National Academy of Sciences

National Research Council

Public Comment Meeting

Concerning

Public Access to Federally Supported R&D Publications

Public Commenters

May 16 - 17, 2013

National Academy of Sciences
2101 Constitution Avenue, NW
Washington, D.C.

Proceedings by:
CASET Associates, Ltd.
Fairfax, Virginia 22030
caset@caset.net

Table of Contents

PROCEEDINGS

DR. HAUSER: Here are my comments as footnotes to the excellent presentations by Altman and Stodden. I will make six observations, but no recommendations. All but the last pertain to my work over some 40 years as a social scientist, who has used public research data with federal support.

There is nothing new about public access to federally supported research data. My own experience includes release of the two major surveys of social mobility in the United States, the last of which, unfortunately, was conducted 40 years ago. I spent 41 years running something called the Wisconsin Longitudinal Study, which has followed the life course of 10,000 plus Wisconsin high school graduates. Almost all of those data are public. If you Google Wisconsin Longitudinal Study, you will reach the tip of an iceberg of data and documentation.

I created a uniform edition of the current Populations Surveys and also helped to create public microdata samples from the Decennial Census of 1940 and 1950. If you want to see what resulted from that, look up IPUMS at the University of Minnesota. Of course, there are dozens and dozens of other social and biosocial surveys that have been created for public use and provide templates for sound data policy.

Second, there is excellent guidance, experience, and technology for protection and use of such data, including data that are quite sensitive. Some agencies have not taken full advantage of such developments, which are increasingly necessary as massive bodies of linked data are created.

Third, access to federally supported research data will depend heavily on the responsibilities of actions of investigators and not merely on those of agency personnel. As in publication, that will cost time, effort, and money and require incentives for compliance. Grant termination is the enemy of compliance. The infrastructure for data protection and dissemination is far better developed in some disciplines than others.

Fourth, data do not speak for themselves. They require documentation. In some disciplines, there are well developed standards for data structure and documentation. However, when data are actually used analytically or otherwise manipulated to yield revised, linked, or combined data products, those manipulations can and also should be documented. For example, some journals now require authors to submit the code that has been used in published analyses. We have heard about that this morning.

Should public access to such documentation be made a part of publication standards or data standards or left to chance? Along with many colleagues, I believe that findings that cannot be reproduced are not science. Look no further than the example of the Reinhart Rogoff paper on national debt and economic growth. Will it be necessary to develop standards for documentation of data management as well as data files?

Fifth, some federal data are already well covered by legislative or regulatory provisions that are highly protective, in some cases, unreasonably so. You may want to consider new or modified provisions that might apply across agencies and that would better encourage legitimate use and prevent misuse.

Some investigators now invoke bogus claims of sensitivity or privacy to keep data proprietary. My favorite example involves claiming privacy as a reason not to release a variance/co-variance matrix. Think about that. The user as well as the use might be considered. For example, commercial users of some kinds of data are not far less subject to regulation than our scientific researchers who may access the same data.

Sixth and finally, with regard to data on human research participants, many of you are, no doubt, aware the advance notice to proposed rulemaking that was issued in December of 2011 yielded well over 1,000 comments and has not as yet resulted in new regulation under 45CFR46, the Common Rule.

The Division of Behavioral and Social Sciences and Education, is undertaking an interdisciplinary study, which I hope will yield sound recommendations for regulatory change that will benefit all of the relevant sciences. A summary of the proceedings of our March workshop will likely be released by July. An enlarged committee is now beginning work on a consensus report. Thank you.

Agenda Item: Public Comment

Session 1

STAFF: Thank you. Next up we have David Fearon. Whenever you are ready, the microphone is yours.

DAVID FEARON: I am a data management consultant for JHU Data Management Services, part of the Sheridan Libraries at Johns Hopkins University. Our service began in July of 2011, providing in-person consulting on data management plan preparation and training on data management best practices. Our department also operates in archives built specifically for long-term public access to research data developed by the Data Conservancy project.

As one of the first comprehensive service groups for data management and archiving among U.S. academic libraries, we are keenly interested in helping implement the goals of the OSTP memo at John Hopkins. I will focus on one suggestion relative to the issue of supporting faculty reviewers of grant proposals, who must evaluate data management plans.

As we consult with researchers preparing data management plans or DMPs, as we will probably end up calling them today, primarily for NSF proposals, many researchers will admit that codifying their plans for data management and sharing at the proposal stage can be a valuable exercise. We have also had researchers tell us that they did not have a clear incentive to produce more than a cursory plan or to seek out what they need to know in order to meet the NSF's basic guidelines.

We feel like the lack of guidelines for grant proposal reviewers may be a significant factor in the overall incentive to produce effective DMPs, not just for the sake of compliance. Reviewers can encourage recognition among their fellow researchers that DMPs are useful and that data sharing is important. We have heard, anecdotally, from some faculty reviewers for the NSF proposal that they do not know what to look for when they are evaluating a DMP. Some simply check off whether one was included with the proposal.

By and large, reviewers may not be receiving encouragement to focus on data management plans and proposals, therefore, they are not passing along the feedback to grant applicants about the quality of their data management plans or of their data sharing efforts.

The White House OSTP memo, section four, item D, addresses the need to, quote, ensure the appropriate evaluation of the merits of submitted data management plans. We wish to underscore that funders consider offering guidance and resources designed specifically for grant reviewers.

As an example of such a resource, we, at JHU Data Management Services, have developed a worksheet for our JHU faculty grant reviewers that provides a simple checklist for evaluating and comparing proposal DMPs. The checklist covers NSF’s recommended content for data management plans and also includes items that might be found on more thorough plans that reviewers can mark as extra credit.

The worksheet also includes guidelines and illustrative text from exemplary DMPs for each topical section. The guidelines are formatted as Microsoft Word template. The worksheet may be used on computers, tablets, and as a printout.

We have had some positive feedback from faculty reviewers that the guide helps make reviewing data management plans more convenient while conveying the main points of an exemplary DMP and also a data sharing plan.

We offer the Grant Reviewers DMP Worksheet as an example to build upon as the funders consider resources for proposal reviewers as part of their overall plans to expand data management and data sharing.

The Reviewers Worksheet is available on our website with a link on our submitted document for this talk or you may find us by searching for JHU Data Management Services. I also brought some copies with me. Thank you.

STAFF: Thank you. Up next, we have Clemencia Cosentino.

CLEMENCIA COSENTINO: Thank you. Good morning. I am a Clemencia Cosentino, a researcher at Mathematica Policy Research, which is a non-partisan organization that conducts policy research, evaluations, and large data collections. Some of you may have even used some of our work through the CSTAT, which is put together at Mathematica.

I would like to thank you for the opportunity to participate in this meeting and share my own personal thoughts on an initiative that I applaud and I think can potentially be of great benefit. I also believe there are several issues that should be considered in developing plans to provide public access to data. I would like to highlight just a few of them from the perspective of the work that I do as a researcher.

First, the plans should take into account that data are collected for different purposes, to conduct a research, which has been talked about a lot today, to monitor projects, to evaluate projects, and to conduct large, independent data collections to support research by others. These data collections are supported using different procurement mechanisms, grants, or contracts.

What is important about this is that it may have implications for whether and how the data are released. Sometimes data collection efforts contemplate the release of data and plan for it, but sometimes the release is not feasible, given capacity constraints, ownership right, or confidentiality agreements.

My point here is that clear rules have to be developed and implemented prior to data collection. They become a real problem if they are not set at the outset. There was a lot of talk earlier this morning from Victoria Stodden about releasing code as well. Again, the same would apply to that.

The second point I would like to make has to do with confidentiality, which must be appropriately protected at all times to be in compliance with national legislation, such as HIPAA, as enforced through both IRB and OMB requirements. This is mentioned in the OSTP memo that was released in late February.

What I would like to say is in addition to protecting confidentiality, we also need to ensure the protection of sensitive information, even if it isn't confidential. It is more difficult to establish rules to do that, but it should be included in the plans that are drawn up by different agencies.

The third point I would like to raise pertains to the creation and maintenance of both restricted and public use data files, which must come with proper documentation so that they can be used correctly by users. Some data collections include this. At Mathematica, for example, we frequently produce such files and documentation, but some grantees and contractors may not have either the capacity or the resources to do this. Assistance has to be provided to them, as well, which brings me to my fourth point, which was also mentioned this morning about resources.

Additional resources are not being allocated to this effort, but appropriate resources need to be allocated to prepare files for release and to make them available. The OSTP memorandum indicates that agencies should allow the inclusion of appropriate costs for data management and access in proposals. This cost can be substantial, depending on the nature and the scope of the data collection. That should be taken into account as well.

Last, I would like to mention, briefly, long-term maintenance of these datasets. There was a new memo released last week on May 9th by the White House, which provides clarifications on a number of points for data collection sponsored by federal government agencies for their own information systems. It mentions that in these cases -- so in some cases -- federal agencies will be the repositories of datasets.

In those cases and in others, I think it will be really important for agencies to consider coordinating their policies and their guidelines to ensure consistency of requirements as it is mentioned, actually, in one of the memos. Thank you.

STAFF: Thank you. Up next, we have Janet van Cleave. The microphone is yours whenever you are ready.

JANET VAN CLEAVE: Thank you. My name is Janet H. van Cleave. I am a Clinical Assistant Professor at New York University College of Nursing. I graduated with my Bachelor's and Master's of Nursing from the University of Pennsylvania, received my PhD from Yale University in 2008, and completed post-doctoral training at the University of Pennsylvania in 2010. My research focus is older adults as they transition across care settings.

I am designated by NIH as an early investigator. Therefore, I am here to support public access and federally supported data. The opinions that I present here are my own and do not reflect those of New York University, not the College of Nursing, nor my collaborators or mentors.

There are two important reasons why I support public access to federally supported data. The first reason is that public access to federally supported data provides important opportunities for early investigators, such as myself, to conduct scientifically studies that we could not afford to conduct ourselves. The findings from these studies, then, enable investigators like myself to advance our understanding of important research questions, generate new hypotheses, demonstrate productivity, in terms of published articles and research presentations, and, thus, generate competitive NIH proposals, which, in this era of decreasing support for NIH funding, is critical.

The second reason I am here to support public access to federally supported data is to increase the portability of data across institutions. As you can tell, as an early investigator, I conducted research across several institutions and have found barriers to carrying this data and carrying my work across institutions. Policies restricting public access to federally funded data creates barriers to portability of data, discourages collaborations, and decreases early investigators productivity at critical periods in their careers.

In conclusion, I am speaking today to support public access to federally supported data and to thus continue advancing early investigators and, ultimately, the nation's scientific endeavors. Thank you.

