U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

Questions for the Record

Scientific Integrity and Transparency

Tuesday, March 5, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

---

1. On February 22nd 2013, the Office of Science and Technology Policy (OSTP) released guidelines, which outlined objectives for public access to scientific data in digital formats. The guidelines defined data as: "*the digital recorded factual material commonly accepted in the scientific community as necessary to validate research finding, including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects such as laboratory specimens*." Do you agree with this definition? Does it reasonably describe what data should be shared?

   The definition of data from the OSTP memo represents a useful starting point for public access to scientific data in digital formats. It is important to balance the needs for public access with the costs and burden that would be placed on researchers, universities, publishers, etc. The validation of research findings, particularly those from published articles, represents an important criterion from which to identify relevant data. It is also important to note that federal funding agencies generally do not provide funding to digitize print or physical materials though there are some exceptions (e.g., Institute of Museum and Library Services, National Endowment for the Humanities). Note that even if there are valid reasons for not offering public access to data, there may be still be valid reasons for preserving the data.

   There are some cases where documentation (in addition to data) is necessary to validate research findings. For example, notes from a laboratory notebook might be necessary to fully understand the processing of data. Even in such cases, the goal of validating research findings remains relevant rather than an overarching policy that could raise costs or burdens unnecessarily. Finally, even if physical data items are not available through public access, it is nonetheless important that researchers describe within their data management plans the means through which they (or their institutions) maintain, provide physical access to and preserve these objects.

2. The guidelines by OSTP outlines ten points that agencies must consider regarding the public access to scientific data in digital formats. Do you have any concerns with anything on this list? Is there any policy recommendation that you would like to see changed?

   The ten points from the OSTP memo describe a useful set of recommendations. There are a few additions or suggestions that I would recommend for the list:

   d) Ensure appropriate evaluation of the merits of submitted data management plans;

   - In order to properly evaluate merits of submitted data management plans, federal agencies should consider instructing their reviewers to comment on the plans specifically. For effective review, agencies should provide general guidelines noting that communities of practice vary by discipline or community.

   e) Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies;

   - Such mechanisms should be oriented toward enforcement of plans and actions as stated within the researchers' data management plans. To this end, federal agencies could support and highlight the development of machine-based mechanisms for compliance, audit, provable possession of data, etc. Additionally, data repositories that have undergo external certification and audit through mechanisms such as the Data Seal of Approval may provide a systematic means for addressing compliance.

   f) Promote the deposit of data in publicly accessible databases, where appropriate and available;

   - I am unsure what is meant by "databases" in this context but it would seem that publicly accessible "repositories" or "archives" would be a better choice of terms. Regardless of the type of system or technology, I believe that deposit of data should ensure the assignment of a unique persistent identifier.

   j) Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities.

   - It would be challenging for federal agencies to develop methods for assessment of long-term needs. Beginning with short-term assessments is more likely, particularly as it relates to metrics for assessing value of data. At this point, one could assert that the only metric is citation within a publication. However, as data repositories evolve and proliferate, there will be value with using, discovering, analyzing, etc. data independent of publications. The development of these metrics could represent

another opportunity for partnership between libraries, publishers, scholarly societies and the private sector.

3. Are there some situations or scientific fields where it would be cost-prohibitive to store and share data? Please explain. How should data be shared in these cases?

   There are nascent or planned scientific projects (e.g., Pan-STARRS and LSST in astronomy) that generate so much data with each individual survey of the night sky that there is not even sufficient hard disk to capture all of the data from the entire project. In such cases it is clearly cost-prohibitive to provide public access. It is my understanding that researchers have developed techniques for analyzing or sifting through such data in real-time. For these types of projects, it is perhaps most useful to document the procedures, processes, etc. that are used to analyze the data and the decisions regarding data acquisition, retention, deaccession, etc. in case there is a need to conduct additional surveys in the future.

   On a smaller scale, it is worth noting that in some situations costs could be lowered if researchers relied on economies of scale offered through community-based data repositories and archives. That is, there should be some third party or community based assertion of prohibitive costs, rather than an individual researcher who may not be using the most efficient options or means for data management.

4. One of the reasons for not releasing data in experiments is that it may contain personal identifying information. Is this a legitimate reason on the part of researchers not to share data? Please explain. How can we promote the sharing of such data while also assuring that confidentiality will be maintained?

   Please note that this response includes input from the Inter-university Consortium of Political and Social Research (ICPSR), which has extensive experience with data possessing personal identifying information. Disclosure: I am a member of the ICPSR Council (or Advisory Board).

