Chris Biemesderfer chris.biemesderfer@aas.org American Astronomical Society

The American Astronomical Society (AAS) is the major association for professional astronomers in the United States, with over 7500 members. One of its primary functions is the publication of the key North American scientific journals dedicated to the dissemination of peer-reviewed research in astronomy and astrophysics, the Astrophysical Journal and the Astronomical Journal. As a society of research and higher education professionals, we have made a concerted effort to conduct our scholarly publishing enterprise with sensitivity to and balance among the need for prompt and inexpensive access to new results, the pressures on the budgets of technical libraries, and the challenges of obtaining grant and institutional funding to support author fees. The Society's mission has a broad public purpose, but its constituency is primarily professional research astronomers. Consequently, public access to data, while an attractive desideratum, is less of a concern than is ensuring access to data among research professionals engaged in on-going investigations. It is access and re-use by other scientists that will improve the productivity of the American scientific enterprise. Most scientific data themselves are not easy to monetize, so public accessibility follows straightforwardly once data are available to professional researchers. The AAS is in general agreement with the Interagency Working Group on Digital Data (IWGDD) that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." Agency policies should support and encourage a distributed system for both access and preservation. Once community-based repositories are in place and in use by a community, agencies and other entities such as learned societies and journals can insist on deposit of digital data. Deploying mandatory deposit policies in the absence of trustworthy repositories exacerbates challenges in communities already struggling with incompletely coordinated efforts to manage the increasing amount of data being produced. Community-based repositories need to be supported first, and soon. Support for (inevitable) deposit fees must be available for researchers whose data will be deposited in community-based repositories for long-term preservation. The IWGDD has also noted that there is a lack of a comprehensive framework for long-term data management in most disciplines. In astronomy, there have been efforts to resolve that shortcoming over the years, most recently in the form of "virtual observatories", and these have resulted in fairly effective channels of communication as well as a collection of standards and procedures for managing digital data across wide scales. National governments (not just the US) have a role to play in ensuring the development of the comprehensive framework envisioned by the IWGDD. This should take the form of continued support for efforts in broad disciplinary organization (like the virtual observatories and international alliances), and also support for the creation of trustworthy repositories in appropriate niches in the academy. In general, citation and attribution of data resources should follow the examples set in the academy for citing articles in the scholarly literature. There is a good deal of discussion in

academic circles about the need for data to be regarded as “first-class objects”, and it is important for that to happen. The broader interests of attribution are served in the community by an efficient and well-understood mechanism for data citation, akin to citing the literature. DataCite is an international coalition whose purpose to build a framework for persistent identification of data sets and for the evolution of policies and practices for citing data so that appropriate credit can be assigned to data set “authors”. The AAS prepared extensive comments in January 2012 for the OSTP. Cognizant agency personnel are encouraged to review those comments.