STAFF: Thank you so much. Up next, we have our first WebEx attendee. Give me a second to switch over the video. Jillian is on the line, but we don't have video from her. Jillian, whenever you are ready, the microphone is yours.

JILLIAN WALLIS: Hello. I am Jillian C. Wallis, a post doctoral researcher at UCLA and the University of Michigan. I would like to start by saying that as someone who has spent the last decade studying data sharing attitudes and practices within the sciences, I am pleased that the Executive Office is engaged with promoting sharing of academic research. I would also like to take this opportunity to share some observations of the points of view of data producers.

This policy proposed in the Increasing Access memorandum would be just one of many already in place to encourage and require publicly funded researchers to share their data with others. There are already policies at the discipline level from journals, research centers, and at the funding level.

Given all of the pressure coming from the top down, all researchers should be sharing their data. Sadly, this is not the case. Scientists are not petulant children who would refuse to share data just to be spiteful, though, so there must be some other pressure of equal magnitude coming from the opposite direction that needs to be overcome. The real question is what are the disincentives from the bottom up that are impeding the top down pressure?

In a recent survey of over 1,300 people, 75 percent of researchers have shared their own data at some point and six percent have made all of their data available. In a much smaller study from my own research group with 43 interviews and ethnographic observation performed in an NSF funded research center, we found that everyone who talked to us was willing to share data, but that only half of them had been asked to do so. When asked to do so, researchers shared their data and they shared it through a personal interaction with the data re-user.

One of the ways that researchers can share their data is through depositing data into some sort of data repository. By depositing this way, researchers are not waiting until they are being asked by other researchers to share their data. They have the added benefit of increasing the visibility of their data to potential re-users and providing long-term support.

In the other survey, 78 percent of researchers were willing to deposit their data in some data repository.

The actual rate of deposit that we found from our much smaller study was closer to a third of our interviewees. We are still seeing not much traction with repositories, even though deposit in them would fulfill the various requirements that are already being applied to researchers.

Another way that researchers can share their data is by putting them online. This is something that researchers do with greater frequency than the repository. According to the other survey, nearly half of their respondents made their data available online. Our smaller study confirmed this result.

For our study population, nearly twice as many of the researchers had made their data available online as opposed to a repository. Unfortunately, there are really well known problems with making data available online, including lack of long-term support and overall issues with accessibility. From this, we can conclude that at least part of our problem is that repository deposit is still too high a bar of entry for the majority of researchers to overcome, given that they are willing to take a lower bar approach, such as putting their data on a personal website.

As mentioned earlier, the majority of researchers we interviewed preferred using a personal interaction to share their data with other researchers. We think this is because of the conditions under which these researchers are willing to share their data. Some of the most popular conditions are interviews provided where retaining the first right to publish from their result, receiving proper attribution as the source, wanting the requestor to be known to them, the ability to negotiate sharing events of exchange, et cetera. These conditions all seem quite reasonable and are all easier to manage with a personal interaction than through a system that might not encode their conditions properly.

From this, we can conclude that researchers are really willing to share their data when presented with the opportunity to do so. There are a few hitches holding them back. First, there are very few requests for data to be shared. Second, data sharing is hard. Researchers have two equally bad options beyond personal interaction. Depositing in a repository is difficult, but yields greater benefits. Putting data online is an easier option, but suffers from various accessibility issues. Third, why would researchers go through the trouble to deposit in a repository when they see very little demand for the data?

Data sharing and data reuse are really two sides of the same coin. We need to overcome the hurdles to both data sharing and data reuse in order for data sharing policies and memoranda to be more useful. What would really help are policies and better funding opportunities to encourage reuse of existing data that would hopefully jumpstart the data sharing engine. Thank you very much.

STAFF: Thank you. Let’s try Amy Nurnberger. Amy, whenever you are ready, the microphone or video camera is yours.

AMY NURNBERGER: Right. Hello. My name is Amy Nurnberger. I am the Research Data Manager at Columbia University. Thank you for this opportunity to offer comment on increasing access to the results of federally funded scientific research. Columbia University is dedicated to advancing knowledge and learning at the highest level and to conveying the products of its efforts to the world. We support the memorandum’s objectives within this context. We particularly appreciate the efforts of the agencies to consult with the various parties that will be affected by the resulting policies and applaud the goal of developing policies consistent in their compliance requirements.

Beyond consistency, the keys to encouraging preservation of digital data and providing public access to them are a definition of research data that serves the objective of making the results of federally funded research useful for the public, industry, and the scientific community, a clear framework for communicating the usage rights for those data, the provision of open data repositories that adhere to common standards for defining, identifying, describing, and storing data by facilitating alliances of existing repositories and creating and funding interoperable repositories managed in partnership with the government, a funding system that supports the deposit, maintenance, and preservation of federally funded research data and provides for the possibly unanticipated costs of the data stewardship, investment in technological infrastructure that makes data management and compliance practical, painless, and intellectually profitable.

To attain these goals, we encourage agencies to work closely with discipline-specific groups, such as professional and scholarly societies, information technology specialists, librarians, and research administrators, creating alliances to fulfill some of the roles that may best be taken on at an agency level, such as creating data aggregating portals that provide a unified point of access to disparately archived data, promoting and incentivizing best practice solutions for data archiving and preservation, reshaping existing grant management workflows to accommodate mechanisms for making research data available in such a way that important stakeholders are minimally burdened by these new requirements, thereby reducing both administrative costs and obstacles to compliance.

This reshaping may take the form of automation based on clearly communicated standards that integrate compliance into existing workflows for granting, research, and publication distribution. Also, providing a centralized index of identification and description standards to facilitate the discovery, reuse, and impact tracking of data is an important role. Such a resource could foster adherence to community practice and reduce barriers to interoperability.

Addressing questions of governance and adoption or development of standards and conventions among disciplinary communities is another area where we feel discipline specific groups can assist agencies. These questions include, at a minimum, issues of establishing baseline metadata requirements for interoperability and discovery, requiring that labeling be done in human and machine readable formats, ensuring the clear labeling of data so that all stakeholders are aware of the use conditions of the given data set, encouraging the assignment of use conditions at all steps of the data life cycle.

As we know, raw data may go through many transformations before they find their way into publication and other end uses, but the ability to trace those data end to end is an essential part of the verification process, assuring that data are clearly associated with the publications that cite them and the code that is used to process them for purposes of validation and reproducibility.

At this point in time, you, the funding agencies, are presented with a variety of possibilities in terms of your potential roles and actions with respect for provisioning access to data. We encourage you to continue to approach these opportunities in accompanying members of your scientific and discipline-specific communities to develop policies that enable consistency, that provide useful definitions of research data, that allow practical funding and compliance practices, and that enable standards facilitating data discovery and reuse.

Working together, we can develop open paths that achieve the mandate’s intent for making federally funded research data publicly accessible. Thank you again for this opportunity.

STAFF: Thank you for your comments. We are going to go to our final WebEx person of this session, who is Marciela Olivam.

MARCIELA OLIVAM: Hi. My name is Marciela Olivam. I teach architecture and environmental design at Los Angeles Trade Technical College. I received my Bachelor’s degree from USC. I have a Master’s degree of

Architecture and Building Science from Columbia University in New York. I run a education atelier full-cycle environment in Los Angeles Trade Technical College where we actually have the students become scientists, architects, builders, and they assess their built environments and campuses.

I am going to be talking about two particular issues. Currently, how are we using federal public digital data for our campuses and an idea of how federally supported research and data can catalyze our communities and our neighborhoods.

Los Angeles Community College -- there are nine of them. They earmark $5.7 billion in approved funding from voters. With this money, the LACCD and the nine colleges are constructing state of the art green buildings and improving ecological literacy.

LACCD educates almost eight times as many Latino students and nearly four times as many African American students as all of the universities of California campuses combined. Eighty percent of LACCD students are from underserved populations. The Los Angeles Community College is the largest community college districts in the United States and is one of the largest in the world. It covers an area of 882 square -- radius. They are very apart.

With this challenge, we knew that we needed to have a synergy and we needed to have a flexible structure and system that could help with our synergy between the natural and built environment, the socioeconomic forces, and the natural resources. We tried to think like Gregory Bateson in the book Steps to the Ecology of the Mind. We tried to think of our campuses as ecosystems where we connect social justice, the built environment, and the natural environment.

Ten years ago -- this is the most impressive thing -- utilizing the untapped talent of our communities, we provided one of the largest scientific spatial data models using more than 200 federal and state guidelines and standards in research. It is the only and most comprehensive campus maps and models from this program.

This scientific repository is now helping us to make decisions. It is also combining efforts for first response, natural intelligence, operations and planning, emergency management, Americans with Disabilities Act, and most important, our neighborhood planning.

The program has standardized information from different points. Really, I am going to be placing all of this information on YouTube. All you need to look later on today is public access to federally supported data and you are going to be able to see the completed scientific models where we actually utilized multiple guidelines from multiple agencies, from NISP, the IFC, Industry Foundation Class, OMNI, and others. We have been following very closely data.gov that has grown from 47 datasets to 400,000 from different agencies.

What we did with this data, basically, is that we connected the life cycle of a building, which is design, build, plan, construct, furnish, equip, and then assess. That is the life cycle. Then we connected it with the federal enterprise architecture.

Back in 2005, we studied the federal enterprise architecture, a memorandum that was given by the heads of all executives, of how we can look at performance, procedures, services, business management, data IP, and natural resources. What we did is we took the federal enterprise architecture and we inserted inside of each of the life cycles of a building. This very quick visual model allowed all the participants, more than 400 architects, engineers, construction managers, and most important, all of the interns and all the stakeholders, to look at the data in a single map, in a single area that we could actually look at variables.

What is very magical is that all of a sudden this new economic model started giving us a hope to see how space could become a vessel of memory, a vessel where you can actually look at how decisions were made and how decisions can be improved.

STAFF: Marciela, thank you for your comments. Sorry I had to cut you off. We had exceeded the five minute limit. Next up, we are going back to the room. We have Rebecca Kennison. Whenever you are ready, the microphone is yours.

REBECCA KENNISON: I am glad to see Columbia is so well represented today. My name is Rebecca Kennison. I am the Director of the Center for Digital Research and Scholarship, which is part of the Columbia University Library's Information Services.

CDRS serves the digital, research, and scholarly communication needs of the faculty, students, and staff of Columbia University. Within our portfolio, we manage Columbia's research repository and through our scholarly communication program, we provide a wealth of information on a number of topics, including the opportunities and challenges of data management, data sharing, and data visualization.