   In certain domains such as social and behavioral sciences, it is not uncommon to collect personal identifying information in the course of doing research. The success of the social science research enterprise relies on the willingness of research participants to take part in experiments and surveys, and researchers are very aware of their obligation to protect such information. Procedures have been developed to protect confidential information during the research process and to assure that subjects cannot be identified in research publications. Disclosure risk is a term that is often used for the possibility that data from a research study might be linked to a specific person thereby revealing personal information that otherwise could not be known or known with as much certainty.

Concerns about disclosure risk have grown as more datasets have become available online and it has become easier to link research datasets with publicly available external databases.

Safeguards can be applied that allow access to data while at the same time ensuring confidentiality. Archive and repository data managers have developed skills in assessing and mediating disclosure risk and now can apply several approaches and technologies to ensure confidentiality throughout the data lifecycle. Working with these professionals, especially in the data collection planning phases, can allay concerns regarding disclosure risk. These approaches include creating public-use files by modifying the data (e.g., removing identifying numbers such as social security numbers), “coarsening” data (e.g., mentioning time intervals rather than specific dates), suppressing highly unique cases, sub-sampling and adding “noise” to the data.

In cases where data cannot be modified to protect confidentiality without significantly compromising the research potential of the data, access to the data must be restricted and stringent confidentiality safeguards imposed.

In these situations, archives require an application, review, and vetting process. Applicants are required to provide a research plan, Institutional Review Board approval, and a data protection plan. Approved users sign a Data Use Agreement, which establishes the rules for acquiring and using the data, a security pledge, and institutional approval and signatures. The agreement is particularly important because it specifies the guidelines that researchers must follow in the release of statistics derived from a dataset. Violations of the agreement are treated as research misconduct and violations of policies governing scientific integrity. Severe consequences are possible, including suspending research grants and legal liability. After an agreement is processed and approved, data are sent securely on CD, made available for secure download, or provided in a virtual data enclave (VDE), whereby the user must access and analyze the data on secure servers of the data provider. Results of data accessed via a VDE are vetted for disclosure risk prior to being sent to the user.

For data that present especially high disclosure risk, access can be provided in a data enclave where researchers must enter a secure facility to access the data. Investigators must undergo an application and approval process, as previously described, and archive staff reviews their notes and analytic output.

5. Would a move towards open-access of published data cause additional administrative costs for Universities and other Institutions that receive federal funding for scientific research? How can we minimize administrative burdens while simultaneously maximize access to data?

   A movement toward open-access of published data would almost certainly cause additional, administrative costs for universities and institutions that receive federal funding for scientific research. There is a challenging and delicate balance that needs to be struck between the benefits of open-access to data and new, additional costs. On the national scale, we may need to consider this balance in terms of how much new science we wish to support as compared to how much value we wish to extract from existing data.

   There is also a time dimension to consider. As noted earlier, the OSTP memo emphasizes data to validate research findings. This tangible goal represents a useful goal with which to make decisions regarding selection criteria for data. Additionally, systematic approaches to data management will almost certainly require lower costs than relying upon individual researchers' to manage their own data. As data infrastructure evolves, economies of scale arise and marginal costs reduce, it may become possible to consider other, tangible goals or classes of data for open access.

6. It is my understanding that a great majority of scientists will want data from a very small fraction of papers in the published literature. This data will most likely be only a specific subset of the entire data contained in the paper. If investigators are required to deposit data in a repository, there will be extra work especially on the investigator's time to make sure he/she is in compliance but there may be no clear long-term benefit esp if there is only a small chance that the data will be used again. There will also be additional costs associated with the storage of data. First do you agree with my assessment? Second, what then is the cost-benefit analysis of having a mandatory open data access policy?

   It is difficult to know the community reaction to open-access data. While it seems likely that scientists will *initially* want data from a small fraction of paper, the availability of such data might encourage greater discovery, re-use, etc. Focusing on specific goals such as verification of results and citation provide a useful, initial set of objectives for identifying data which should be deposited into repositories or archives. It is important to remember that federal funding is supposed to result in reproducible, citable science. As scalable, more efficient data infrastructure becomes available, both costs and time related to data management should diminish. With more data available, the prospects of unanticipated uses may increase over time. One of my Data Conservancy colleagues once said: "one scientist's noise is another scientist's signal" referring to the conventional wisdom of "one person's garbage is another person's treasure."

It is also worth noting the public's potential interest in scientific data. The experience of PubMed Central has demonstrated that the public does indeed refer to scientific literature for various reasons. The experience with the Sloan Digital Sky Survey (SDSS) provides evidence that similar trends may apply with data. There are approximately 10,000 professional astronomers but there are nearly 1 million registered users of the SkyServer that provides access to SDSS data.

