Fri 1/13/2012 1:46 AM
Response to RFI on Digital Data

Name/Email: Abigail Goben

Affiliation/Organization: I am writing this as a private citizen, though I am employed by the library at a publicly funded university which receives federal grants. Views herein expressed are my own as my employer has submitted their own response already. I am also writing as a blogger at HedgehogLibrarian.com, where I have been exploring data in my Friday blog posts with the advent of the data science movement. Show citation box

City, State: Chicago, IL Show citation box

(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?

The NSF Data Management Plan is an excellent start but it needs improvement. It needs teeth and there need to be gold standardsThere should be immediate adoption of an open access approach to data that is federally funded. This shouldn't wait for a private company to eat up data and hide it from researchersThere needs to be a policy on citation so that researchers creating the data have a way to be recognized for their workMetadata standards should be created so researchers have an idea of what is requiredShow citation box

(2) The primary stakeholder is the taxpayer who is funding this research--certainly not the publishers. That they were listed first, even before the scientists, is really troubling. Federal funding is already paying for creation of this data and as federal employees are not able to claim copyright and IP, these researchers acting as federal employees with the grant money should be similarly bound.
Show citation box

(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?

There are many leaders in the various scientific fields who, I'm sure, would be happy to consult on developing appropriate needs for the different types of data. Those types should be made as coherent as possible though so that data can move across disciplinesShow citation box

(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?

Again, I'd speak to experts in the field. I would avoid passing those costs to private vendors at all costs though, they're more likely to raise prices and cut off our access down the road. Show citation box

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

From this I'll speak from a librarian perspective, as that is my present role. Librarians have a number of roles that we already play that could immediately assist with this. Librarians are already teaching about data management plans and trying to raise the questions about preservation, storage needs, metadata, short and long term access, open data, etc. I'm preparing a tutorial for my faculty this spring and will be teaching it over the summer. We are already experts in creating metadata and have thousands of catalogers who could assist in tagging data in such a way that removes the burden from the researchers (granted--we would need funding for this, library budgets have been incredibly slashed). Libraries can help with storage, again if there is funding available to set this up and maintain it, and with access. Our missions are about providing access to information, we would be happy to take on data. Scientists and researchers already see us as repositories of books, is it that much of a move to have us be seen as repositories of data? Many librarians are already research partners working with the various scientific fields, so we know what kind of data they are generating. Finally, librarians, because of our role outside of the lab, tend to see the big picture and have excellent ideas about the translation of data, which would promote translational and collaborative science. Show citation box

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

Funding needs to be built into the grants, not tried to be scraped together by the institution after the grant is finished. I would suggest working to create hubs based around major public universities that are already working on much of this. If each university takes a subject area, then that could be funded there and data made available without each university or each scientist trying to reinvent the wheel. Show citation box

(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?

Make it as EASY as possibleProvide a federally funded way to clean up the data and apply metadata. If that means the US Government has a division of Data and Metadata, that might be the best thing for it. If the scientist need only turn over their data and someone else will do the cleaning and tagging, they are more likely to comply. Those funded locations would fit nicely into academic research libraries. Show citation box

(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?

Offer awards to young scientists for use of old and publicly accessible data. Too many young scientists are vying against highly published scientists who have moved on from their old data. Find funding ways to encourage people to reuse data rather than only value new data. Encourage academic institutions to grant tenure based on science using old data and/or Find a way to scrape data, particularly medical data of personal information so scientists and researchers can use it without having to go through IRB every five minutes. Show citation box

(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?

Get the experts in the field to develop what standard citation is. Require that publishers verify as part of their work (though they'll just hand it back to researchers)Find a way to convince administrators at R1 institutions that creation of data sets is valued by the federal government. Perhaps offer awards for data sets that have successful reuse? Show citation box

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.

Open and FreeShow citation box

(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?Show citation box

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

While I have no doubt it will be seen as naive, I think we need to lead the way to say that we have good information that we're willing to share. Being open and willing to share will start conversations. If we can establish excellent practices for data storage and access, we can also offer to assist overseas in storing their data and processing it for them. Show citation box

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

Publishers cannot be allowed to privatize data or say that by publishing it they have created some intellectual propertyA standard copyright agreement that data funded by a federal grant, while represented by the publisher, is not copyrighted. URIs/DOIs would help

I recommended for further reading this excellent post by Heather Pinowar: http://researchremix.wordpress.com/2012/01/11/nsf-data-vision/

I well recognize my own library bias in this and hope you can appreciate it. Libraries, particularly academic ones, are already trying to solve these problems. We have the expertise but we are struggling with time, funding and federal support. If you could provide those, we'd be more than happy to pick up the ball and run with it. We collaborate well with each other, we work hard to support our faculty, and we're here to provide access.

Thank you for the opportunity to provide feedback and please let me know if I can provide any further information.

Abigail Goben

--
Abigail Goben, MLS