Adrian Pohl
Thu 1/12/2012 5:07 PM
Response to the RFI on Digital Data

Dear people at the OSTP,

below are my answers to your questions on Digital Data. Again, I am responding as an individual working in an institution which provides information (research tools as well as licensed content) to academic libraries. Also, I am coordinating the Open Knowledge Foundation's "Working Group on Open Bibliographic Data" but cannot and do not speak for this group.

All the best
Adrian

(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?

Response:
In summary our response advocates:
Immediate release
Disclosure of broad estimation of acquisition cost Proper open licensing Adoption of open standards for data files Adoption of extensible standards for metadata

Immediate Release
Federal agencies funding scientific research must establish policies by which the data acquired in federally funded scientific research
(FFSR) must be made immediately and fully available in public data repositories while ensuring subjects privacy.

The policies should follow the model of the Bermuda Principles
(<http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml>).
In particular on:
- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute and reorganize the data.

Disclosure of Acquisition Cost
When reviewing proposals for funding opportunities, federal agencies should require that the sections requesting public funds for data acquisition activities provide a clear estimation of the cost of acquiring the data. If funded, researchers should be required to make data available in public repositories immediately after acquisition, and in the metadata used to describe a dataset, researchers should also be required to include the cost of acquisition.

The goal will be to develop a sense of the economic cost of not releasing data. For example, not releasing a dataset that cost $1M to be acquired is a loss for the federal government of the $1M funds provided by taxpayers. This is the direct value lost from the overall economy; the actual value lost is much larger since it should include the missed opportunities that could have resulted from the exploitation of the data.

The European Commission, for example, recently adopted a policy of open data dissemination (<http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=EN&guiLanguage=en>).
The principle, rooted in the arguments that Yochai Benkler makes in his book “The Wealth of Networks” is that data is more valuable when shared; in economic terms, data is an “anti-rival good”. It is a good that becomes more valuable, when more people have access to it and use it.

Proper Open Licensing
Current copyright legislation has been strongly focused on protecting the creators of artistic works, and in the process have created an inhospitable environment for the daily sharing of scientific information. The litigious behavior that many institutions have developed around copyrighted materials, results also in a reaction of over cautious behaviors on the part of the potential users of data and documents resulting from scientific research activities.

To dispel this environment of uncertainty, it is fundamental to clarify the rights of the public to make use of data acquired as a result of FFSR. The most effective way of achieving this goal is to affix to every released dataset, a clear statement of licensing indicating what the recipients of the data are legally allowed to do with the data. Licensing issues are expanded on in the Panton Principles for Open Data in Science (<http://pantonprinciples.org/>).

Federal agencies should identify a set of licenses that ensure the rights of the general public to deal with the data, in particular to copy, distribute, and create derivative works, and in this way ensure that the data get to reach their maximum economic potential to foster the growth of the U.S. economy.

Adoption of Open Standards
Federal agencies must ensure that data are released in a usable form.
The first step in that direction is to require the adoption of open standards for file formats, and forbid the use of proprietary formats that could prevent the general public from having access to the data.

Standards file format used for digital storage of scientific data are abundant and vary greatly from one domain to the next. Therefore, the scientific community will have to be engaged with the federal agencies in identifying the proper open standard to be used on each discipline, and to create new standards in the cases where no suitable standard file format exists yet.

For standards to reach their full potential, it is fundamental to have an open source reference implementation of the standard, and to encourage the development of a ecosystem in which commercial applications implement the standard as well. In this way, it becomes possible to maximize the use of the data acquired as a result of FFSR.

(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?

In addition to the stakeholders listed in this question, it is critical to note that the general public is one of the primary (if not the primary) stakeholders to be considered here. Given that in the context of federally funded scientific research, it is the public’s tax dollars that are paying for the scientific research being undertaken, and thus the public's interest is the first one that should be considered when making trade-offs between available options.

Scientists who gathered data in federally funded scientific research did so as part of their job duties, and therefore under U.S. copyright laws they were performing "work for hire." This means that their employers are the copyright holders of any creative aspect of that data gathering (as pointed above, that only include the organization of data collections). Given that the scientists' employers received funds from the federal government, it should be expected that they will be subject to the same demands of the Federal Acquisition Regulations (FAR) as other contractors of the federal government. In particular with respect to the licensing of data acquired as part of federal contracts. The data should be published using an open license
(<http://opendefinition.org/licenses/#Data>) and at best follow the Panton Principles.

(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?

Working groups should be established for different disciplines, involving representatives of leading research institutions for each discipline.

