Fri 1/6/2012 7:08 AM
Response to questions on Public Access to Digital Data

Submitted by: Walter S. Snyder

Department of Geosciences

Boise State University

Based on 10 years experience in geoinformatics, including being involved is starting the geoinformatics program at NSF, development and management of several community data systems (GeoStrat (NSF supported); Geothermal Data Exchange (GDEx) (DOE and NSF supported), the National Geothermal Data System (DOE)), and continued national and international collaborations on data issues.

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?

The biggest mistake that the federal government can make is to assume that the agencies themselves will always be the best stewards of the data generated by federal funds. There is at least a two-tier issue here. First, data generated by personnel of the agency and through procurement contracts are perhaps best preserved by the agency itself (and therefore the agency must provide access to these data). Second, for data generated by extramural funds, through grants and cooperative agreements, the agency itself may not be, and perhaps in most cases won't be, the best stewards for these data. This issue is one of who can best determine the user's needs re the data - starting from data generation to management to working with data to publication to long-term, open access? Typically agency personnel are not the best people to make those decisions, and whereas their input is valuable, the decisions of how to construct and operation external data systems must reside with the user communities. It is important to note that these users are precisely the ones targeted by the COMPETES Act, and it would be presumptuous for agencies to assume they know better than the users what needs to be done. The approach to this two-tiered problem varies by agency and within each agency - and this is not a surprise. For example, in general NSF understands the importance of the users, and almost errs on the side of being too flexible and allowing unsustainable approaches to data on a project-by-project basis. For many other federal agencies, is completely understandable that they want to have and serve all data their funding has paid for and put these data on agency-controlled servers. servers and/or data sites - including that generated by extramural funding.

Two things here - this is fine for internal agency data, but is not optimal for data derived from extramural funding. This will lead to lower quality and incomplete data simply because it is a forced approach to data acquisition and management versus one that comes from and is for the user and data producing communities.

In short, policies must recognize that the agencies themselves have different missions and mandates than the user communities, and no matter how much one writes or talks about it, agencies are not representatives of the user communities. The agencies have to first and foremost worry about their own existence as a business entity, placed second is their role (depending on the agency or group within the agency) as a servant of the public, the user communities. That is not necessarily a bad thing, it is what it is and agencies should be the stewards of the data they produce for themselves, but they should not control (but can and should participate in) facilities that manage data generated by extramural funds. These need to be agency-supported community systems.

Policy: Data generated internally by an agency can be hosted by the agency; data generated by extramural funding of grants and cooperative agreements should be funded by the agency, but hosted by community-based data sites if at all possible.

(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?

Within a document, there is a need to distinguish between the written word and data contained in the document. The data, if it is the result of federal funding, should be immediately available to the public, and at a minimum be able to be used by anyone free of charge or other restrictions. The interpretations of those data however should be given the same copyright protection as any written document. These interpretations may have been possible because of federal funding, but they are the intellectual creation of individuals beyond the boundaries of a particular batch of funding and should follow standard approaches to intellectual property, including the individual passing the copyright to the publisher.

Policy: the data generated by federal funding should be freely and openly available within a reasonable time frame, but the interpretations of these data as presented in published and unpublished documents remain the intellectual property of their author or the author's assignee.

(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?

This is a very insightful question and focuses on a critical issue - is there one best way to handle all digital data? Is there a singular approach that will work for all needs and communities? The answer is "no" - although there are many who think a single way can be created. The problem with that singularity view is that it would have to be forced, and in the process loose critical knowledge value of the data - this is a topic for a longer discussion. At a high level, data sharing among agencies will be possible by developing high-level standards for data discovery and sharing, but that should not dictate how data are captured, stored or even served to particular user communities.

Policy: each agency must assess the context and use goals of the data generated by their funds for two distinct groups of data: 1) internal data, and 2) for data generated by extramural funding. For this second group, the agency must consult with or assign this assessment process to users outside of their or any agency (i.e., to the user community). The agency should then compare their needs to what other agencies, groups and institutions are doing, and establish their own best practices.

(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?

This is a very difficult question, because assessing the value of data is such a vague endeavor. Is it all about the cost of re-generating the data? Is it about its intrinsic value to understanding something else? Data coming from machines and sensors is far easier to capture and store - the problem being the quantity of the data - but that too is not longer a major problem. Some data have very long "shelf lives", e.g., geologic data, others very short life, e.g., medical research data. So one could ask agencies to perform a qualitative assessment of the data groups they handle - the emphasis here is "qualitative" - if you try to force a quantitative assessment, then you will, from the start, under-value some groups of data.

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

These "stakeholders" or "users" need to self-organize, and that may require some funding from the relevant agency. They need to address all the questions poised here and many more. Then their recommendations need to be implemented by the relevant agencies - at least addressed in

an open way that also allows stakeholder rebuttal to a higher authority. OSTP could establish a clearing house for such input. Why is this important? Because the agency that provides the funding to the stakeholders can have undue influence and control. An a priory notion that the agency always knows best, while true the vast majority of the time, is not always the case, and the stakeholder has no recourse; OSTP should provide that recourse.