David Marques d.marques@elsevier.com Elsevier

Research Data: Needs and Solutions David Marques, SVP Research Data Services, Elsevier Anita de Waard, VP Research Data Collaborations, Elsevier Big vs. Small Data Problems? To move science ahead faster and more efficiently, the research community is increasingly making available research data (raw and summarized), both linked to publications and directly into open repositories. Funding bodies and government organizations are calling for massive increases in the sharing and availability of research data, with the hope that 'big data analytics' will greatly increase the speed and efficiency of science (e.g., see http://www.nih.gov/news/health/dec2012/od-07.htm). One of the key difficulties in building sustainable services around research data is that scientific research is built mostly on many small niches of research, a long tail effect, with many thousands of small datasets every year. In our view, the three biggest needs for research data today are: • Increase Data Sharing and Preservation: Help increase the amount and quality of data preserved and shared. In many scientific disciplines, little research data are currently made available. Reasons for the low participation include a lack of mechanisms for assigning credit, lack of distribution control, fear that others obtain key insights the scientists themselves have they overlooked, lack of standardized nomenclatures. • Increase Data Value: Help increase the value of the data shared by increasing annotation and interoperability. Because of the effort and informatics expertise needed to standardize, normalize, and add sufficient provenance and descriptive metadata required for domain-specific data repositories, data sharing today often means loading isolated datasets into generic repositories, which require little metadata or informatics support. These repositories do not integrate datasets on the same topic, are not easily discoverable, and provide little opportunity for analytics. • Develop Sustainable Models: Help measure and deliver credit for shared data, and enable sustainable infrastructures. The cost of data preservation and especially annotation for reuse is high, and will only increase dramatically in the coming decade. The research community as a whole – scientists, funders, repositories and publishers – need to find long-term models to pay for this. Research Data Services We submit that these needs are best addressed by a combination of skill sets and collaboration across the community of products and services that comprise scientific communication. Specifically, what needs to be done is summarized here. More detail is provided at a longer version of this paper at http://elsevierconnect.com/presentation-in-big-data-era-small-data-management-is-critical/ . Deposit with DOI Obtain peer-review where appropriate and requested. Normalize, anonymize Track and report downloads, usage, citation Perform standard checks for accuracy Link between datasets and papers Annotate with standard descriptive metadata. Enhance discoverability – adding standardized and variant nomenclature and registrations in data catalogs Annotate with provenance metadata and descriptions Create solutions as appropriate for specific tasks Publishers have much to contribute in all of these areas and organizationally have expertise in: • High-quality/-volume annotations, and particularly in creating

efficiencies in workflows at scale • Dealing with scientific information and data • Linked-data datasets and the skills and technology for joining multiple datasets by mapping nomenclatures • Long-term preservation and sustainable models The two most important principles for such services are: 1. Data must be open and shared, with distribution controlled by the creator of the data. 2. The model must be derived by the research community and funding agencies, not driven by publishers, though publishers can help. Recommendations Disclosing and sharing research data is a labor-intensive job, requiring special data science, process management, annotation and informatics skills, though there are many tools available to help. We recommend establishing pilots of Research Data Management at the university or research institution level containing clear explicit data management plans covering guidelines for all disciplines, with the goals of sharing more data usefully annotated into subject-specific repositories, and indexed in data catalogues or directories and documenting exceptions with the appropriate justifications, and full reporting of progress to all of the relevant funding agencies. The authors of this document are involved in 5 such pilots specifically to understand the skills, resources, and costs of implementing such plans, and to understand how the services needed scale up across disciplines. We propose forming a series of nimble expert advisory groups to develop and advise on the reuse and simplification of discipline-specific or repository-specific metadata and infrastructure standards, to improve cross-disciplinary data integration. This group will establish pilots of linked data repositories and data directories to explore and make recommendations for the scaling, value, cost and funding of extracting further insight from data-rich but fragmented disciplines, such as proteomics and metabolomics. These are currently built by commercial enterprises using the data without license or royalty, not helping support the repositories. This model should be examined further. We further recommend a working group followed by bounded pilot projects formed from a public/private partnership to explore and make concrete recommendations for long-term sustainable funding models of discipline-specific data repositories.

Jillian Wallis jwallisi@ucla.edu University of Michigan & University of California Los Angeles