Working groups should define differences with how the data are represented, indexed, stored and exchanged, but should not have the latitude to restrict in any way the free dissemination of information. All the policies should consistently have as a common factor the requirement for immediate and full release of data, unconstrained by any embargo periods or licensing restrictions. Credit for the acquisition of data could be ensured by data publications (eg
http://datacite.org) that can be cited by further works.

In this process, it is vital to invest in and commit to the emergence of standards that enable interoperability of, and thus reuse of, digital data. Linked Open Data standards for publishing (meta)data on the web build on central features of the Internet and the World Wide Web. As long as those data are in a tower of babel of formats, incoherent names, and might move about every day, they will be a slippery surface on which to build value and create jobs. Federal policy could call for a standard method for providing names and descriptions both for digital data and for the entities represented in digital data using URIs for identifying datasets and RDF and vocabularies like the DataCite metadata core for their description.

Standards also make it far easier to provide credit back to scientists who make data available, as well as increasing the odds that a user gets enough value from data to decide to give credit back. Embracing a standard identifier system for data posters will make it easier to link back unambiguously to a researcher as well as to make it easier for grant review committees and universities to receive a full picture of a scientist's impact, not just their publication list.

(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?

The working groups in the different disciplines (from Question 3) should establish guidelines on practices for dissemination and storage for different types of data. For example, in genomics, it may be reasonable to store the secondary sequence information but not the primary sequence (given their great difference in data size).
Analogously, the guidelines may require primary sequences to be stored only for 2 years, while the secondary sequences should be stored for
10 years.

In astronomy it may be required that certain types of images be stored for different periods of time. Some images may be required to be stored with different compression ratios, and therefore correlate their storage cost with the potential expected benefit for future studies. In this cost-benefit evaluation, the original cost of acquiring the data should be taken into account. For example, a project that invested $50M in acquiring data should not attempt to make savings of a few hundred dollars in storage.

Economists must be involved in the working groups charted with the mission of providing guidelines for storage and dissemination, given that this is a problem in which the trade-off for the benefit of society at large must be continually evaluated.

The policies of federal agencies should be affected by the constant advances in storage technology and the rapid decrease in the cost of storage. The federal government should stimulate the development of storage technology, either by creating large storage decentralized facilities, creating consortia to manage data storage services, involving the public in facilitating distributed (and redundant) storage systems based on peer-to-peer technology that has already proven to handle large amounts of data.

All these guidelines should be prepared following open and transparent procedures in order to prevent proprietary standards and vendor lock-in situations that would prevent the policies from maximizing the utility of federally funded scientific research to the general public.

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

They can join the working groups established in their respective disciplines of interest that will define practices for data management, including consortia combining universities, commercial companies and government agencies.

As standards and agreements are developed, working groups can help implement and test such plans in pilot projects. It will be of great help if federal agencies provide seed funding for these pilot projects.

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

Preserving and making digital data accessible is closely related to the issue of preserving and making scientific publications accessible.
If libraries and other non-profit organizations take over these tasks from the current commercial publishers as suggested in my answers to the RFI on scientific literature, there will be more than enough funds available from the current publisher profits to allow libraries to store and make digital data publicly accessible.

Once data and literature are stored in a database where both are linked semantically, innovators have a bounty of opportunities to provide commercial services and develop new applications and drugs/therapies to then generate a profit from.
In the current system, this information is restricted to a small set of academics, with innovators largely barred from access.

(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?

Once all data and literature are available to innovators, market forces should be allowed to take over without any additional policy interference, as the government is already funding the establishment of this resource.

(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?

The DataCite initiative has been working for some years on providing answers to this question.

(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?

I have also heard of Minimum Information for Biological and Biomedical Investigations (<http://www.mibbi.org/>) and Minimum information about a bioactive entity (MIABE) (<dx.doi.org/10.1038/nrd3503>).

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

A. Ensuring that Internationalization (of language and "locale") is made an integral part of the standards.

B. Starting with simple standards that can progressively be improved, instead of spending a lot of time in top-down design, committees and long-term procedural approaches to the definition of the standard. In other words, following the Agile methodologies that have proved to be successful in open source communities.

C. Working with existing international organization that have already defined standards in different disciplines, e.g. DataCite.

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

At best, the Linked Open Data approach should be used. Publications should get a HTTP-URI as identifier and also their different parts.
Datasets should be assigned a HTTP-URI, e.g. a DOI like used by the DataCite project. OAI-ORE is a well-known standard for representing a complex publication containing datasets etc. in RDF.