutions and organizations with memory driven missions (e.g. libraries) should be considered. Agencies should consider funding libraries to perform more research "in the field" on making specific data types conform to standards and archived for maximum searchability.

Comment 7:

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

Interagency standards that offer transparent and practical workflows and mandate deposit in publically accessible repositories will improve compliance, facilitate verification, and reduce burden. Ideally, such a systematic approach would require that:

- All data are deposited in publically accessible databases in conjunction with manuscript acceptance.
- Standardizations of record formats and minimum metadata are applied and verified.
- Several submission workflows are supported, including third-party deposit.

- Data and manuscripts are assigned persistent, linked identifiers.
- Compliancy is demonstrated through key events in the research enterprise via persistent, citable data identifiers.

In contrast to data management and access policies organized around the individual researcher or lab, a systemized approach reduces burden by enabling home institutions, libraries, and scientific communities to build effective support services.

Comment 8:

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

In cases where technical and legal openness are supported, examples of innovation and market growth abound. Open access to US National Weather Service data underpins a large and diverse industry, estimated by the American Metrologic Society to exceed $1.5 billion per year.[5] "Between 1988 and 2010 the human genome sequencing projects, associated research and industry activity—directly and indirectly—generated an economic (output) impact of $796 billion, personal income exceeding $244 billion, and 3.8 million job-years of employment."[1]

To stimulate innovative use and grow the economy, technical infrastructure must support the sophisticated needs of human and machine readers; and, legal infrastructure must support liberal re-use rights, including non-exclusive commercial development. For example, data should be archived according to standards that support multiple formats, using standards metadata and complying with current best practices of data sharing and integration over the Web (e.g. Semantic Web standards and Linked Open data). Licensing frameworks, such as the Create Commons CC-BY, offer a starting foundation for building a license for data that will stimulate investment in new capabilities and applications.

Finally, agencies and institutions should promote their data wealth and encourage its use. The World Bank opened its data in April 2010; in October 2011, it launched the Apps for Development contest, challenging the developer community to create tools and applications using World Bank data. The contest rules ensured that contestants would retain the intellectual property rights of their software. Developers from 36 countries responded, submitting software, mobile apps, games, and widgets aimed at policy makers, educators, health care providers, and the public. Agency promotion of open scientific data would be sure to spur similar participation and innovation, and small funding opportunities or similar contests could be created to do so.

Comment 9:

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

As with manuscript publication, secondary results should cite primary data. Standardized data identifiers will provide the means to identify and link the resource to other relevant documents,

data, persons, etc. The use of controlled author and institutional identifiers (e.g. ORCID registry) will be critical to support disambiguated and resolvable attribution. Furthermore, use of a common metadata standard to tag various kinds of data with appropriate attribution in a standardized way will ensure proper attribution. It is not always enough to know whom the data came from, but also the version, from where, and how is it related to other documents, data, experiments and grants.

Comment 10:

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

It would be a burdensome and unproductive to identify, set, and enforce discipline-specific content standards. Rather, what should be pursued are standards that optimize discovery and use by human and machine readers. This includes format standards, minimum metadata requirements, Semantic Web standards, and Linked Open Data. A minimum metadata standard for any kind of content should be created - whether it is a data set, a publication, a patent, a grant, and ontology, a blog, etc. Anything that is reportable as linked to grant funding activity should meet this minimum metadata standard.

Comment 11:

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

There are examples of successful standards development in various domains. The W3C standards development process has successfully produced HTML, XML, RDF and other languages. Key to the process is its openness and community participation. Successful standards development relies on the contributions of a diverse population of experts, including scientists, information professionals, and technologists.

Comment 12:

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

All federal agencies should pursue a technical infrastructure for data that utilizes international standards for interoperability and re-use, such as Semantic Web Standards and Linked Data. Agencies can and should leverage the work of organizations focused on international data sharing and utility, such as CODATA, the Global Biodiversity Information Facility, the Open Archives Initiative, and the Digital Curation Center. It would also be worthwhile for federal agencies to participate in and support international efforts to connect data collections and build

collaborative data infrastructures that aim to deliver cross-disciplinary data services. Similarly, adoption of other international efforts to standardize metadata, for example, coordination between VIVO and CERIF, the European organization for international research information, will facilitate data integration internationally. Promoting such coordination as part of existing granting mechanisms or via new ones to promote international collaboration will be beneficial.

Comment 13:

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

In order to support manageable and meaningful linking, standards and practices must speak to the required use of persistent, unique identifiers for data, publications, authors, and institutions. Unique identifiers will strengthen the visibility of each item and the links between items, as well as enable re-use and the development of new services. This will require a new age of semantic awareness on part of the researcher, the reviewers and the publishers of manuscripts and data. Publishers and granting agencies need to improve their standards regarding unique identification of research entities, and guidelines to authors need to be generated in support of these new standards. Furthermore, enhancing current research training to include these modern information management strategies will be key.

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

1. Battelle Technology Partnership Practice. (2011). Economic Impact of the Human Genome Project. Retrieved December 13, 2011, from http://www.battelle.org/spotlight/5-11-11_genome.aspx
2. European Commission launches Open Data Strategy for Europe | Open Knowledge Foundation Blog. (n.d.). Retrieved December 27, 2011, from http://blog.okfn.org/2011/12/12/european-commission-launches-open-data-strategy-for-europe/
3. Approach | EUDAT. (n.d.). Retrieved December 27, 2011, from http://www.eudat.eu/approach
4. Oecd Follow Up Group. (2003). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal, 3 (29)*
5. Annex 1 – Best Practice and Emerging Evidence - Economic Growth | data.gov.uk. (n.d.). Retrieved December 27, 2011, from http://data.gov.uk/opendataconsultation/annex-1/economic-growth#_ftn9