Along with our research data manager, Amy Numberger, from whom you have already heard, we work closely with our colleagues, both in the libraries and in the Office of Research Administration, to provide valuable training and education about data management and the research data life cycle. It is within this context that we welcome the opportunity to respond to the White House memorandum on increasing access to the results of federally funded scientific research.

We appreciate, in particular, that the memorandum calls for agency plans to be developed in consultation with all stakeholders, which include universities and their libraries, who share common interests with the federal government in promoting broad public access to and the productive reuse of research data. I would like to present, here, four recommendations to begin -- and I do mean only to begin to achieve this goal.

First, agencies should permit principal investigators to request funding to cover their research data management costs as part of the data management plan requirements that already exist. Only by integrating the cost of storage, data curation, and long-term stewardship of data into the granting process will it be possible to gather enough information to allow for a proper consideration of the genuine cost to share and preserve various data types, information that will be crucial for an evaluation of the full benefits of these data to other researchers and to the public. We urge adoption of this requirement by all agencies.

Second, given the variability of agency funding, we believe the wisest policy is to encourage the growth of existing repositories and the development of new ones that will be managed by academic institutions, consortia, scholarly societies, or a combination of these in partnership with government agencies rather than any individual agency trying to go it alone.

Every researcher funded by a federal agency should be required to submit their datasets to a suitable repository upon completion of the grant in order to help ensure consistency and compliance. Allowing researchers to deposit data in the repository of their institution or a collaboratively run university-sponsored or other federal agency approved data repository and providing a persistent link to that data in reports to the federal agency providing funding would, we believe, permit maximum compliance with minimal confusion.

Third, no matter where the data resides, final peer-reviewed scholarly publications should be linked openly and persistently to their source data to allow for reuse and replication of results. As much as possible, underlying datasets should likewise be linked to the publications that arise from them.

Agencies should require the use of persistent, unique identifiers for datasets in order to facilitate discovery and reuse of data, development of new services, and demonstration of the impact of sharing data in ways that align with existing discipline norms.

Fourth, but perhaps most significantly, we encourage, in fact, urge the involvement of the scholarly and professional societies in the identification and development of domain-specific digital data standards and of the data repositories, themselves. As both liaisons among and representatives for their constituencies, societies are equipped to deal with the inevitable idiosyncrasies of the data in their domain and should be vital partners in the development of any agency policy on research data. Thank you.

STAFF: Thank you. Next up, we have Michael Tanner.

MICHAEL TANNER: Good morning. I am Michael Tanner, Vice President of the Association of Public and Land-grant Universities, speaking on behalf of the AAU, ARL, and APLU. Tuesday, John Vaughn spoke about publications on behalf of our three organizations. Today, I will address public access to data.

The February 22nd White House memorandum on increasing access to the results of federally funded scientific research provides new opportunities for partnership between research institutions and federal agencies. Enhanced access to digital data accelerates the pace of scientific discover, facilitates independent confirmation of scientific results, promotes innovation, and supports education. Open data policies must also be harmonized with existing policies, such as those protecting the privacy of research subjects.

Today, practices and the infrastructure for curation, description, storage, and preservation of digital data are considerably less developed in some disciplines than they are for peer-reviewed journal articles. It will be important that whatever policy or policies federal agencies propose minimizes the cost and complexity to avoid increasing the administrative overhead of compliance with grant requirement for both principle investigators and research administration.

Since resources must be found within existing agency budgets to implement plans, the value of full data preservation and assured ease of access must be balanced against the cost involved, as foreseen in the first sentence of section four of the memorandum.

Coordinated federal agency policies can lead to lower barriers to digital data access, discovery, sharing, and reuse. Similarly, agency policies should ensure that public access to digital data occurs through well managed, sustained preservation archives that enable a legally and policy compliant peer-to-peer model for sharing.

While the OSTP memorandum reflects U.S. national policy, we recognize that research is a global enterprise. The principals, tools, and solutions to accessing scholarly literature and sharing digital data will be global. We are in agreement with OSTP that significant advantages can proceed from the open deposit of selected digital research data. However, much discussion is required to define research data before plans to build upon existing repositories and/or create repositories to receive it can profitably proceed.

The definition provided by OSTP in its directive is data is defined consistent with OMB circular A110 as the digital recorded, factual material commonly accepted in the scientific community as necessary to validate research findings, including datasets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communication with colleagues, or physical objects such as laboratory specimens.

This is a helpful starting point for the discussion, but elaboration of this definition is required. Given the varied nature of experiments and the data gathered in different fields and the metadata necessary to interpret, correctly, data gathered in a range of representations, commonly accepted standards may not yet exist in all areas covered by the directive.

We suggest that a task force be impaneled to give operational meaning to this definition and to address the challenges and resolve the issues OSTP enumerated in subheading A-J of point four in the directive. The membership of the task force should include representatives from research universities, scholarly societies, and federal funding agencies. In the interim, we suggest that every federal agency subject to the OSTP directive develop consistent policies for researchers to include data management plans as part of their research funding proposals as soon as is feasible so that data retention and preservation when appropriate will be universally practiced.

A data task force should remain on call in the future to provide advice on revising the operational definition of research data and other aspects of data management subject to the directive since the dynamic nature of research will surely require continual updating of the definition. Thank you for this opportunity.

STAFF: Thank you so much. Next up, we have Todd Vision. Whenever you are ready, the microphone is yours.

TODD VISION: Good morning. Thank you for the opportunity to speak. I am Associate Professor of Biology at the University of North Carolina at Chapel Hill and the

Associate Director of Informatics at the National Evolutionary Synthesis Center. In full disclosure, I am also on the Board of Dryad Digital Repository and of Orchid. I was able to put one affiliation up in the sheet so I put Dryad. I am a researcher as well as a cyber infrastructure provider.

I want to begin by commending OSTP for recognizing legitimate public interest in the dissemination of research data produced with public funds and the critical and unique role of researcher funder policy in making it happen. My own perspective is that the motivations are manifold. Data should be available for scrutiny to ensure confidence in scientific findings. It should be available for simple reuse to avoid the cost of duplicated data collection and promote methodological refinement. Where feasible, it should be available for more complex repurposing, allowing synthetic research that transcends the data collection efforts of any individual grant.

The relative importance of those may vary depending on the data. Particularly for irreproducible data about natural or historical events unique to time and place and circumstance, the data, itself, will frequently be of more lasting scientific value than the initial publications that derive from it.

The ideal data infrastructure, in which all data of value is accessible and carefully curated with all the context needed for open ended reuse, is clearly not achievable with present day resources or knowledge, but it shouldn't dissuade us from taking pragmatic steps, now, that will greatly improve the state of affairs.

I would like to recommend ten pragmatic steps that I believe could be applied fairly broadly across different agencies and disciplines. Many of these could and probably should also be applied to software developed in the context of publications to process or analyze data.

The first, reiterating what we have heard before, requiring data to be archived in the public repository. I would like to emphasize that that be done at the time of first publication when researchers are best able to provide and document their data. The evidence is clear that data sharing upon request is ineffective and inefficient and archiving at publication and list journal editors and reviewers, et cetera in achieving compliance.

Two, exceptions for sensitive data of various forms could be asked to either meet certain exemption criteria, which could vary by agency or program or could be justified in advance by a data management plan. Implicit in this is that data management plans would be more broadly used across the agencies.

Funders could define what is minimally required of a public repository, including accessibility, preservation, reuse terms, and possibly desired features, as well, such as the identifiers for datasets and researchers that we have heard before. These checklists can then also be used to inform the features of new data infrastructure projects that agencies choose to fund to ensure that they are fit for purpose.

Four, consider allowing researchers to place time limited embargos on the release of their data if it is supported by the repository. This can greatly help overcome researcher resistance to archiving data at the time of publication. Recent studies inform the length of embargos that have the maximum benefit.

Funding agencies can do a number of things to align research incentives with the larger goal, although, they are all relatively minor. Researcher incentives will be aligned when making available a high impact dataset contributes more to success in obtaining future funding than having produced a low impact article.

That is hard to achieve, but one way to promote this is to explicitly allow the evaluation of the reuse of an applicant's past public data products along with their publications. The NSF is to be applauded for its recent decision to allow research products other than traditional publications to be listed in proposer biosketches. Metrics on reuse would also greatly help this.

Give data management plans teeth. That means giving them weight and evaluation and having them listed in -- having data listed in grant reports to promote follow through. Ensure that grant budgets include funds for the execution of the plan. There are very good models in the Welcome Trust in the UK that take onboard the costs of dissemination as integral to doing research. That statement is to be commended.

Promote development of a market for providers of data services. The infrastructure that we need can potentially be developed and sustained, provided that funding organizations allow flexibility in how researchers and institutes cover data management costs out of their grant funds.

Nine, license terms for data should not exclude commercial and derivative products. The primary effect of those restrictions is simply to undercut business models for innovative products that could add value to primary collections.

Lastly, data policy should, itself, be the subject of targeted research so that future decisions can be informed by evidence. With that, I thank you for your time.

STAFF: Thank you so much. Next up, we have Jared Lyle. Whenever you are ready, the microphone is yours.

JARED LYLE: Hi. I represent the Inter-university Consortium for Political and Social Research, which is a research center and data archive at the Institute for Social Research at the University of Michigan. We strongly support the recent OSTP memorandum.

For over 50 years, ICPSR has been providing access to and archiving of research data, as well as supporting the sharing of research. I want to share six things that we have learned and recommendations to pass along to the federal agencies, including that data should be discoverable, meaningful, persistent, trustworthy, confidential, and citable.

As context, I want to show you a punch card from a recent study that we received at ICPSR from a study conducted in the 50s that was stored in filing cabinets for about 40 years, a very influential study of a retired population. While they had done a good job documenting the data, the data were never distributed outside their research team. We, as an archive, have gone and tried to replicate their collection and make that available.

I certainly applaud the original research team's efforts, but also want to recognize the difficulty and challenges of trying to go in and recreate the data and make that available. I want to reemphasize that while agencies are developing public access plans to make research data available that they make data discoverable, making sure that data can be found through search engines, as well as across collections.

In the social sciences, we have metadata or data about date using standards called the Data Documentation Initiatives. Other disciplines have similar standards.

We recommend that data are made meaningful and useful. Access involves not just finding the data, but making sure that you can understand it and interpret it. Data are often messy or can be messy or incomplete or incorrect. What we try to do is curate it or enhance it to make sure there is documentation there and that it matches the data and that it can be used for future users.

Third, we recommend that data are made persistent and that you can continue to access them over time. To do that, you need digital preservation. That is the proactive and on-going management of the data. Especially in electronic format, you need to make sure you are guarding against file loss, file corruption, as well as physical loss, physical corruption.

