Thu 1/12/2012 10:27 PM
Data RFI – input

Professor Victoria Stodden
Department of Statistics
Columbia University
New York, NY

http://stodden.net

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?

Federal funding agencies must require the digital datasets and computer instructions needed to reproduce published computational results be made openly available. In the pre-computational days, empirical scientists would describe their methods carefully in published papers with the intent that others would then be able the reproduce their work. Now, with the pervasive use of computers in scientific research, the steps taken in generating published findings are immensely complex and impossible to capture in the short methods sections as before such that the findings can be replicated. This is causing an enormous credibility crisis in computational science. It is impossible to reproduce the vast majority of results published in journals or presented at conferences today.

Sharing the information necessary to replicate the computational aspects of the experiment is a necessary response to this crisis. This means revealing the computer instructions, the code and scripts, as well as the datasets these instructions acted upon to produce the published results, at the time of publication. There is an emergent movement to create reproducible computational science, people and groups voicing concerns and creating sharing solutions from fields as diverse and geoscience, signal processing, statistics, bioinformatics and –omics research, MRI processing and neuroscience and many others.
These folks are working against a collective action problem: sharing is extra work for the scientist and not likely to seen as personally beneficial. Given that the incentive structure faced by computational scientists is heavily influenced by funding agency requirements, responsive funding agency policy is a necessary part of the solution.

One size does not fit all research problems across all research communities, and a heavy-handed general release requirement across agencies could result in de jure compliance – release of data and code as per the letter of the law – without the extra effort necessary to create usable data and code facilitating reproducibility (and extensions) of the results. The National Science Foundation now has a database of Data Management Plans and can collate this information to learn what is a reasonable sharing requirement in each field. These data would permit federal funding agencies to craft release requirements that are more sensitive to barriers researchers face and the demands of their particular research problems, and implement strategies for enforcement of these requirements.

This approach also permits researchers to address confidentiality and privacy issues associated with their research. I would hope for the funding agencies to move aggressively, then adjust if problems are encountered. The standard must be replication of the results by a contemporary in the field, without having to contact the original authors.

Data and code sharing requirements are not a foreign concept to scientists. The prestigious journal, Science, now requires authors to relinquish code and data from all articles they publish to any enquirer, for example.

To make the concept of data sharing coherent, it must connect to the reason scientists have the norm of sharing methods in the first place:
Reproducibility.

(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?

The Supreme Court has established that raw facts are not copyrightable. The sharing of datasets would presumably contain raw facts, although the datasets may have some residual copyright if they meet the standard of "original selection and arrangement." Thus the intellectual property status of datasets is not as clear as for written scientific articles. It is exceptionally important to preserve open access and reuse of the datasets, and I feel certain publishers will not do this adequately since they have not done this for published articles and it is important that the data and code that underlie published results do not become the property of the publishing houses. These must be openly available to facilitate reproducibility as well as transfer the knowledge and methods behinds the results beyond the ivory tower. I have previously published an intellectual property framework for scientific research, called the Reproducible Research Standard (cf. V. Stodden "Enabling Reproducible Research: Licensing For Scientific Innovation" at http://www.ijclp.net/issue_13.html) to untangle intellectual property rights associated with research release and clarify requirements.

The Reproducible Research Standard (RRS) realigns the Intellectual Property framework faced by computational researchers with longstanding scientific norms. The RRS suggests a licensing structure for research compendia, including code and data, that permits others to use and re-use code and data without having to obtain prior permission or assume a Fair Use exception to copyright, so long as attribution is given. The RRS utilizes existing open licenses that permit the free use of licensed work, so long as attribution is given, and is satisfied if the following four conditions hold:

1. The full research compendium, including code and data, is available on the Internet,

2. The media components such as text or figures, (including original selection and arrangement of the data), are licensed under the Creative Commons Attribution License 3.0 or released to the public domain under CC0,

3. The code components are licensed under one of Apache 2.0, the MIT License, or the Modified BSD license, or released to the public domain under CC0,

4. The data have been released into the public domain under CC0 or according to the Science Commons Open Data Protocol.

Using the RRS on all components of computational scholarship will encourage reproducible scientific investigation, facilitate greater collaboration, and promote engagement of the larger community in scientific learning and discovery.
Moreover, in evaluating compliance, we would also want to encompass the ability to build, run, and verify any source code. This might be accomplished using

* spot checks of the repository
* automated checks akin to unit tests
* tests run by a separate reviewer at the time of inclusion

(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?