Finally, greater availability of data could inspire the development of tools and services by a host of stakeholders such as scientists, publishers, professional societies and even the general public.

7. What specific infrastructure-technology requirements are required for the storage of scientific research data? Are University libraries or National Laboratories currently equipped with this type of infrastructure technology? Would an entirely new infrastructure need to be developed for the massive storage of data?

   I can only speak to the experience that my colleagues and I have gained through our process of dealing with Sloan Digital Sky Survey (SDSS) data for over a decade. Through our evaluation of storage systems, we have identified that current systems have limitations in terms of data preservation. For example, current storage systems do not possess formal auditing that is necessary for full-fledged preservation. Through personal interactions, I have heard similar concerns from other large-scale storage users such as the Internet Archive and the Science and Technology Council of the Academy of Motion Picture Arts and Sciences. I believe that development of new storage hardware and software based on these data infrastructure requirements represents an ideal opportunity for private-public partnerships that respond to federal funding programs. These funding programs should require working systems in operational environments as an outcome.

8. What are the potential cost-drivers for storing data? What are other costs that need to be considered?

   It is important to note that storing data is necessary, but not sufficient for sustained data sharing, access and preservation. In addition to storing data, archiving (e.g., protection such as checksums or computer generated codes to check integrity of data), preserving (e.g., format migration), and curating (e.g., adding value for re-use) are required.

   Regarding costs of storage, there is an unfortunate perception that storage is cheap so therefore we can store data easily. Not only does this perception ignore archiving, preservation, and curation, it also ignores the reality that storage *management* is not cheap. For example, the costs (in the form of computing cycles) for generating checksums can be significant or for migrating from one format to another (e.g., jpeg to tiff), depending on the amount of data.

There have been systematic attempts to measure costs associated with managing digital assets though the emphasis on data is more recent. For example, the LIFE project (http://www.life.ac.uk/) in the UK has "developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next 5, 10 or 20 years." The Australian National Data Service (ANDS; http://www.ands.org.au/) has developed a business plan. More recently, the OpenAIRE project and the European Commission has announced a tender seeking input for a Sustainability Model and Business Plan for digital infrastructure.

9. Are there any countries that have successfully implemented open-access data-sharing? Could the models used in those countries be used here in the US? Why or why not?

   It is fair to assert that, in many ways, Europe and Australia are both better organized than the US with respect to open-access data sharing. In the UK, some funding agencies require deposit of data into publicly accessible repositories. In Australia, the Australian National Data Service (ANDS) provides a national discovery service for open data deposited throughout their country. Arguably, these countries have also implemented data systems at the institutional, community and national levels, understanding that diverse "ecosystem" of approaches and systems are necessary for different functions related to open-access data.

   It would be difficult to imagine adopting these models verbatim within the US. There is a difference in scale and diversity of funding sources with the US. That is, there are fewer researchers, universities, etc. that generate data and fewer funding agencies that provide funding in Europe and Australia, many of which share common data management plan requirements. There is much the US can learn from our colleagues in Europe and Australia. We may possibly adopt elements of their approach.

   Having noted this, one could make a reasonable argument that while other countries are more advanced in the deposit, discovery and access realms, they are not more advanced in the data preservation realm and, in some cases, US-based data centers such as the Inter-university Consortium for Political and Social Research (ICPSR) and the National Snow and Ice Data Center (NSIDC) have long-term track records with data preservation (at least for certain types of data). Additionally, some new US-led data infrastructure development efforts such as the one I lead at Johns Hopkins (the Data Conservancy) have focused specifically on data preservation. Given this situation, there is comparative advantage to working with our colleagues in Europe and Australia.

   NSF (and perhaps other federal funding agencies) often seeks international collaboration as part of solicitations but do not allow use of funds to support international participants. Understandably, this reality makes it challenging to secure international partnerships. There have been joint NSF/JISC and NSF/EU funding programs but these programs can

lead to greater administrative burdens in terms of reporting, oversight, etc. Streamlined programs that foster international partnerships would be worthwhile. The Research Data Alliance (rd-alliance.org) has been launched with a goal of fostering collaboration on a global scale toward data sharing and interoperability. At this point, NSF and NIST are the only two federal agencies directly supporting the Research Data Alliance (RDA). Disclosure: I am involved in RDA, particularly as the leader for the task force planning the 2nd meeting of RDA in Washington, DC from September 16-18, 2013.

10. What specific support could the federal government contribute towards a permanent community-maintained archive for storing research data, that non-federal organizations could not provide?