Fourth, we recommend that data are stored in trustworthy repositories. Data producers need to make sure that the data they provide access to is the same over time and that the repositories that are providing access are to be trusted. They are organizationally, procedurally, and technologically sound as data custodians.

Fifth, we recommend that data be made confidential when applicable. A growing number of studies are sensitive and confidential, as we know. Robust methods are available, including those distributed by the American Statistical Association for treating confidential data. We also know that repositories have systems in place, such as physical and virtual data enclaves for making confidential data available and for protecting the safety.

Six, we recommend that data be made citable. Properly citing data encourages the replication of scientific results, improves research standards, and guarantees persistent reference. We include recommending a minimum that each citation include basic elements such as title, author, data, version, and persistent identifier. There are established citation standards available, including those promoted by Data Cite.

Lastly, we know that providing access and preserving scientific data can be expensive. We are encouraged that the memo allows for the inclusion of appropriate costs, although, we wonder what existing, additional, and new funding tied to proposals will support access and preservation of data. We advocate long-term funding for specialized, long-lived, and sustainable repositories that can mediate between the needs of scientific disciplines and data preservation requirements. Thank you.

STAFF: Thank you so much.

(Break)

Agenda Item: Public Comment

Session 2

DR. BROWN: With that, let's proceed to the afternoon session. The ground rules are the same as before. Each speaker will have five minutes to comment. The time will be monitored by the staff and the lineup will be visible. At least the next speaker is visible on the board so each of you can get ready.

With that, let me welcome this afternoon's first speaker and we can proceed.

FREDERICK DYLLIA: Good afternoon. I am red Dyllia. I am the executive director of the American Institute of Physics. If I look familiar, I also spoke yesterday. I have a broad interest in public access both as a working scientist of most of my career and now for the last six years of head of American Institute of Physics, which is a science publisher.

But I have also served and continue to serve on two major trade associations for scientific publishers: the Professional/Scholarly Publishing Division, the American Association of Publishers, and International Association of Scientific, Technical and Medical Publishers. All three groups, American Institute of Physics, and those associations have a broad interest in helping solve the access problem to data.

Compared to publications is subject of the last two days dealing with access to data is far less controversial. As we have heard, data as a collection of facts are not copyrightable. That particular form of data in a very sophisticated database might be, but the fact that it is not copyrighted sort of lowers the temperature of discussing this subject. That is not to say that it is not complicated. It is just less controversial.

I believe there is fair agreement among all sectors of academia. In our funding agencies, our library's institutional archives, academic institutions, and the scientific publishing community that we all need to work together on the issue putting some order to the data business in terms of persistent identifiers how we allow data to be collected and put in formats for use and reuse and how it is preserved and archived.

And the simple reason that it is more complicated than publications is because a journal article is a well-defined object for the most part, where data sets are not. They can range from very complicated, very large data sets from big science, that scenario. I am familiar with being a physicist. I think there are agencies in this room that can be congratulated for the work that they are doing as part of energy to preserve high energy and nuclear physics data sets and NASA, NSF, and astronomy data sets and NIH, NSF, Department of Agriculture and biology data sets.

The larger data sets are well ahead of the other end of the spectrum where we saw examples this morning of data sitting in a little notebook that may leave with a graduate student or on a thumb drive leaving the same way.

I refer the audience to a very useful report that was published a little over a year ago called the Opportunities for Data Exchange, ODE. It is an EU-funded report. It has a very nice mnemonic in their compliments of Jim Gray called the data pyramid. It is a hierarchy of what we consider the most valuable data in the world, which with the peer review data that ends up in scientific journals down to that lower level of raw, uncatalogued, and often lost data.

As publishers were prepared to help the community work on all of those types of data, particularly starting with the top of that pyramid with the data associated with peer-reviewed publications. I make note of an ongoing and

National Science Foundation-funded study with the American Astronomical Society and the American Institute of Physics where we are polling authors of several journals about their attitudes for collecting data and attaching it to publications. It is a good start and one of many we are interacting the peer review data with the journal.

There are some publishers who publish data journals. We are one of them. We have a joint project with NIST on physics and chemistry data. Publishers stand ready to help the community to work on these issues.

I just issue one caution. Since it is so complicated, let's move deliberately and cautiously because we put a lot of requirements on our researchers. We do not want to add a burden. We want to remove burdens.

Thank you very much.

STAFF: Thank you so much. Up next, we have Julie.

JULIE MCCLURE: Good afternoon. My name is Julie McClure and I work in the Science Policy Office for the American Society of Agronomy, the Crop Science Society of America, and the Soil Science Society of America.

First, I would like to thank the sponsors for the opportunity to participate in this forum and provide comments on this very critical issue. The Agronomy, Crop, and Soil Science Societies are the premier international scientific societies focused on food, agriculture and natural resources research.

The societies meet members' needs through publications, recognition and awards, placement services, certification programs, meetings and student activities. Our members obtain research funding from an array of federal agencies including USDA, NSF, DOE, USGS, NASA, EPA, as well as support from corporate partners.

In 2012, the Agronomy, Crop, and Soil Societies launched a digital library that now holds all of our publications including 9 journals, 320 books, extension and teaching resources, including guides and digital media, newsletters, and other general content.

One of our strategic goals for the digital library is to publish data sets in two forms: data that is directly linked to specific journal articles and data sets that have exceptional scientific and societal value that are independent of individual journal articles.

Philosophically, the Agronomy, Crop, and Soil Science Societies favor full and open sharing of data for research and educational purposes. However, we see publishing data as an additional way of adding value to the digital library for our members and institutions as well as a source of revenue through which we can provide other services to our members.

We are in the process of developing the policies and procedures to implement a data archiving and retrieval programming within our digital library.

The following are a few points that our societies are looking at in reference to data management. We believe that the Agronomy, Crop, and Soil Science Societies along with other professional societies may have a role in helping to develop profession-specific data standards. Like many fields of science, research activities in the agronomic sciences are dispersed and few standards are formalized. The societies hope to play a key role in determining standards for minimum data sets.

We also see our societies as having an educational role assisting in creating and implementing best management practices for data in agronomic and agricultural sciences. Many principal investigators have mixed feelings about publishing their data.

There is a sense of ownership that can be discouraging that can discourage investigators for making their data widely available. They worry about the perceived threat of data misuse and during reuse. We believe that change to the scholarly credit system for publishing data could help incentivize this process. We also believe that we should provide tools to make the submission process as seamless as possible.

Data from the agronomic and agricultural sciences will be far more heterogeneous both temporally and spatially than much of the data currently being archived in databases like GenBank. Accurate and complete metadata will be key to understanding and reusing these data.

In many research programs, graduate education is a key part of the mission. And the students collect the vast majority of the data. However, like faculty, graduate students have vastly different data habits. Our societies may have an educational role in creating and implementing practices and programs that foster good data habits a priori among students.

On federal agency integration and cooperation, it is critical that processes and procedures for data be seamlessly integrated over all federal agencies in order to have consistency and coherence. This could be a potential issue for authors and publishers who receive funding for multiple federal agencies that have different processes to deposit, archive, and embargo data. How will this be managed?

Some of the research done by our members is funded by nonfederal sources, corporations, foundations, NGOs, data policies that recognize this diversity in funding and with its sponsor policies regarding data ownership and publication are critical to continue success of these important private public partnerships.

A realistic timeline for data program implementation is key to long-term success of this effort. Giving agencies only six months to come up with a plan may result in a haphazard, poorly thought out policies that fall short of our shared vision for this important program.

I would like to thank you for your time and opportunity to comment on this issue.

STAFF: Thank you so much. Up next, we have John Downing who will be coming to us via the WebEx here. Just give me one second. We have you up there. Are you planning to do video?

JOHN DOWNING: I thought not. I am casually dressed today. Thanks.

STAFF: Whenever you are ready, the microphone is yours.

JOHN DOWNING: Thank you so much. I will be making just a few comments on using, creating, and publishing data that I hope could be considered in building plans to comply with the OSTP memo.

For some background, I am a scientist at (?)13:17 University in the United States where I use and share large amounts of data from all the world. I am also president of the Association for the Sciences of Limnology and Oceanography, a publisher of scientific journals. The opinions I will express or the ideas I put forward are really my own, however.

I have been a user and archiver both of data for 30 years. I am extremely supportive of a concept of data sharing. I perform global analyses of water resources that have combined data to create a global vision that is very important to issues of climate change and eutrophication and a variety of other things. Therefore, efficient data sharing is extremely important to me and others in the earth and aquatic sciences.

I have published numerous publications on secondary and meta-analyses and these have been amongst my most popular and useful publications. I also have been an archiver of public data on water quality collected through our certified water quality laboratory.

More recently, I am involved in an NSF macrosystem's ecology project where we concatenating data on something like 30,000 lakes worth of information, which is a very large job in which the challenges of data archiving have become quite obvious.

I began my career in Canada and at that time was encouraged politely to use a thing called CISTI, which is an early federally-funded data archive for federally-funded research results. This stands for the Canada Institute for Scientific and Technical Information. It has been in existence for about 30 years or so.

In using it and other data storage systems, I have encountered as a user several barriers to data-to-data use. One is determining what the relevant data are. And the most useful that I have found has been a reference in a publication that needs to be some sort of stable link. In the past, these have basically evaporated as institutions have waxed and waned.

Also determining what the data mean is a challenge. Metadata are very important, but they are often hard to find, harder to understand, and even harder to write and often mean nothing a decade later. Standards of metadata are very important.

Actually locating data is quite important and actually being able to use them. What format is used and keeping those constant and current is also important and critical. Oddly, the most useful has been paper tables because of changers and computer formats. There is obvious maintenance that needs to be done.

Also, questions of whether the sharer of data can afford to put the work in to make the data usable and available. Even building a single database to send to someone as cost. I would like to recommend as a data user that stable URLs are federally maintained databases be referenced and publications and the data management be well and securely managed.

I also have had concerns as a data gatherer and an archiver. I ask what reasonable period there is for exclusive use. I wrap up a lot of intellectual property and the decision of what data to collect and how to make the measurements. Failing to honor that, I think could squelch innovation. Some embargo period would be useful and deciding how to do that is an important consideration. Also, how we square data sharing with the Bayh-Dole Act that permits a university, a small business, or a nonprofit institution to elect to pursue ownership of intellectual property arising from federal funding and preference to the government is important. I would recommend that a suitable embargo period be determined to allow scientists to fully develop the data that they have struggled to gather.

As a president of a scientific society with publications, it will need to link to data sources. I have some publisher's concerns. It seems that the data will either need to be linked to the article or included in some kind of a database with the article.