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

We can speak from personal experience here. First, the agency needs to realize that preserving and making digital data accessible is not inexpensive, and a rule of thumb may be a minimum of about 3% of their operational budget for funding outside community data centers and perhaps 9% (total cost assessment) internally. Second, the agencies need to fund several key "community data centers" that not only handle data from extramural funding, but could also handle some of the agency's internal data at a cost lower than the agency can do internally (on a "total cost" basis). For extramural funds, if the agency funds one or more community data centers, then subsequent cooperative agreements and grants can utilize these centers for their data management, and include modest amounts in their budgets to pay to the center for this service. Thus, the agency funds the core operations of a data center, and each subsequent award pays for its data management needs in an affordable way.

(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?

Each project, whether internal to the agency (including procurement contracts) or extramural funded projects could produce a data management plan that details the data they intend to generate. This is not a difficult or burdensome exercise, and could take as little as two hours per project. Then there is a metric for comparison. These plans need to be open to amendment as the project proceeds. Then, if systems exist that can work with the data producers to capture the data as close in time to when they are generated as possible, you decrease the burden on the data producer, improve the amount and quality of data captured, and allow for an ongoing assessment of progress on data capture by each project. This need not be burdensome if it is integrated into the daily workflow.

(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?

This can primarily be done through the establishment of "community data centers" - these can produce jobs for each center, but also potentially spin off collaborations with industry and business that may want to use the data in innovative ways. Simply put, if all the expenditure for preserving and making digital data accessible is contained within the agency there will be little stimulation of outside jobs and innovative outside use of these data.

(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?

Again, this is from personal experience. In our data systems we now have established a first tie to a professional science society where the society, through its publications, will provide Digital Object Identifiers (DOI) for datasets. In addition our data system will host all data referred to in their publications (journals and books), including the said datasets and provide open, free access to these data. Thus the creator of a dataset gets publication credit for the data as well as the published paper, and each time another person references that dataset in subsequent publications receives another citation. This becomes a win-win situation that reduces the costs to the science society, and provides an easy and consistent way for the data producers to manage and get credit for their data.

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.

The underlying problem is that there are numerous "community-driven data standards efforts" - so which ones do you use? Also, there is lots of hype around particular approaches, and although OSTP seems to have bought into the full story of the "Semantic Web" - other experts (see recent PEW report) have noted that it has not been as successful as billed. Nevertheless there are many "standards" that are now commonly used - for example the semantic web's web services, OGC/FGDC geospatial standards, etc. Some data groups are easier to set standards for than others - for example sensor/sensor array data is relatively easy, but science data based on field and laboratory analyses is far more difficult. Again, the key is what is implied in the question; to allow user communities to come together and work, over time on developing and refining the way they approach data. Agencies should follow these community-based groups, not tell them what to do. Convergence will happen as several studies point out, if you let it happen naturally and do not force it. A note of caution: defining a user community is difficult as well. What we have discovered is that the academic user community is quite distinct from that of state and federal agencies and that, whereas you want dialogue and to promote a path

towards convergence among academia and the agencies, initially these communities should acknowledge that they have different mission and mandates and that they need to get their own houses in order first - or you run the risk of the more powerful partner forcing "standards" which are not the best for the particular community and don't last.

(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 Environmental Information Exchange Network (www.exchangenetwork.net) which has been operating for 7 years, was developed such that each node on the network can continue to do what its individual mission and mandate requires, yet a series of data exchange templets have been developed based on network-wide (i.e., community) input and agreement that allows the seamless exchange of data among the nodes on this network. An example of one that has not worked well is the U.S. Geoscience Information Network (USGIN); the story is in the details, but it has been imposed from effectively a single source and does not represent development in an open, true community environment, and will survive only as long as it has political support. The official written descriptions capture correct sentiments, but the key is how it has developed and its implementation. It seems to fit the needs of select state agencies and the US. Geological Survey, which is fine, but when it is forced onto academic and industry communities it fails because of the lack of openness and inclusiveness. The lesson here is that when you develop community-based groups where the smallest has as much say as the strongest member, where there is a spirit of collaboration and true exchange of ideas, you can build community standards, protocols and best practices that not only work theoretically, but practically and that can be sustained. A further lesson is that you can learn as much from failures as successes.

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

U.S. Federal (and perhaps some state) agencies need to be at this table, but no more so than as equals to the nongovernment user communities. If the agencies are not producing what users need - why are they doing it? There must be a balance in power, votes and representation. There have been and will continue to be opportunities for such international collaboration. When they arise the relevant federal agencies should see these as opportunities to participate, and even support the efforts with funding, but be directed to not attempt to control the process or outcomes. In short, typical government agencies have difficulty cooperating within the U.S., much less internationally; user communities much less so. When opportunities arise they should join the efforts and participate and follow the lead of the user communities rather than being the leaders A notable example of a successful international effort is OneGeology (www.onegeology.org), which is a collaboration of mostly the national

geological surveys from over 117 countries. Its focus is geologic maps, and for the scale of the maps they are working with, it has been highly successful. (With one caveat - it effectively does not include, therefore represent the non-government geoscience communities).

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

See the discussion of question #9. What you don't want to do is have agencies take over the role of publication - that simply is too expensive, competes with private industry, and culturally won't work. What you have to do is provide mechanisms that allow for the natural convergence on this issue. In particular accept the fact that the words of explanation and interpretation about data are separate from the data themselves. Require the publication of data, then the linking of data to the publication will happen naturally, championed by the user communities.