Thu 1/12/2012 7:04 PM
Response to RFI FR Doc No: 2011-28621, digital data

General comments:

If publications are the currency of science, then data is the collateral behind the currency's value. By mandating the sharing of this collateral, it changes the way science is transacted. Some scientists have embraced this and fantastic discoveries have emerged. Other scientists are not as enthusiastic. As Wilbanks (2012) puts it "The ugly reality is that sharing data represents a net economic loss in the eyes of many researchers: it takes time and effort to make the data useful to third parties (through annotation and metadata) and that is time that could be spent exploiting the data to make new discoveries….There is pervasive fear that other scientists will "scoop" them if their data are available before being fully explored" (para. 5). Before embarking on technological or financial resolutions, there should be recognition that sharing data may violate long-held beliefs. Only clearly articulating policies, incentives, and minimizing undue burdens on researchers and institutions can overcome this cultural barrier. In addition, "We do not have the sociotechnical infrastructure required to answer questions of data stewardship with any authority" (Wilbanks, 2012, para 8). However, because you ask, I will try to answer.

(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?

While the current data sharing mandates are laudable for their intent to increase exchange of information, which in turn should increase innovation and economic prosperity, they need further clarification. First, the goal should be scientific reproducibility and data re-use, not making data available for the sake of availability. Second, there should be continuing review of policy based on examples of effective data re-use. If policies are to be informed by evidence, then evidence must be collected and evaluated by economists, computer scientists, information managers and others who are qualified to determine the required innovations, costs and trade-offs required to meet the goal. Funding should be provided to study the effectiveness of current data sharing practices and the best use of resources for future data sharing.

(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?

Current intellectual property tools have a difficult time accommodating all expressions and uses of data (spreadsheets, computer code, database queries, etc.). While freely available data would be the ideal for creating new innovations and stimulating the economy, this is not always desirable. As well, even freely available data needs to be attributed correctly to protect the scientists' efforts. Placing the burden of responsibility on the scientist leads to confusion regarding issues such as the use of derivatives and designations of non-commercial versus non-profit. This confusion may result in unnecessarily conservative copyright and/or licensing. Alternatively, practices that are too liberal may lead to the loss of commercial potential for the institution or scientist, run contra to export control regulations, or endanger vulnerable populations. The federal government should encourage research institutions to craft intellectual property tools and educational programs for their researchers. This would enable scientists to apply appropriate copyright and licensing to their output. Each institution should create a clear and unambiguous policy on when and where data can be freely re-used, specific to the unique

potential of the data and in alignment with the spirit of data sharing. As well, institutions must provide legal consultation services to scientists as standards and mandates change.

(2) 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?

The current situation is that researchers do not necessarily come equipped with expertise in databases and curation activities and information management professionals may lack subject-specific knowledge about science data. Indeed, it is highly improbable for any one person to have expertise in all physical and life science data requirements, the infrastructure necessary to handle the data, and the curation activities that makes the data findable and re-usable. The only solutions are interdisciplinary.

To facilitate interdisciplinary collaboration, the federal government could sponsor interdisciplinary working groups. These working groups may not follow traditional lines of discipline separation (i.e. departmental hierarchies) and may be best identified by asking professional associations. Professional associations tend to have specialty or interdisciplinary subgroups that may represent discrete data practices. As groups of researchers are identified to have common data practices, they may then delineate how those practices meet, or fail to meet, the data sharing initiatives. Each working group should have access to expertise from the broadly associated disciplines. These broadly associated disciplines include metadata specialists, infrastructure experts, and legal counselors.

It may not be possible for certain groups to release data without significant detriment to commercial goals, research programs, or other contrary regulations. These groups may need waivers or accommodations when faced with data sharing mandates. One suggestion is that "Individual disciplines and communities can opt-out of funder-wide approaches if they make a strong public case that the principles and goals are not applicable to their area, or that they plan to achieve the same goals in a different but equally-effective way " (Piwowar, 2012, #3). Should disciplines not be prepared to either share data or defend why, then they need to elucidate current practices and explore future options. On the other hand, disciplines that have currently have 'dark repositories' or that desire specific data sharing services would be discovered. Specifically requesting that major professional associations report on their constituents' data sharing practices would identify discipline specific differences. Alternatively, simply making data management plans publicly available would allow information managers to get at discipline specific practices and to suggest alternatives.

Lastly, once discipline specific practices have been identified, they need to be unified. In other words, their standards, languages, and metadata schemas need to be interoperable. Without intervention, these standards "will not spontaneously emerge …as long as 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" (Wilbanks, 2012, para. 13). Funding agencies need to cite examples of verified emerging standards and stimulate new interoperability through challenges, prizes, and expanding grant opportunities. In particular, the ability for new repositories to federate with existing ones may drastically increase their survival.

(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?

Policies need to be informed by working groups that are primarily populated by the scientists. They know best what the relative costs and benefits are of long-term stewardship and dissemination of their particular data. Polling their professional societies and involving economic analysis would go a long way to answering this question in a discipline specific manner.