If they are included, then we revert to the same issues of open access that we treated over the last few days as we spoke of open access publications. There are major costs involved in reviewing and preparing data for publication and free access is really not free and recommendations need to be made about who will be responsible for those costs.

If the data are simply linked then there needs to be a stable URL to put into the publications that will last for decades, not just one. Things that are here after one decade will be gone after two and so on. The stability is almost unimaginable given the fast pace. The need for stability is unimaginable given the fast pace of change and technology.

I would recommend that data archiving not be relegated to publications, but be well funded and centralized. Funding needs to be continuous. A down to earth example is if you bring food home from the grocery store, you need to pay the cost to keep the refrigerator running or you lose the groceries.

I would ask that perhaps some of these --

STAFF: Hi John. Thanks for your comments. Sorry I had to cut you off there, but we reached your five-minute limit. We greatly appreciate you calling in.

Next up, we have Nathan Bell. Just give me one second to queue him up here.

STAFF: Nathan Bell is substituting for Felice Levine.

NATHAN BELL: Good afternoon. I am Nathan Bell, the associate director of Education Research and Research

Policy at the American Educational Research Association. AERA is a national scientific association of 25,000 members dedicated to advancing knowledge about education, encouraging scholarly inquiry related to education, and promoting the use of research to serve the public good.

AERA applauds both the principles and the objectives for public access to scientific data in digital formats. The AERA code of ethics mandates data sharing and acknowledgment of data use and allows for data use under restricted access provisions when necessary to protect privacy and confidentiality.

Authors in AERA journals and education researchers more generally are expected to make accessible data related to their publications.

For more than 20 years, AERA under its grants program funded by the National Science Foundation has fostered use of federally supported data sets. This long-term project has led to important scientific discoveries and methodological advances and has contributed to building scientific knowledge cumulatively through analyses of such data.

In collaboration with the Inter-university Consortium for Political and Social Research, ICPSR, AERA is also promoting data sharing and respectful, responsible use of data sets.

We urge OSTP and related agencies to develop macro-level plans that not only require data management and sharing from grantees, but also more broadly take steps and allocate resources to foster and facilitate a culture of data sharing and use. Knowledge about data access, the range of data amenable to sharing and mechanisms for providing access varies within and across federal agencies and within and across fields of science to ensure not just more policy on the books, but more meaningful incorporation of policy and action requires implementation steps that can deepen and widen appreciation of the scientific value of data sharing, access, and use.

We offer the following comments to facilitate that end. One, federal policy for data sharing should include access to digital data that encompass voice and video data or other forms of big data harvested from diverse sources and preserved in digital form. Data sharing should also include the sharing of data collection instruments such as interview protocols, measures, coding guides, and manuals.

Two, data management and sharing plans already required by agencies like NSF are essential. Funds should be provided in awards to support archiving and data repositories to maximize data standards, access, and preservation.

Three, to maximize meaningful access and contain costs, agencies should require use of data archives as a default and investigator or institution provided access as the exception. Agencies might offer a certified list of data archives with appropriate capacity and expertise.

Four, funds need to be provided to support repositories to expand their capacity to make accessible an expanded body of federally-funded data as well as to prepare for a larger, wider number of users and innovative mechanisms of access and use. Fields of science with no or only limited repositories may need to support to launch such entities.

Five, educational materials, webinars, or courses should be supported by science agencies particularly in partnership with scientific societies to provide deeper knowledge about data sharing and the value and use of third-party data archives like ICPSR. Emphasis should be placed on data sharing and principles of sound use.

Six, accessible guidance on data sharing and alignment with consent, privacy protection, and data confidentiality would be valuable. Knowledge, expertise, and views about data sharing vary widely among investigators, institutions, and institutional review boards and limit or inhibit data sharing and use. An entity like the NRC might prepare a general guide for data sharing for federally funded research.

Finally, OSTP, federal agencies, and the office for human research protections should develop a statement to foster responsible sharing of identifiable as well as linked data as long as scientists use such data under restricted conditions and as long as they are legally bound to honor consent agreements and face stringent penalties for disclosure. The NRC, federal agencies, data repositories, or scientific societies could assist in this task.

In conclusion, AERA urges attention to these issues and where necessary to the investment of cost effective funds that can reap major scientific benefits.

Thank you.

STAFF: Thank you so much. We appreciate your comments.

Up next, over WebEx we have Andrea Schneider. We have you on the screen here. Let me just make sure you are unmuted. It should work. Andrea, whenever you are ready, give it a try. We are not hearing you right now. I am not sure you are actually connected over the phone. Did you dial into the -- you might have to redial in because we are not hearing any audio from you. Do you have a microphone that goes to your computer rather than the phone? We are not hearing anything from you. What we will do is I will send you an email with the call in information and we will come back to you after Jeanne Holm, who is our last scheduled presenter. Sorry about the inconvenience.

We are going to move to our next caller who is Chris Moore-Barbosa. Chris, are you there? Chris does not seem to be on the line.

We have one more person calling in who is Gene Public. Gene, are you on the line?

Moving forward, I do not see Martha on the line either. Martha, are you out there? No Martha either.

We are going to move back to our in-person presenters. I am hoping Jeanne Holm is here. If you are here, the microphone is yours. Sorry about getting to you so much more quickly than expected.

JEANNE HOLM: No worries. Hi. I am Jeanne Holm and I am the evangelist for data.gov, which is the US government open data initiative. You can find over 400,000 data sets on data.gov thanks to the work of many of those government folks here and contractors at various agencies. 180 agencies and sub-agencies are participating in data.gov today.

I wanted to make sure that people were aware that as a resource both for researchers and analysts to come to find information. The information on data.gov ranges vary wildly from genetics and genomics data on agriculture, which has been a particular focus recently, public safety data, information related to laws and findings at different agencies to information related to energy.

We also have a new community on data.gov that might be of interest called research.data.gov. Mike Stebbins is our chair of that community. It brings up over 900 research-oriented data sets. NITRD was helpful in pulling that information together in the group with George Strawn.

Research.data.gov is a go to place where you can find context, apps, tools and resources for researchers who are looking at using all kinds of government data that has been made available and accessible.

All of the data on data.gov is validated and vetted by the agency. There is a variety of ways to give feedback through social media, a variety of developer tools, and other ways for people to both contribute back to the data as well as be able to comment on the data and suggest new data sets. If there is data that we are missing, we are welcome to hear from you as well.

Particularly of interest to this group is anybody who is working with a federal agency is welcome to publish that data to data.gov. We reach out to a wide variety of national partners at cities, counties, and states who are sharing local information sometimes as research focus, sometimes it is local information about populations or crime or safety.

We also work with a wide variety of international partners. There are a lot of other places across the globe that are sharing their data and federating and linking that data back into data.gov.

We open sourced the code working with the governments of India, Canada, and Ghana as a way of making it available to anybody, not only to improve the source code on data.gov, but also to mean that if you wanted to manage your own data sets internal to your organization with your city or university or just a small company that you can also use that source code.

It is a resource available to everybody. It is free of charge. All of the data is made fully accessible. Our data set pages allows us to be able to link back to professional journal articles that are using that data, to apps that are using that data, to businesses that are using that. We provide a whole set of resources for any agency or any other organization with your university, a national lab, or an independent researcher working with federally funded research data.

Thanks.

STAFF: At this point, we are going to go back to Andrea, who I believe is on the phone now. Andrea, are you there?

ANDREA SCHNEIDER: Can you hear me?

STAFF: Yes, we hear you loud and clear. You can go whenever you are ready.

ANDREA SCHNEIDER: Hi. I would like to introduce myself. I am Andrea Schneider and I just want to give you a frame. I am a political scientist working in Silicon Valley and I am talking to you from a garage. I just wanted to read the rules of the garage, which is we believe you can change the world, work quickly, keep the tools unlocked, work whenever, know when to work alone and when to work together, share tools, ideas, trust your colleagues, and it goes on.

I want to expand the idea a little bit about open data and how people like myself who are applied researchers can take information from research and translate that research into projects and program designs that are actionable and that our projects then become integrated into local and state programs so that there is a continuous flow between the researchers and research that is interesting and important that can back up policy that can back up practice and then can get designed and redesigned for real communities in real places and community settings.

Important to me is that I have done a lot of work where it really depended on research to back it up. I will give you an example. When I did a major set of projects on high-risk use and I worked with the University of Washington and others who were aggregating data on high-risk use that could be used to design projects and programs. The data I was interested and it was not just statistics, but was also in what they learning, how it was coming from different disciplines, how that could then get the utilized. For example, if reading were important for all third graders, how would we integrate that into a program design that would be able to help kids, that would prevent problems later on.

I want to say that I think that a lot of the open data initiative right now in general is very tied to the use of technology. It is a little confusing for me to understand what we mean. I think we need to define by public. Who are we talking about and are we talking about multiple markets? Are we talking about the community program? Are we talking about philanthropy? Are we talking about other scientists? Are we talking about universities? Also, what do we mean by data?

A lot of the projects that get funded could be equally important because they are stories, there are narratives. The context is very important. The process, the design, the strategy, the theory of change, and things that were other than just hard quantitative data that would help us do a much better job if we knew how to access it.

I want to give another example. I was assigned to evaluate under the Clinton administration 96 community policing agency grants from the Department of Justice. I can tell you even though that was ten years ago that the information that came out of that would be incredibly useful to police departments today that are also looking at community policing models. In fact, Commissioner Davis was one of our leaders after the Boston bombing. You can tell that they used the community really well.

How do we get that information out to people who are suddenly discovering community policing might be important? How can we build on programs and projects and science that we have previously paid for so that we are not being redundant or duplicating our efforts and we are leveraging the funds we have already spent and ideally getting to spend the dollar once?

I agree with people who have said that we need to have any tools that have been developed as well, if there is training that has been involved, et cetera. And I also see this as an opportunity to make available to challenge the federal agencies to meet the other part of the open government directive, which talked about collaboration between agencies and that it can be really confusing if you have multiple funding agencies funding similar ideas, but nobody has any real idea about how to find it. It makes it hard for us to build on it even if the research is sound and even if we can find it.

I think we should suggest some kind of strategy for developing a federal infrastructure that would help us be smart and strategic in the collection of analysis and drive the transformation of public services. And how would we use this new executive order to connect our resources and reduce redundancy and duplication and increase our effect --

STAFF: Hi, Andrea. Sorry. We have reached your five minutes there so I had to mute you. Sorry about that. We greatly appreciate your comments.

We will now be moving to actually the open session. We have a participant on the phone who is interested in speaking. I believe Eric Kansa is on the line. Eric, are you there?

ERIC KANSA: Yes. Hi, how are you?

STAFF: Great. Thanks. Whenever you are ready the microphone is yours.