Redundant Pressure with Little Hope of Release: A response to the recent OSTP memorandum on the sharing of research products 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 in academic research. And then I will add a rather longer caveat about how this does not really help the effort and may actually just add more pressure to an already difficult situation being faced by researchers today. The policy proposed in the Increasing Access memorandum [1] would be just one of many already in place to encourage and require publicly funded researchers to share their data with others. The NSF [2] and NIH [3] have had data sharing requirements since 1989 and 1990, respectively [4]. The NIH policy already requires a "data sharing plan" to be submitted along with a proposal, like the "data management plan" requirement recently added by the NSF. According to an upcoming study from the UK [5], about 50% of all academic journals, like Nature, have some data sharing policy in place. And this is on top of the already pro-social policies within academic communities. For example, seismologists are supposed to begin sharing data from an experiment six months after the last piece of equipment is out of the ground. Given all the pressure coming from the top-down, all researchers should be sharing their data. Sadly this is not the case, or we would not be escalating to policies sent down from the Executive Office of the President. Basically, the teacher has told researchers to share and they did not, the teacher got their parents involved to little effect, so now the principal has be summoned. Scientists are not petulant children who would refuse to share to be spiteful, so there must be some other pressure of equal magnitude from the opposite direction that needs to be overcome. Thus the real question is: what are the disincentives from the bottom up that are impeding the top-down pressure? Just in case there was some doubt, I would like to present some evidence to support my claim that data producers are not petulant children who are just refusing to share their data out of spite. Conveniently there are some researchers studying just how willing other researchers are to share their data. Through the next few points I will be using some complimentary results from a recent article and a forthcoming article, both in PLOS One, which capture data sharing attitudes and practices. In a survey of over 1300 people Tenopir, et al [6] found that 75% of researchers have shared their own data at some point, and 6% have made all of their data available. In a much smaller study from my research group [7], with 43 interviews and ethnography performed at an NSF-funded research center, we found that everyone we talked to was willing to share data, but that only half of them had been asked to do so. And when asked, researchers shared their data through a personal interaction with the data reuser. One of the ways 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 asked by other researchers to share their data. Repositories also have the added benefit of increasing the visibility of data to potential reusers and

providing long-term support. The Tenopir survey found that 78% of researchers were willing to deposit some data in a repository. Unfortunately, the actual rate of deposit from our much smaller study was closer to a third of the interviewees. We are still not seeing much traction with deposit in repositories, even though deposit would fulfill the various data sharing requirements. Another way researchers can share their data are by putting them online. This is definitely something researchers do, and apparently with greater frequency than they deposit their data in a repository. According to the Tenopir survey, nearly half of their respondents made their data available online, and our smaller study confirmed this results. For our study population nearly twice as many researchers had made their data available online as through a repository. Unfortunately, there are 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 still has too high a bar to entry for the majority of researchers to overcome, given that they are willing enough 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 were willing to share. Some of the most popular conditions our interviewees provided were: retaining the first right to publish from their results, receiving proper attribution as the source, wanting the requestor to be known to them, the ability to negotiate sharing in advance of exchange, etc. These conditions all seem quite reasonable and all easier to assure with a personal interaction than through a system that may not encode their conditions. From all this we can conclude that researchers are willing to share data when presented with the opportunity to do so. There are just a few hitches holding them back. First, there are very few requests for data to be shared. Second, sharing data is hard. Researcher have two equally bad options beyond personal interactions: depositing data in a repository is difficult but yields greater benefits, and putting data online is an easier option but suffers from serious accessibility issues. And third, why would researchers go to the trouble to deposit in a repository when they see very little demand for their 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 these data sharing policies and memoranda to be meaningful. To return to the Increasing Access memorandum, as you can see this is just another push from the top-down, like the funder, publisher, and even scientific community policies, that does not really get us any closer to the goal of sharing data. Everyone to whom this memorandum applies has other policies that have applied to them for decades, and still we have seen little movement towards a culture of sharing data. What would really help are policies or better yet funding opportunities to encourage reuse of existing data and jump-start the data sharing engine. References: [1] Increasing Access Memorandum, OSTP, 2013-02-22, http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [2]

NSF Award and Administration Guide, Chapter VI, Section D.4.b., http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp [3] Final NIH Statement on Sharing Research Data, 2003-02-26, http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html [4] Marshall, E (1990) Data Sharing: A Declining Ethic? Science. P. 954. [5] JISC, http://jordproject.wordpress.com/2013/02/01/a-rather-long-post-but-quite-a-brief-summary/ [6] Tenopir, et al (2011) Data Sharing by Scientists: Practices and Perceptions. PLOS One. [7] Wallis, et al (in print) If We Share Data, Will Anyone Use Them. PLOS One.