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

While data management plans have served to bring data management issues to the forefront, they reflect the high variability of available options and willingness to share. Since the plans themselves are so highly variable, the requirements for implementation are highly variable as well. To date, librarians have provided consultations, examples and tools for data management plan creation (DMPTool, etc.) and continue to explore options such as data repositories. The Association of Research Libraries has provided a structured course for research librarians to explore these topics and provide recommended actions to their institutions. However, no one entity will be able to answer all the challenges.

Research communities must contribute or effective solutions will not evolve. Universities must actively support their faculty with data sharing by providing legal consultation, infrastructure, and information management expertise. Scientific publishers need to provide avenues for data sharing and work with institutions to apply appropriate copyright and licensing. In particular, publishers must clearly state how they are handling data copyright and ensure that it is compatible with institutions' and scientists' needs. As well, scientific publishers should require and provide unique identifiers and citation of data sets. There are several endeavors currently underway that would facilitate this (datacite, doi's, etc.). There is work for everyone in this endeavor.

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

Funding agencies should promote digital data sharing by clarifying existing funding mechanisms and targeting new ones. Specifically, repositories need a funding mechanism for that period of time between start-up and when they have accumulated a critical mass of information, value-added services, and a strong user base. Across campuses, there are researchers who compile databases in an effort to organize and use their own output more efficiently. At times, their colleagues wish to contribute to this effort and a repository is born. Since funding agencies typically evaluate grant proposals based on novelty and the ability to generate new ideas, there is rarely funding for maintaining and improving existing repositories (Bastow & Leonelli, 2010). As researchers struggle to find funds and expertise, these 'dark repositories' languish in obscurity.

The logical entity to incubate a burgeoning repository is the institution's library. However, library budgets are not increasing and their service expectations are not decreasing. Therefore, an investment in a data repository must be carefully considered as their cost is "an order of

magnitude greater than that suggested for a typical institutional repository focused on e-publications" (Beagrie, 2008, para. 7). There is evidence that this increased cost is largely attributable to staff efforts in documentation, formatting, and ingest - not necessarily to the archival storage (Beagrie, 2008, Chapter 10). The expertise to properly document and ingest documents usually exists in the library, an entity that funding agencies typically consider an indirect cost, and therefore not eligible to charge fees directly to an individual grant award (OMB Circular A-21, F8). This classification and cost structure is a dilemma for libraries. If a data repository is to be available widely, it would be a major function of the institution, and should be covered under facilities and administrative (indirect) costs. If the data repository will only be used by a few disciplines then it should be charged to those projects that require and use the service (a direct cost to specific grant funds). In reality, any data repository will likely start with a few heavy users and then either generalize to accommodate a whole community or specialize to a particular discipline at a national or international level. How then, should the cost of such services be re-captured? Successful data subject repositories typically subsist on several sources of income, including private and public funds, and even subscriptions. The best solution is to provide a separate budget line for all activities surrounding data sharing (including proper documentation and ingestion) and to allow those funds to go to whatever entity, public or private, that provides the required services. This may also discourage the current practice of eliminating all funds for data dissemination when the proposal isn't fully funded. Otherwise, the letter of the mandate may be met, but the ultimate goal, re-use, will be hampered by inadequate documentation.

(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?

The simplest approach is to add an information field to the existing funding agency reports. In other words, require that data management plans be verified by a digital object identifier (doi) or uniform resource locator (URL) of where the data is shared. These doi's or URL's could even be published with final reports and summary publications. Safe places to deposit data will be preferentially used, creating new data repositories or increasing the use of existing ones. Proper repositories will specify proper citation techniques, and as data sets are cited, they can be tracked. This is analogous to tracking journal article citations. As we know, the Impact Factor from the Thompson Reuters Science Citation Index has been used to determine tenure, promotion, and publication preferences. Perhaps similar measurements of data re-use will evolve and be used as an incentive for data sharing. (This also applies to question (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 well, funding agencies need to develop guidelines for reviewers to evaluate data management plans. The inability to distinguish a good data management plan from a bad one negates their value. Grading data management plans and data sharing efforts will help define best practices and improve 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?

There are three major barriers to wide-spread data re-use. The first is ensuring that the proper intellectual property tools are developed and used, so that scientists are informed and protected. The second is federating existing data repositories through interoperable standards. This increases the ability to locate data. The third is development of analysis and visualization tools. For industry, the most important barrier is problematic intellectual property rights.

Thank you for your time,

Amanda K Rinehart

Very Interested American Citizen

References:

Bastow, R, & Leonelli, S. (September 17, 2010). Sustainable digital infrastructure. EMBO reports (2010) **11**, 730 – 734. doi:10.1038/embor.2010.145

Beagrie, N., Chruszcz, J., & Lavoie, B. (May 12, 2008). Keeping research data safe (Phase 1). Retrieved from http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

Piwowar, H. **(**January 11, 2012**)**. A View Of The Rights And Responsibilities Of The NSF Wrt Data. Retreived from http://researchremix.wordpress.com/2012/01/11/nsf-data-vision/

Wilbanks, J. (January 11, 2012). Response to RFI on digital data. Retrieved from http://del-fi.org/post/15710035064/response-to-rfi-on-digital-data