ERIC KANSA: Great. Thank you so much. Thanks for this opportunity to comment on this new and important policy development.

My name is Eric Kansas. I manage and direct a program called Open Context, which is an open access, open licensed data publishing venue for archeology and related fields. I have also worked on text mining initiatives and the digital humanities. I want to really sort of highlight the issue that text mining shows that distinctions we have between data and text or publication are somewhat artificial and that is increasingly the case as these analytic techniques become much more popular. Many of the requirements we have around open data and interoperability with the respect to open IT and data are also going to be increasingly important for research with regard to text.

I want to put the thrust of my comments on this issue with data thinking about an information ecosystem that I think needs to be cultivated around how we manage data in the research community.

With open context, we focus on the editorial and peer reviewed services on data sets contributed to us. We work closely with colleagues at the University of California, California Digital Library, an institution that provides us with essential digital repository and persistent identity services.

We are grateful for grant support from the National Endowment for the Humanities, especially the Office of Digital Humanities, but also from the National Science Foundation and from private foundations. I think this is a good example of how the lines between the humanities and the sciences are increasingly blurred and that is a very good thing.

But it also shows that there are many efforts that are receiving support from multiple federal agencies. I think coordination across these agencies is really vital because research would suffer if we stove piped it into artificial silos.

One of the issues to think about too is that there are many agencies that are involved in research, but they are doing it through enforcement of laws and regulations. I am especially thinking about historical preservation or environmental protection. And the data practices relating to those efforts need to compliment the data practices that are coming from agencies that support mainly academic-oriented kinds of research.

The other issue about this is that data needs are really constantly evolving. We need to really encourage that dynamism by welcoming new -- and new ideas in approaches to data management. There is often a tacit assumption that data are a residue of research and that a researcher's primary responsibility with respect to data centers mainly in preservation.

I think that is really limiting and in some circumstances data can and should be valued as a primary outcome of research. And to borrow a phrase from my colleagues at the California Digital Library, data can be a first class citizen of scholarly production and can play a central role in new modes of scholarly communications. Here, there are lots of different approaches in innovation including thinking of data sharing as publication or as exhibition or even data sharing as an open source reuse and release cycle.

The point is that there are many and expanding roles for data in research and communications and policy should not assume that data is only going to play the role of a secondary supplement outcome.

The last thing I want to address with this issue about dynamism is it also relates to -- it needs to inform when thinking about financial sustainability. Public policy needs to recognize that sustainability of particular organizations and practices in the research endeavor is only a means to an end in promoting the public good. Sustainability of particular interest should not be an end to itself.

I think that resiliency might be a better term here since it might better capture obligations for data and knowledge stewardship without locking to a particular set of institutions or practices. In other words, notions of data openness need to expand beyond the technical and licensing concerns, but also to organizations, people and communities that are practicing in the information ecosystem. Especially the next generation of students will have their own needs and priorities with respect to data.

True resiliency will require real funding. This is an issue that the OSTP policy memo falls somewhat short. I urge agencies to work with the research community, libraries, and others to honestly understand funding requirements. We need this to make a better and clearer case to the American public about investing and unlocking the richness of research data.

I appreciate the opportunity and thank you so much for letting me make some comments here. Thanks.

STAFF: Thank you so much.

Up next, we actually have a walk in here live. We will go back to the lectern for our next speaker. Please remember to state who you are for the official record and if you are speaking on any one or any organization. Thank you so much.

JOE HOURCLE: I am Joe Hourcle. I work for the Solar Data Analysis Center and the Virtual Solar Observatory, but I am here speaking for myself. I did submit a written statement, but I did not think I would need to speak.

There have been a lot of folks talking about that we need to look into certain things and we need to look in other things. Data citations were mentioned a lot of times. I am not sure how many people are aware that the National Academy of Sciences, Board of Research Data and Information actually ran a meeting about a year and a half ago on developing data attributions and citation practices and standards. I know a few of you were at that meeting. Anyone actually read the report who wasn't at the meeting? Not a whole lot.

Part of the problem is that a lot of the communities are working on things. I think they would be applicable to other fields. But the information that it happened just is not getting to the other communities. We have a lot of work being spent on big data, but it is not trickling down to the folks that are doing mid-sized data. I do not know how much duplication of effort is being done by the different agencies of everybody looking into citation when maybe you can piggyback on the work that has already been done rather than reinventing the wheel and all that.

I know the UK has just in the DCC, which is Digital Curation Centre, to sort of help centralize some of that and spread the information about what is going on. I do not know that we really have anything. We have data.gov for the actual letting people know what data sets are out there, what the good practices are. I do not know that we have that sort of thing.

There has been talk in ESIP, the Earth Science Information Partners, which I think Carol is going to be talking for them tomorrow about maybe it is time to look into decadal survey to look at data practices. What is out there? What has already been done? What fundamental questions still need to be answered? Who can fund it? And where do we really need to concentrate our effort?

Thank you.

STAFF: Thank you so much. If there is nobody else here in the room that is interested in making a comment. We do not have anyone else on the phone.

PARTICIPANT: I thank all of the speakers this afternoon. So far, we have arrived at a pause in our program. The program calls for us to reconvene at 3:30 in case there are any walk-in registrants who wish to speak. That is not clear there will be. I wanted to let all of you know. Anyone who does enroll as a walk-in registrant between now and then will be welcome to speak and will be filmed and viewed by whoever is here.

I wanted to just remind you. We will begin the session at 3:30 and then adjourn if there are no further speakers in order to keep to our contracted and advertised plan for public comment.

I want to remind you all that we will have another session tomorrow morning beginning at 9 o'clock. We do have a list of commentators who have expressed interest in commentating in making comments.

We will have a session of public comment tomorrow morning followed by according to the plan followed by a report from the rapporteur who will try to summarize and organize some of the comments you have heard. Tomorrow morning at 9 o'clock you are welcome to return. If you keep your badge from today's session, you can come into the building tomorrow without having to re-show your ID and check in. I hope to see anyone who wishes to stay until 3:30 and I hope to see many of you tomorrow morning. Thank you.

(Session adjourns)

4National Academy of Sciences

National Research Council

Public Comment Meeting

Concerning

Public Access to Federally Supported R&D Publications

Public Commenters

May 17, 2013

National Academy of Sciences
2101 Constitution Avenue, NW
Washington, D.C.

Proceedings by:
CASET Associates, Ltd.
Fairfax, Virginia 22030
caset@caset.net

Table of Contents

P R O C E E D I N G S

Agenda Item: Public Comment

Session 4

CAROL MEYER: Good morning. My name is Carol Meyer and I am the executive director of The Foundation for Earth Science. The Foundation provides management services to ESIP Federation and 150 plus member organization that fosters collaboration across diverse earth and environmental science interest in a neutral non-political setting.

Our breadth and depth among data and technology practitioners is wide, crossing federal agencies, academia, commercial, and non-governmental organizations, as well as the science domains. Our community has been at the forefront of developing consensus on a variety of data management issues including the use of DOIs, data citation, and creating a reusable data management training – series of training assets.

The ESIP Federation was born out of an NRC recommendation to NASA, to form a community stakeholder organization that advise the agency on the evolution of EOSDIS, its satellite data and information system.

I would like to talk today about the value of science data and how we might maximize it. Imagine a future where data not only supports the science for which it was collected, but could be utilized in everyday decision making. We already see this happening in the application of weather data in disaster management settings and in the application of translational medicine.

How can those advances be extended to other disciplines? Whether the focus of science is for discovery, for problem solving, or for behavioral change, the value of science information increases when it informs many context beyond its original intent.

A diverse group of ESIP Federation partners has been meeting since January to discuss how our community might move ahead in light of the president's recent initiatives to open up government data. We applaud these efforts and are willing partners to help evolve policy and practice to maximize the benefit of our public investments in science.

Much progress has already been made through agency contributions to data.gov, international cooperation through the group On Earth Observations, and on-going interagency activities that leverage capabilities across the government. But I ask is it enough?

Our community has witnesses and even helped lead the evolution data management practices during the past 15 years, and in thinking about long term data management challenges created by an increasing volume and variety of data, and the demands for increasing data integration across traditional science disciplines, the Academy can help science agencies evolve a short, mid, and long-term coordinated vision to help realize the full potential of science, data and information.

Building on the administration's Earth Observation assessment, its recent OSTP memo on opening research results, and the recently released open data policy, opportunities abound for putting science data to work. The full potential cannot be realized unless there is a coordinated effort among federal agencies and there non-federal partners.

The administrations Big Earth Data Initiative is a bold next step to promote such coordination. And the Academy's decadal survey process could be an effective, longer term approach to align science data outputs with the growing interest in their use. Not only by scientists but by non-traditional users, as well.

We urge the Academy to think big. That is look beyond existing practices and envision a future that enables science data and information to be impactful to research, decision making, and public behavior.

We are working on a formal recommendation that would encourage the Academy to study the necessary framework and investments needed to achieve this grand vision.

The US continues to be a leader in science and technology development, and the opportunity before calls for bold steps that support increasing demands on data and information. Our community is excited by the prospect of greater collaboration across government and beyond, and that would allow the full potential of data usage to be realized.

Thank you.

ALEXANDER EVANS: I am here to speak out figshare. Our founder Mark Hahnel could not be here unfortunately. I would like to first introduce you to him before I introduce myself. I am going to read a little bit from a press release from Digital Science, just to be sure I am citing the right sources.

Mark began work on figshare while he was PhD student at Imperial College. Frustrated with the duplication and waste in research due to inadequate data openness and visibility, figshare allows researchers to publish their data in a citable, searchable, and shareable manner. The data can come in the form of individual figures, datasets, or video files. Users are encouraged to share their negative data and unpublished results, too.

All data is persistently stored online under the most liberal, Creative Commons license, waving copyright where possible. This allow scientists to access and share the information from anywhere in the world with minimal friction.

So to introduce myself, I am the director of a non-profit, Alvb Limited. What we do is work with artists and scientists to bring them together to share their data and create art. We create aesthetic data visualizations. For several years we have been trying to find a platform that is easy to use, because we are not computer scientists we are artists, and something where the data can be made readily available and stored securely.

We began working with figshare not long ago, to introduce figshare, it is a cloud based repository in data management system. They are working with amazon web services. Data is stored online under a CC license. Even with the CC license, the author retains copyright on all outputs made openly through figshare. CC licensing allows people to build out of or on top of previous research. It is private and secure with one gigabyte of free storage. Or if you are ready to make your data public, there is unlimited free space. As I said, it is secure. It is cloud based, is scalable, and also it makes your research outputs citable, shareable, and discoverable.

They are working in partnership with a group called OCID, to create digital identifiers for researchers.