Tamara Wagester tamarawagester@cfare.org The Council on Food, Agricultural and Resource Economics

On behalf of the Council on Food, Agricultural and Resource Economics (C-FARE), we are pleased to offer comments on plans by the White House Office of Science and Technology Policy to review "Increasing Access to the Results of Federally Funded Scientific Research." We applaud OSTP for offering this opportunity for the community to respond. We appreciate and value your leadership on behalf of the science community. In general, the C-FARE supports the key provisions of the policy memorandum from the Office of Science and Technology Policy. C-FARE is a non-profit organization dedicated to strengthening the national presence of the agricultural economics profession. C-FARE promotes the work of applied economists and serves as a catalyst for incorporating economic thinking into the analysis of food, agricultural and resource decisions. We serve as a conduit between the academic research community and Washington, DC policymakers and agency personnel, matching expertise to public needs. Agricultural economics is the study of the economic forces that affect the food and fiber industry. Specific areas of study in agricultural economics include: (A) Community and rural development, (B) Food safety and nutrition, (C) International trade, (D) Natural resource and environmental economics, (E) Production economics, (F) Risk and uncertainty, (G) Consumer behavior and household economics, (H) Analysis of markets and competition, and (I) Agribusiness economics and management. As a result, our research areas reach the mission of multiple agencies and granting programs and our scientists utilize public data in a wide variety of important ways that help feed the United States and the World. Good decisions—public and private—require good data. The United States is fortunate to have some of the world's best data. While our rich data sources provide valuable information that informs countless decisions, we recognize that there is always room for improvement. C-FARE supports the concept of requesting all researchers receiving Federal grants and contracts to develop data management plans. This will help scientists increase the ability to replicate results, which can strengthen the peer-review process, scientific integrity, and long-term economic growth. Structured evaluation has become increasingly important in these current economic times, and we applaud the provision to encourage cooperation with the private sector to improve data access and compatibility. As long as privacy is maintained, the ability to work across sectors to obtain the most reliable information will benefit society. However, as an alternative to developing another costly data and/or publication repository, we urge the agencies to consider providing a standard acknowledgment to be added to internet documents so that private search mechanisms can return the requested results. Otherwise, we question the sustainability of the reporting system and have additional questions including whether the government is going to provide servers for storage, and whether the long-term preservation and access to the content will be provided without charge to individual researchers or professional societies? C-FARE also supports

the provision for an assessment of long-term needs for the preservation of scientific data in fields that the agency supports, including identifying options for developing and sustaining repositories for scientific data in digital formats. Public data reaches all aspects of our lives—including food, water, and economic development, just to name a few. Public data has recently come under increased scrutiny and threats of elimination. Prior to cutting any public data sets, we strongly recommend an evaluation be made of the linkages that this could break throughout an industry. For example, the elimination of a particular data set could impact the linkages between the producer level price of food and the choice of food by the consumer. Making improvements such as these will only enhance public and private decision-making capacity and help ensure society's future well-being. Thank you again for the opportunity to provide comments. Sincerely, C-FARE Board of Directors

Brenda Lonsbury-Martin blonsbury-martin@llu.edu Acoustical Society of America

I would like to support the view held by the Acoustical Society of America's Editor-in-Chief, Dr Allan Pierce. This strategy opposes a blanket 'open access' policy and suggests, instead, that certain university libraries, distributed over the country's geographical boundries, be designated as access sites. Such libraries will allow the public's on-line access to the medical and scientific journals they desire to consult. With this approach, the various related professional societies along with the commercial publishers will still be able to market their journals to their members and paying customers, respectively.

David Fearon datamanagement@jhu.edu Johns Hopkins University Data Management Servcies