    The federal government can and should provide funding toward the development of community-maintained data archives. There is value to building infrastructure at scale (i.e., beyond individual universities). While the private sector has an important role to play, certain functions such as preservation – while essential – are unlikely to be profitable. It is worth considering the role of the federal government with other types of existing infrastructure that rely upon a combination of federal, state, university and private funding and resources. If one considers other forms of infrastructure to support data-intensive science such as high-performance computing, there is a diversity of options ranging from university-based or company-based services. Some of these options such as supercomputing centers receive federal funding support. However, even in cases of federal support, there should be a real sustainability plan that does not rely upon additional rounds of federal investment.

11. A 2007 GAO Report entitled "Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research" states: "*The scientific community generally rewards researchers who publish in journals, but preparation of data for others' use is not an important part of this reward structure*." What are your suggestions to change this structure?

    This matter relates to the reward and recognition structure that is part of universities' academic policies and practices. There is a tremendous diversity and complexity to this framework that the federal government cannot address. Having said this, there are existing mechanisms within the federal funding environment that can be leveraged effectively. For example, NSF recently changed its guidelines such that instead of mentioning "five most relevant publications" within the NSF-compliant two-page bios, one can know list "five most relevant products" ostensibly to include other output of research such as data. Similar mechanisms should be leveraged as well. If data are included in this manner (e.g., NSF two-page bio), then they should be cited using a persistent identifier to ensure reliable, sustained ability to discover and review such data.

12. What specific technical standards need to be considered when storing data for open access?

    There are many existing standards. Consider the growing list that the Digital Curation Centre in the UK maintains at http://www.dcc.ac.uk/resources/metadata-standards/list. Each scientific community has its own set of metadata standards. There are attempts to map between these standards but it is perhaps more important to focus on data types. The aforementioned Research Data Alliance (RDA) has two working groups focused on persistent identifier types and data type registries. These groups are considering the various *types* of data (e.g., images, videos), the salient or representative properties of these types, and the role of persistent identifiers with these data types. This type of foundational work focused on data types and identifiers is necessary before considering a universal set of metadata standards that may be applied across a variety of domains and contexts.

13. What federal agency and/or other entities would be appropriately suited to determine standards for storing data?

    As mentioned, the Research Data Alliance has undertaken global community-driven and guided work in this regard. The National Institute of Standards and Technology (NIST) would seem to be an appropriate agency in this context. Various federal funding agencies have natural connections to various scientific communities (e.g., NASA with space sciences and earth sciences) in a manner that facilities development of community-based standards.

14. On July 29, 2010 Dr. David Lipman testified before the House Subcommittee on Information Policy, Census and National Archives. While most of his testimony centered around open-access issues, he noted that the National Center for Biotechnology Information (NCBI) produces more than 40 databases, including GenBank and dbGaP. He also mentioned other data intensive activities that his center is currently handling. Based on his testimony, and other publically available information about the activities at NIH Pubmed Central and NCBI, do you think that they have the technical capability and infrastructure to store, archive, and handle large amounts of data (i.e. achieve the purposes of open-data)? Please explain. If there was a movement towards a national repository for scientific data, would it not be better to build off of existing infrastructure at NIH and NCBI? What are other issues that should be taken into consideration when going towards a single repository model? Finally, based on your experience, do you see any potential cross-agency issues (for example between NIH and NSF) that might make a single federal repository inefficient or not worthy of pursuing?

    I do not know enough about the technical capability and infrastructure of NCBI to comment in detail. I was the Principal Investigator of an NSF-funded evaluation of pros

and cons for a potential open-access repository of publications resulting from NSF funding. Based on this evaluation, I can offer the following observations or comments.

One of the main reasons that NIH can provide infrastructure for publications and data is the existence of the National Library of Medicine (NLM), which is itself a type of infrastructure. Noting that other funding agencies such as NSF do not have an equivalent resource, it is worth considering whether NIH or NLM could provide relevant infrastructure or services. Having said this, while the approaches and processes that NIH or NLM have undertaken might be useful, it is not clear that the specific choices and workflows would apply effectively to other scientific domains or communities.

As with other infrastructure development, there needs to be a balance between national or centralized approaches and community or decentralized approaches. A national repository could offer significant economies of scale (e.g., for storage) but might result in too rigid a framework to effectively describe or share data across a diverse set of domains or communities.

It may be more effective for the federal government to identify cross cutting, common components of data infrastructure that could be applied across different funding agencies. For example, referring to the aforementioned discussion of data types and identifiers, the federal government could require funding agencies to mandate the use of persistent identifiers but not prescribe the specific choices. This type of approach represents a balance between an overarching national approach that recognizes the need for flexibility within scientific communities.