With cloud service it is available for global access. Very harvestable, discoverable datasets, and as I said before, great capacity for scalability.

In the context of sharing you can make your output citable with the DOI, with the OCID ID, as I mentioned. You can visualize your academic formats in the browser in any format, including machine readable data. You can track the metrics of your research. You have access to dynamic embeddable badges to illustrate the impact of your research. Also, within the platforms, there are embeddable widgets to display your outputs on your lab website or your blog.

As I said before your storage is private for one gigabyte and unlimited is free. It is accessible from anywhere, easily filtered on elements such as tags, files types. It allows for collaborative workspaces, and also has a new feature of a desktop uploader with multiple formats of PC, Mac, and Linux. On board and IPI to automate your research management.

So if anyone would like any more specific information please contact Mark at figshare and he will create what you need.

Thank you.

BRIAN ATHEY: Hi, I am Brian Athey. I am founder and CEO of the tranSMART Foundation. tranSMART is an open source, cloud based platform to enable biomedical research and sharing across the public and private sector. The tranSMART Foundation has been created to sustain this effort and think like the Linux Foundation. We maintain the code base and work with partners and organize the community across the world.

The tranSMART Foundation believes that data in the governance sector, data created with government funding, with sponsored research, and data created for the public good in the private sector should be shared, should be made liquid and available for analytics and for knowledge creation.

We are part of the open data and open science movement here in this country and abroad. We are welcoming the opportunity to work with the OSTP guidance and the federal government here in the United States, as we are in Europe, to enable data liquidity and sharing in the biomedical research sector. This is especially between pharmaceutical companies, amongst themselves, biotechs, and between biotechs and the academic sector, which is really an emerging area for R&D in the future.

And will potentially avoid hiding of results that are not available to those of us relating to proprietary concerns, and will allow us to actually accelerate new discoveries and translate these discoveries into use in the biomedical sector.

We are deeply embedded and engaged in a public/private partnership model. It means that for things like data life cycle management, we understand that the government investment is not enough to sustain this. That the private investment is not enough to sustain this, so we are really looking for a partnership. In this country, for a government investiture in these kind of efforts.

We are really looking to lower the barriers for translational research for knowledge creation, and we currently are being used by several major projects in Europe, which is helping us to continue the development of software platform. There are two in particular, one funded by the Innovation Medicine's Initiative, which is a public/private partnership which is an umbrella that enables translational research. It is a two billion euro investment, a billion euros from the private sector, a billion from public sector, the European Union. And underneath this to support 35 clinical and translational research projects, is the tranSMART platform, which is funded by 25 million euros of support. It has just really started over the last six months

We in the foundation are working with the folks at the ETRIX(?) Consortium to make sure that the code base continues to evolve and is useful globally. We have over

20, 25 installations now at tranSMART, in pharma, in academics, in not-for-profits, and the list is growing.

There is a translational information technology initiative in the Netherlands that is funded with 16 million euros of support, which is embraced and is using the tranSMART platform.

Here in the United states the actual tranSMART code base itself, was created through contributions from the NIH National Biomedical Computing Centers effort, I2B2, the NCIBI, and NCBO, have contributed code to this. We are hopeful that the NIH NCATS, FDA, which is already evaluating tranSMART, and other government agencies will contribute to the public part of the public/private partnership in the United States, to balance that that is happening in Europe.

TranSMART Foundation is excited about the upcoming open science champions of change initiative that is coming from the White House. One can nominate folks like us who are contributing to the public/private partnership on open science and open data sharing and liquidity efforts. The nominations can be made at the whitehouse.gov website and are available to us until May 23rd, for a June 20th event.

I know that we at the tranSMART Foundation are hopeful to announce our code release 1.1, our Amsterdam meeting of the foundation is on the 17th to the 19th of June. So the timing is good for that event and we hope that others will participate as well.

Thank you.

STAFF: Say, next we have Timothy Vollmer. Just give me one second Timothy. If you can hear me, can you make sure your phone is on line or unmuted? I need to unmute it here. When you're ready, you can start at any time.

TIMOTHY VOLLMER: Can you hear me okay? I'll go ahead then. Thank you very much for having me on the program today. My name is Timothy Vollmer, and I work at Creative Commons. Creative Commons is a non-profit. We're headquartered in the US, but we operate all around the world. We're ten years old now. Over those ten years, we've been developing three copyright licenses that allow creators to share their work on more open terms than what's called the default "all rights reserved". We like to call our approach "some rights reserved".

CC licenses are used around the world really by anyone who wants to share any sort of creative work, whether it be musicians or photographers or scientists, authors, and even government bodies are using it. You might have noticed creative commons licenses in use on big websites like Flickr and Wikipedia, even WhiteHouse.gov.

The executive order on open data that was released last week even recommends that Creative Commons licenses can be used as an option for sharing some of this federal public sector data.

We're really glad to see that the Obama administration is supporting the principle that the public should have free access to publicly funded data. We think this makes sense. We think that federal agencies can take the next logical step by removing permission barriers as well as those price barriers. We believe that the Creative Commons licenses, and even more appropriately CC0, which is our public domain dedication instrument, can really help the agencies meet their requirements set out by the directive.

As you build your agency's public access plans in the next few months, we urge you to consider supporting those that wish to release their data as open access. I'd like to touch briefly on two points that we think are important for you to consider. First, since the goal of the directive itself is enabling broader use of publicly funded research data, we think that it's important that agencies make it clear to the re-users of that data the rights that are available to them. From a legal perspective, you can do this by permitting researchers to deposit their data directly into the public domain, using it to select the CC0 public domain dedication.

Now of course, in cases where making data available on a public domain is not suitable or applicable, you could allow for the data to be marked with a liberal license, such as a Creative Commons attribution license. This license allows full use and redistribution of the data only with the obligation that the user credit the data creator. Second, you can require researchers to deposit data in a scientific repository such as Dryad, figshare, or DataONE, or other similar storehouses that allow these researchers to easily upload and make their data available in the public domain.

In conclusion, we believe that the communication of a clear, unambiguous rights to federally funded research data, combined with depositing such data in repositories that make it easy for others to access and use will really increase the utility of this data for science and help you meet the overarching goals of the directive itself.

Creative Commons is standing by to offer any assistance, and we're happy to talk or email with any of you. I also note that we've submitted a written statement for the record. Thanks again for your time, and I look forward to the rest of the meeting today.

STAFF: Thank you so much. Up next we have Jonathan Markow.

JONATHAN MARKOW: Good morning. My name is Jonathan Markow, and I’m from the DuraSpace Organization. We are an independent, not for profit organization that’s committed, as you are, to a shared digital future. We collaborate with academic research, cultural, government, and technology communities by supporting open source repository structures, and in turn help knowledge communities ensure that current and future generations will have access to our digital heritage.

We also provide hosted services that include DuraCloud and archiving and preservation services and dSPACE direct, a turnkey repository solution that allows organizations to archive and preserve content with minimal maintenance. We have a dual role out of being stewards for the open source repository applications and supporters of their communities and a provider of archiving and preservation services.

We provide leadership guidance and infrastructure to encourage community development of dSPACE and Fedora open source repository systems that are used by over 1,500 institutions worldwide for disseminating and preserving digital content. Digital resources managed in dSPACE or Fedora repositories include theses, dissertations, research data sets, audio and visual files, many types of imagery, and more. These institutions include federal government organizations such as the Smithsonian Institution, the National Libraries of Medicine, the National Aeronautics and Space Administration, the Food and Drug Administration, the Department of Agriculture, and others.

DuraSpace supports the initiative to promote open access, dissemination, and long term preservation of publicly funded research. Our message today is that we strongly recommend that any technology solutions deployed for this initiative be based on open source software applications, which have a number of advantages relevant to our current needs.

For one thing, the start-up costs are much lower than proprietary software because the licensing expenses are non-existent. Open source software comes with freely available source code as well, and is supported by active and engaged communities of practice as well as by many commercial service providers. You have a choice. You can support it yourself, or you can pay to have it supported.

Government agencies and departments deploying open source applications like dSPACE and Fedora are able to join a global community of developers to add or change features to meet specific requirements. Changes may be contributed back to the community so that others can take advantage of them and help maintain them, or they may simply use the software without any obligation to write program code themselves.

Finally, open source software is most often based on open standards, which facilitate interoperability with other applications that adhere to standards. Most importantly, users of open source software may invest in its use without any fear that changes to proprietary code will someday stop an application from functioning, or even worse become obsolete and simply disappear from the marketplace, stranding users without a growth path. It seems to us that this kind of assurance is critical when one is considering the preservation of our nation’s research data and publications.

We are eager to connect you with our communities of practice so that you can learn more about our community repository projects, our out of the box products, and our host of services to help implement flexible and durable open access content management solutions.

I’d like to thank the National Academy of Sciences for providing the opportunity to comment. Please feel free to contact us through our website at duraspace.org for more information. Thank you.

STAFF: Thank you. Up next, we have another WebEx participant, Nettie Lagace.

NETTIE LAGACE: Good morning to our hosts, the National Academies, and distinguished colleagues. My name is Nettie Lagace, and I’m associate director of the National Information Standards Organization, often called NISO. Thank you for convening this distinguished group and for allowing me to provide a bit of the technical perspective on these important issues. I’m sorry not to be in Washington in person, as is Todd Carpenter, executive director of NISO, but we’re very pleased to be able to make a comment virtually.

First, a brief word about our organization. NISO is a standards development organization accredited by the American National Standards Institute to develop standards, specifications, and industry best practices in the fields of information creation, distribution, collection, management, and preservation. We are a non-profit community comprised of more than 150 organizations split roughly in thirds between publishers, libraries, and systems developers.

Since 1939, NISO has been serving the needs of content creators, libraries, and end users by building efficiency into the systems surrounding content distribution. We represent the community in a variety of ways including representing the US national interests to ISO as the US technical advisory group for the ISO committee on information and documentation.

Over the past 40-50 years of digital content distribution, the committee has developed many systems for storage, descriptions, discovery, and access of traditional content. This foundational layer is fairly well established for textual content, particularly journals. While there remain enhancements in areas of research that are necessary to advance the capabilities of these systems, much of this foundation is in place for scholarly content. NISO is one among a small number of organizations responsible for the creation, consensus, education, and maintenance of these specifications and best practices.

Data publication, however, is a relatively new form of content distribution, and here community best practices are less well defined or agreed upon. In addition to being newer, data publishing is considerably more complex than traditional text-based forms. For example, data sets are not necessarily fixed in the way that text is when it is published. Data can be regularly updated and expanded upon. It can be processed and actionable.