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 largely in response to NSF's data management planning requirement, and in anticipation of assisting our researchers with the expansion of data sharing and management requirements that we are discussing today. We provide consulting on data management plan preparation, and training on data management best practices. Our department also operates a research data archive, working with the Data Conservancy, a growing community promoting data preservation and re-use across disciplines with tools and services. Our archive, which is built specifically for long-term public data access to research data, exemplifies the type of institutional archive called for by the OSTP. So as one of the first comprehensive service groups for data management and archiving among US academic libraries, we are keenly interested in funders' plans stemming from the OSTP memo, and are committed to supporting and helping implement their goals at Johns Hopkins and the broader scientific community through Data Conservancy. Drawing from nearly two years of experience helping researchers with data management and sharing, I am sure I could echo a number of issues and suggestions that others today will raise. I will focus on one suggestion relevant to the issue of supporting grant proposal reviewers as a factor toward improving researcher's incentives for producing better data management plans. For our service model, we conduct in-person consultations with faculty and researchers who are preparing data management plans (DMPs) for grant proposals, primarily for the NSF. We have heard many researchers admit, sometimes begrudgingly, that it can be a useful exercise to codify their plans for data management at the proposal stage, and to consider which research data might be valuable to share beyond the publication. Therefore, as we monitor data management plans being produced in the last several months, we have been concerned that researchers have not had clear incentives 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. The need for more clarity of guidelines is one factor implied by the OSTP memo. We have created a DMP questionnaire worksheet to supplement guidelines, and other academic libraries have created similar resources. We feel that a lack of guidelines for grant proposal reviewers may be a significant factor in the overall incentive to produce effective DMPs, and 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 faculty members asked to review NSF proposals that they do not know what to look for when evaluating a DMP, and some simply check off whether one was included in the proposal. By and large, reviewers may not be receiving encouragement to focus on DMPs in proposals, therefore they are they are not passing along feedback to grant applicants about the quality of DMPs or their data sharing efforts. The White House OSTP Memo (in Section 4, item d) addresses the need to "ensure the appropriate evaluation of the merits of submitted data

management plans" suggesting that the quality of plans and data sharing should more directly factor into overall grant decisions, and for compliance for implementing plans after grants are awarded. We believe that NSF's Social, Behavioral and Economic Sciences Directorate is among those addressing support for grant reviewers. We wish to underscore in funders planning that guidance and resources designed specifically for grant reviewers be considered. As an example of such a resource, we at JHU Data Management Services have developed a worksheet for our JHU faculty on NSF review committees that provides a simple checklist for evaluating and comparing proposal DMPs. The checklist covers NSF's recommended content for DMPs, 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. Formatted as a Microsoft Word template, the worksheet may be used on computers, tablets, or as a printout. We have had some positive feedback from faculty reviewers that the guide helps make reviewing DMPs more convenient while conveying the main points of an exemplary data management and data sharing plan. We offer the Grant Reviewer's DMP Worksheet and Guide as an example to build upon as funders consider resources for proposal reviews as part of their overall plans for expanding data management planning and data sharing. The reviewers guide is available on our website, with the link in the submitted text or you may find it by searching for JHU Data Management Services. Link to JHU DMP Reviewer Guide and Worksheet: http://dmp.data.jhu.edu/assistance/grant-reviewers-worksheet-for-data-management-plans/ JHU Data Management Services website: http://dmp.data.jhu.edu. datamangement@jhu.edu

Tribikram Kundu tkundu@email.arizona.edu University of Arizona

Easy public access of research results is possible by one of the following two ways 1) Giving the public access to journal articles free of charge or with nominal charge. It is possible through open access journals. 2) Giving them access to journal articles through university libraries free of charge or with nominal charge. Problem with the open access journals is that they have very high charge (often over $1000) for publishing an article. It is a big barden on researchers and some journals are eager to publish even low quality papers for getting the publication charge and making profit. They will not care if anyone will read that article as long as they make their money. Some researchers with lot of research funding will be willing to spend the publication charge to get their low quality work published in a journal to get the promotion. Thus overall quality of the published papers will be lowered. Therefore, my suggestion is let the interested public have easy access to the journal articles through the University libraries for nominal fee and come up with a mechanism for the university libraries to get compensation from the government for giving public the access to their libraries.