The National Science Foundation has collected field-specific information on difference in data sharing through their Data Management Plan. I believe standard must come from the scientific community but they are foundered by a deep collective action problem – sharing data and code is a burden on the scientist and at the moment is not perceived as providing a payoff. The federal agencies should aggressively pursue data and code sharing policies that made the data and code that underlie published results conveniently available, such that the results can be replicated. Different communities can decide how to implement these standards but working toward these standards is imperative and federal policy leadership is a key part of doing so. Researchers with exceptionally different to share datasets or code bases should be apply for temporary waivers from federal sharing requirements, if more time is needed to create repositories or cloud access for example. Permanent waivers can be applied for on the basis of confidentiality or national security.

(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?

More costly sharing solutions, such as for very large datasets or codes, would reasonably be expected to take longer to implement.
Otherwise this question misses the point of sharing: reproducibility of results. Data is not shared because of potential industrial applications or because downstream users may find it beneficial. These are important effects, but they are corollary. The reason to share is to ensure that what we purport to be scientific facts are indeed reproducible. This is why we are facing a credibility crisis in computational science today and why data and code sharing, such that the underlying results can be reproduced, is imperative, and federal agency policy must take a clear leadership position.

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

In my experience there is enormous goodwill and desire to move toward greater reproducibility in computational science. The federal agencies can coordinate meetings between these stakeholders to implement data and code sharing plans. If a researcher does not receive further funding if data and code are unshared, this is the best contribution toward data sharing, and toward reproducible science.

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

Pilot projects and case studies. Fully fund some grants to be fully reproducible: sharing the data and code that generated the published results emerging from the study. Here are some examples of fully reproducible research that was very inexpensively shared:
http://sparselab.stanford.edu and
http://www-stat.stanford.edu/~wavelab . Both are enormous success stories with many downloads and many citations. The solution for many researchers does not need to be expensive or fancy, the policy just has to demand it. For other researchers, primarily those with very large datasets or codebases, they made need additional funding for cloud resources or repository creation for example. Research that was done with no sharing standards, other than the final published paper, is quite difficult to share in a reproducible way, but new research, where researchers are well aware that the data and code will be shared with publication, are much easier since efforts will be made as the research is ongoing. Choose several new grants and fund the applicants to create really reproducible research. This will provide measures of costs and needs beyond those described in data management plans.

(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?

Check if the data and code are openly available. This is trivially straightforward, for example following links in published articles, although does not verify the whether the data and code actually reproduce the published results. Leave it to the community to verify the results, but provide an avenue for downstream users to report whether or not they were able to replicate the results. This provides accountability and further evidence 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?

After the data and code are made openly available, then we can take a look and see how best to proceed on increasing access. Perhaps interfaces to the data and code will be the right way to proceed. At this stage it is imperative to just get the data and code open, without additional encumbrances for 3rd party usability. That can come later.

(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?

Scientists follow the norm of attribution. It must become standard practice to cite data and code use. Federal agencies can help provide stable URLs where data and code can reside through federal repositories.

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.

These are important but, as in the example cited in the question, must emerge from the community. Where federal agencies can help is facilitating the production of reports like the one in question by providing a mechanism for scientists in communities that lack agreement to apply for funding and engage a community meeting on the site of the funding agency to create standards.

(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 Protein Databank (PDB) was created in 1971 and thus has 38 years of experience in becoming a standard within the structural biology community. It is funded by international agencies with hubs in three countries. A PDB “accession number” is a precondition for publication in computational biology, meaning your data is available in PDB. Phil Bourne, one of the founders of the PDB, has noted that some tweaking in the policy may be in order now since some researchers appear to be tempted to get the accession number very early in their work and then feel they then have a license to publish. One remedy I might humbly suggest is the inclusion of the concept of reproducibility of published results: accession numbers for the data the results were derived from, and another resource locator for the code that will permit others to replicate the published results, using the data.

The PDB has enabled a new field of statistical studies and molecular dynamics research around the deposited structural data, impossible without access to the data as part of each publication.

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

I believe the most effective tool is convening meetings as described in the answer to question 10. Funding can be provided to bring international community members into the discussion.

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

We could experiment. If linking is required for publications arising from federally funded research for, say, the next 5 years, and repositories provided for researchers without access to communities repositories, this approach can be tested. This is straightforward for funding agencies to verify, by randomly checking the links in articles published from federal funds.