The scale can be exponentially larger compared to text-based content. Authorship is not always obvious. Preservation requires more technological dependencies. The meta-data that needs to describe a data object are more significant and will make more of an impact if disassociated from the data. All of these elements are barriers to researchers' willingness to share data, they pose challenges to repositories trying to store the information, and cause sharing and integration of data to be considerably more challenging.

NISO has been engaged in many issues related to data exchange and preservation. As an organization of publishers, libraries and software providers, NISO is well placed to advance conversations related to data exchange standards and community best practices. We stand willing to support the public, private, academic, and philanthropic communities in their overall interests to support greater data interoperability and reuse. Here are a few NISO projects of particular note that are related to data exchange.

The ResourceSync initiative is a data synchronization protocol for web-based repository synchronization being developed in partnership with the Open Archives Initiative with support from the Alfred P. Sloan foundation. NISO's membership recently launched a project on meta-data and indicators for open access content. Key foundational standards such as the digital object identifier, DOI, and the Dublin Core Metadata Standard have been formalized within NISO. At the ISO level, NISO helped to lead the standardization of the International Standard Name Identifier, a system for unambiguously identifying the public identity parties including institutions, authors, and content creators.

Finally, NISO’s leadership has been engaged in a variety of international data exchange efforts underway within ICSTI, CODATA, and the W3C. NISO is beginning to explore other areas of work, such as formalization of data citation practices, author-type ontologies, and research on data equivalence.

However, despite the progress within NISO and a variety of organizations represented here, there’s a great deal more that needs to be done. Too many communities of interest are developing domain-specific repository systems, metadata structures, and infrastructures. While some domain specificity is obviously needed, a core set of identification systems, metadata, discovery services, preservation practices, and interchange protocols are necessary for a robust data exchange ecosystem.

Many research questions related to these standards remain, such as addressing conceptually the relation between data objects and describing the transformations. Once those specifications are created, a long road of education, adoption, and compliance-assessment will be critical to ensure those standards are put into practice. The NISO leadership is willing to discuss any of these initiatives in more detail, or the community’s needs or expectations when it comes to data exchange standards.

We welcome community input into all of our activities. Thank you for the opportunity to join you today and discuss NISO's activities and roles.

STAFF: Thank you so much. Next up we have another participant via the phone, Anil Srivastava.

DR. SRIVASTAVA: Thank you for the opportunity to make a statement at this meeting, which is very vital to our work.

Open Health Systems Laboratory is a non-profit organization focusing on global shared cyber infrastructure for medical research. We've been very pleased to listen to the speakers and the statements because that is very relevant to the work we do, but we are more of a user institution, user of big and open data, for our research purposes, the consortium that we built, so I'm bringing in a user's perspective.

As Professor Victoria Stodden said yesterday, open data is crucial to science and computation, and is becoming central to scientific research. I think what the speaker before me said is very relevant to what we do. To that end, we have brought together an international consortium, which we call ITT Biomed, which is ITT for biomedical global research collaboration of life sciences and supercomputing centers in India, Poland, Gothenburg, and other centers are joining us, including our high-performance computing first city that we're building here in Rockville, Maryland connected by an advanced network.

It's a grid of great computers with emphasis on infrastructure software, too, and people, which make data interpretive science possible. As I mentioned in the beginning, what we're doing is rebuilding global consortia for data intensive science.

Not long ago, there used to be meetings hosted by the United Nations on trans-border data flow. To data flow across the border is a reality and a necessity. Today that is what the Internet is all about, the free flow of data. We have accepted open data and public access to even federally supported research and development not as a mere necessity but a virtue. That is the rationale for this meeting.

The emphasis that I want to make and we have heard others mention the European network and initiatives elsewhere-- the emphasis I want to make is that the United States need to help spread the word to join the open data initiative, because disease travels freely. The climate changes are interrelated and interdependent. Large cohorts are needed to study the etiology of diseases, and so on. The words are interconnected; therefore for scientific progress we need open data to understand our planet and its inhabitants.

I think that if there is an effort which is focused in the United States alone, it is going to be lopsided because we need the data that reaches across, and there is informally less of an effort in making data accessible and curating data for sharing elsewhere in the world. That is where most of the population of the world resides. I think that's where the initiative needs to look at.

The years of investment of time and money in data science in the United States is very valuable knowledge and experience. We need to share this knowledge and experience with the world and in return ask them together to create an open global data initiative.

One of the international development agencies interested in the impact of big data is the World Bank, which happens to be co-resident with the National Academies and the rest of them in Washington, DC. Is a partnership performing a global quest for leveraging big data for development not a possibility? I'm not talking of big data just for economic development, but ideas like the Open Source Drug Discovery.

While trying to learn the science of data, we owe it to the world to co-develop the field of data science and address the world with open data sets for new understanding and new discoveries. After all, the Human Genome Project began here, and that achieved a quantum jump in the understanding of biomedical science. We are working on a note for the Academy of International Cooperation in open data.

I’m hoping that all of us here would focus on those issues, because not the US money-- we know that’s a scarce commodity-- but the US knowledge and expertise would prod people all over the world to be able to curate that data and develop the data for sharing, and benefit from the data that the US is making available in the open data space. With that, I would like to end my statements, and thank you.

STAFF: Next up we have another person on the phone, Joshua Rosenthal*.

DR. ROSENTHAL*: Good morning, I’m Joshua Rosenthal, a member of the National Committee on Health and Vital Statistics’ Work Group on Data Access. The following statement is my personal opinion based on experience having founded start-ups from public data with the Bush’s road map the W(?), and successfully scaling them across the market on the basis of using data to solve real world problems, such as improving quality across a continuum of care while reducing cost and creating patient-member engagement. It’s from that experience as an entrepreneur that I’d like to structure my comments on the policy goals of accelerating goals, breakthroughs, and innovation promoting entrepreneurship, and enhancing economic growth and job creation.

In a nutshell, as we shift from a fee-for-service paradigm to one that's performance-based, an important focus for a prospective entrepreneur is not so much around data as the donations. Applying the data to specific use cases, if you will, whether it's market-based or more broadly towards public and social good, releasing the data is important as the raw fuel, but the key to transforming that on a broad scale to do innovations comes from making not just the data transparent, but its meaning, making a clear code without this domain activity. With weather and geolocation, the data is relatively simple and self-explanatory, and the market application is pretty straightforward.

With healthcare, the data itself is not more complex per se, but rather its meaning. What is a DRG or HRI and other things that make up our alphabet soup can be evolved by the very motivated, navigating how they apply this to the in gap drug co-pay for HMOs in a county with relatively low (?) performance (?) adherence in high statin(?) populations is another matter altogether.

Healthcare start-ups fail at an astronomical rate, disproportionately compared to other variables, partially based on perverse incentives from a historic fee-for-service model, partially based on the opacity of meaning.

To that end, I'd humbly ask if you could incorporate the following components into the various access plans: taxonomy, paucity(?), and community. Taxonomy not only in the technical sense of metadata, but specifically in clearly defining the relationships to data element begins a series of common definitions and entity relationships. This is critically important to the current market policy context.

Fee-for-service paradigm challenges revolve around size, latency, and adjudication. In the performance paradigms, it's really about reconciling a large number of small files. Some of the most exciting innovation comes from some very different types of data sets, and here spending additional resources has disproportionate impact. Tax code examples of how to do this include not only data dictionaries but even publishing ERDs, which are entity relationship diagrams.

Paucity(?), some data's restrictive use if necessary, and likewise the barrier from outside fields, you need not only access to become better familiar with the data and its structure and taxonomy, but also from an entrepreneur's perspective, working raw material with which to build a prototype is the basis to gain a pilot or secure funding, or to create informed applications for restricted use. Specific tactical examples that need to be included are the synthetic files such as a decent top file.

Finally, community, there are essentially two approaches to helping entrepreneurs navigate all this. One is from the top-down; an agency releases the data that users work through individually. The other is bottom-up where users interact with one another; compounding their progress and creating assets that persist that help one another. I would submit that the agencies would be better served both in terms of reviews in resource taxation as well as getting valuable exposure to users directly by creating and actively cultivating community. Actual examples include online learning centers, data browsers, and the ratings and data comments about the data itself, as well as relationships of meaning, taxonomy, and even as applications, usage.

In short, I propose liberating not only the data but also the meaning. Whereas weather and geolocation are widely understood, healthcare as most here know, can be a little bit tricky. Thank you for your time and attention.

STAFF: Thank you so much. Next we have another speaker on the phone-- oh you made it, great. You're up next.

KENYON CROWLEY: Good morning. My name is Kenny Crowley, and I serve as deputy director for the Center for Health Information and Decision Systems at the University of Maryland where our work is focused on how do we facilitate the adoption and use of information systems in healthcare. I also serve on the National Committee for Vital Health Statistics' working group on health data and access. My comments this morning are my own and do not represent the official comments of the University of Maryland or the NCVHS.

The collection, aggregation, and sharing of data holds tremendous potential, but hopefully the effective use of this data to achieve the numerous desired outcomes depends in large part on how usable the data is and how easy it is for community and users to collectively leverage the knowledge that's being created across the different stakeholders that are attempting to use the data not only in the science, but how the data is creating value and how it's being used and applied by these different groups. I support a learning system and an infrastructure that enables that translation of data through analysis, linking, and collaborative knowledge development.

Integral to a learning health system or a learning system in general for data is the social learning aspects, so allowing users to provide feedback and assessments of the data, community meta-tagging of data, the ability to share experiences on working with specific data sets and facilitating collaboration, providing a directory or repository of products that use certain data sets. These activities can be poor facilitators to incite innovation. There are many successful models of social communities.

The open source community may serve as an exemplar. The usability of data is very important, so making data available for use in machine readable formats for computer programs or apps; by providing application programming interfaces for developers, scientists, entrepreneurs, students, and others can help lead to many creative things. The committee may consider a common API structure, data interoperability standards for federal health data that may help ensure its use and reuse. Also with the data sets, it’s important to provide information such as data dictionaries and include learning and training with data release.

Thank you again for the opportunity to comment this morning. I’m very excited about the prospect of the open data revolution we’ll have for both science, innovation, entrepreneurship, and the social good. Thank you.

STAFF: Thank you. I believe we still have one person on the phone. If there is someone who’s interested in speaking who’s not yet spoken on the phone, the microphone is now yours. Then if there’s anyone else in the room who has not yet spoken who’s interested in making a comment, now would be the time. If there are no other commenters this morning, we’ll adjourn until 11:00 a.m. when Professor King will be on the podium to give a wrap up rapporteur report. See you at 11.

* (?) in the Joshua Rosenthal remarks indicates words or phrases that were not understandable because of a poor telephone connection