Jonathan Markow jjmarkow@duraspace.org DuraSpace

DuraSpace Written Statement for NRC DBASSE Meeting on Public Access to Federally-Supported Research and Development Data and Publications By way of introduction, this statement is from DuraSpace--an independent 501(c)(3) not-for-profit organization providing leadership and encouraging innovation in open source repository technologies that promote durable, persistent access to digital content and data. We collaborate with academic, research, cultural, government, and technology communities by supporting open source projects to help ensure that current and future generations have access to our collective digital heritage. In addition, we offer hosted services for organizations that would like quick access to archiving and preservation solutions with minimal maintenance. DuraSpace supports the initiative to promote the dissemination and long-term stewardship of research results funded from Federal science agencies. The open access, open source repository applications we promote, DSpace and Fedora, are used by over fifteen hundred institutions world-wide for disseminating digital content. These institutions include many 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 strongly recommends that technology solutions deployed for this initiative be based on open source software applications, which have a number of advantages relevant to the current needs. For one thing, licensing expenses are non-existent compared to the often steep costs of commercially licensed software. Open source software comes with freely available source code, as well, and is supported by communities of practice. Government agencies and departments deploying open source applications like DSpace and Fedora are able to join a global community of developers in adding or changing features to meet their specific requirements if they care to. Changes may be contributed back to the community so that others may 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. DuraSpace is happy to elaborate further on the advantages of applications built by open source communities as well as provide additional information about DSpace, Fedora, and DuraSpace hosted services.

Timothy Vollmer tvol@creativecommons.org Creative Commons

Creative Commons (http://creativecommons.org) applauds the White House directive supporting universal access to publicly funded research articles and data. It is a productive step toward speeding up scientific discoveries, promoting information sharing, and increasing the return on investment of public monies. The Administration is "committed to ensuring that... the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community." Creative Commons (CC) would like to help agencies fulfill this aspirational plan. CC has created free copyright licenses and public domain dedication that serve as the legal plumbing enabling innovative sharing of creative works and data by authors, publishers, data providers, and countless others on terms that are more open than "all rights reserved." CC's legal tools along with forward-thinking data sharing plans ensure science is more collaborative and participatory. And by making data, the raw material of science, available to everyone freely, increasingly scarce resources can be directed toward scientific analysis and discovery instead of duplicative data collection activities. As agencies build their individual public access plans in the coming months, we recommend that they take as progressive an approach as possible with respect to data. This would mean: 1. Requiring that any data that support published research conducted using federal monies are marked clearly as being in the public domain and immediately deposited in a scientific data repository; 2. Requiring that organizations and researchers that use federal monies to collect or aggregate data without any specific research project in mind also make their data available freely in the public domain; 3. In cases where making the data available in the public domain is not suitable or applicable, data may be marked with a liberal license such as CC BY, which allows full use and copying of data with no more obligation than giving credit to the data creator. We believe that CC licenses and public domain legal tools can help federal agencies meet the requirements set out by the White House directive. The goal of the White House directive is broad reuse of publicly funded research. The Administration has taken the important first step by removing price barriers to this research data. Federal agencies can take the next logical step by removing permission barriers as well. By requiring that researchers make their research data available immediately as open access, federal agencies will be clarifying reuse rights so that downstream users know the legal rights and responsibilities in reusing those data. This is an important and useful public service. But why is communicating reuse rights important? As Creative Commons board member Michael Carroll writes, "Granting readers full reuse rights unleashes the full range of human creativity for translating, combining, analyzing, adapting, and preserving the scientific record" (N Engl J Med 2013; 368:789-791). When permission is granted via a public domain dedication or via standard public licenses, researchers can more easily understand what they can do with the datasets. A public domain dedication or a very liberal attribution-only license ensures disparate datasets can be mixed and combined without running afoul of creators' permissions.

This is vital in today's global scientific research environment, where massive amounts of data--from potentially thousands of sources--are produced, analyzed, and reused in new experiments. The communication of clear, unambiguous rights to data break down barriers to reuse. And of course, data creators still receive the credit they deserve because the norms of science have always dictated proper acknowledgment and attribution along with citation. Scientific data repositories such as Dryad, figshare, and DataONE allows researchers to upload and make their data available under a CC0 Public Domain Dedication (http://creativecommons.org/publicdomain/zero/1.0/). CC encourages federal agencies to fulfill the letter and spirit of the White House public access directive by crafting their agency policy to require immediate deposit of data to such repositories. By marking data clearly, other researchers will know exactly what they can do with the data. CC is standing by to help.

Joe Hourclé joseph.a.hourcle@nasa.gov Solar Data Analysis Center

As one of the items in the OSTP memo directs agencies to "develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan" (4.h.), I wanted to bring to their attention the recommendations of the Technical Breakout of the BRDI meeting on "Developing Data Attribution and Citation Practices and Standards" in August of 2011. [http://sites.nationalacademies.org/PGA/brdi/PGA_080121] As many of the problems with citation of data are related to the problems of identifying the data, [Wynholds, http://dx.doi.org/10.2218/ijdc.v6i1.183], building a registry of the data with sufficient specificity to establish identity would not only serve to solve task 4.h., but could also be used by the public to specify which data should be given priority in making it available (in the case of this in dark archives) or that need to be made more accessible for those already available. For more background on the topic, please see "Linking Articles to Data" [http://virtualsolar.org/citation/], a poster and handout that I co-authored with members from that breakout group and presented at the 2012 Research Data Access and Preservation (RDAP) Summit [http://www.asis.org/rdap/].

Juan Arvelo juan.arvelo@jhuapl.edu Johns Hopkins University

As a researcher, I'd suggest that the Government take steps to enable persons from the public to have general access to at least one university library (presumably within geographical proximity), for nominal fees, and that the Government take steps to insure that the results of research funded by the Federal Government be assessable via online access to university libraries. In the event that the results appear in peer-reviewed journals, the Government may decide just which peer-reviewed journals are to be included in a recommended list of journals that should be generally available to the public via online access to university libraries. Preference should be given to those journals that are associated with recognized professional societies and whose subscription prices are within reasonable bounds.

Joe Wolfe J.Wolfe@unsw.edu.au The University of New South Wales

I am not a US citizen, but am an associate editor of J. Acoustical Soc. America (JASA), which is the world's leading acoustics journal and, de facto, the principal international journal. My observations are likely to be relevant to most fields of science and to most countries. Much of the work published in JASA comes from outside America and much that comes from the USA is not supported directly by the US government. FIrst, it would be great if scientists from institutions that cannot afford subscriptions to a wide variety of journals could access all the important research, regardless of who funded the research. Second, I think that science as a whole would be better served if work not supported by the US government were not disadvantaged (i.e. be less readily accessible) as a result of changes in policy. US funded labs do much of the world's best research, but excellent research is done elsewhere, too. It would be difficult, but not impossible, to find a compromise. I recognize that the costs of publishing are large and it would be difficult to achieve both aims completely. However, I urge decision makers to take these two issues into consideration.

The Alexandria Archive Institute /Open

Sarah Kansa skansa@alexandriaarchive.org Context

Comments for the meeting on Public Access to Federally Supported R&D Data Prepared by Eric Kansa and Sarah Whitcher Kansa The Alexandria Archive Institute / Open Context May 8, 2013 We would like to contribute the following comments to the meeting on Public Access to Federally Supported R&D Data, following the Office of Science and Technology Policy's memorandum on open access. The Alexandria Archive Institute (AAI) commends the OSTP for further promoting access to data, which we see as critical to improving research and education. Our comments are based on ten years of exploration of issues around open access to digital data and data reuse in and beyond the scholarly community. The AAI (http://alexandriaarchive.org) is a non-profit organization that works to promote the dissemination and curation of digital scholarly resources. To this end, we developed Open Context (http://opencontext.org), a free, open access system for the publication of editorially-vetted and peer-reviewed research data sets. Open Context demonstrates readily achievable ways to cultivate a distributed foundation for digital scholarship. Its methods for data portability enable researchers to work across silos and use a host of visualization, search and analysis tools. By leveraging archival and identity services offered by the University of California's California Digital Library (CDL), Open Context gains a strong institutional foundation for permanent citation and archiving. 1. Cultivate a dynamic information ecosystem. Policy-makers should avoid locking-in to particular institutions or approaches to managing data. Data management needs are diverse and ever-evolving. We need to encourage innovation and participation from the widest possible community. 2. Managing data often extends beyond archiving. In many cases, researchers may value data as more than a residue of their studies. While necessary, archiving in many cases may not be sufficient for the effective communication and reuse of data. Agency guidelines should accommodate and promote multiple approaches to the management and dissemination of data that extend beyond preservation requirements (i.e. data publication, exhibition, and other innovative forms of scholarly communication). 3. Software is an important part of the picture. Data and software work hand in hand in research. Agency guidelines need to recognize that data access and reuse require appropriate software and software documentation. To replicate research findings, software critical in the lifecycle of data creation and analysis need to be open source. 4. Recognize and reward data excellence. Agencies can play a role in fostering positive incentives for the production, management, and reuse of high-quality research data. These can be specific granting programs, awards, and (as proposed) encouraging budgeting around data management. 5. Metrics need context. There may be strong temptations to institute uniform metrics within or between agencies involved in different research domains. While impact and other metrics about research datasets can be invaluable, agencies need to apply such metrics with caution. Metrics can be difficult to establish because not all data management

should be measured in the same way. For instance, certain efforts to document and preserve irreplaceable cultural heritage information may have long-term value that will be difficult to measure with short-term metrics. Metrics about data in different domains need to be crafted in collaboration with stakeholders, and may need revision and updating as needs change. 6. Accommodate change. Policy needs to recognize that scholarly communication practices need to be able to evolve. Data dissemination is part of a larger, evolving world of scholarly communications. 7. Coordinate agency policies. Some research areas receive funding from multiple agencies. For instance, both the NSF and NEH fund archaeology. Agencies should coordinate data management policies and allow enough flexibility in the interpretation of such policies so as to avoid administrative burdens and duplication of effort. 8. Research data as a public good. Ideally, data should be available without burdensome licensing or contractual limitations. To maximize the research and public value of data, data should be accessible according to general licensing and technical recommendations advocated by Creative Commons and the Open Knowledge Foundation. Furthermore, policy implementations need to avoid privileging commercial interests over other public interests with respect to data (including search, visualization, aggregation, data integration, and other applications). At the same time, concepts of the public domain and data openness are culturally situated and should not be imposed arbitrarily, especially with regard to culturally-sensitive information (such as archaeological site locations and information regarded by indigenous peoples as sensitive or sacred). Ethical data management requires involving stakeholder communities in navigating culturally-situated data sensitivity and privacy issues. 9. Identify funding gaps. Open access and open data can lead to fundamentally greater equity in the conduct and outcomes publicly funded research. However, creating public goods typically requires public financing. Agencies need to be in an honest dialogue with representatives of research communities (inside and outside universities), libraries, and other stakeholders to clearly identify financial requirements for creating, using and preserving high-quality open research data. Better accounting of these financial needs will help better inform future policy making and budgeting.

Richard Buckius rbuckius@purdue.edu Purdue University

R&D Data Access Comments from Purdue University May 8, 2013 White House Memorandum "Increasing Access to the Results of Federally Funded Scientific Research" Purdue University is a doctorate granting, land-grant university established in 1869. Purdue's West Lafayette, Indiana, campus has 39,256 students, 15,612 faculty and staff , and in FY 2012, system-wide Purdue faculty received nearly $354 million in sponsored funding for research. Purdue is classified as having "very high research activity" by the Carnegie Foundation. The University supports the principles of the policy outlined in the February 22, 2013, memorandum from the Office of Science and Technology Policy (OSTP), which provide public access to the outputs of federally-supported research and development, and in particular, to enable the reuse of research data. We appreciate that the OSTP and National Academies have offered the opportunity for stakeholders to give input and to serve as collaborators with those agencies as their individual or collective policies are developed and implemented. Universities house major stakeholders in this system, with the faculty and researchers receiving grants, conducting research, and disseminating the findings that are collected and stewarded by libraries and archived in a variety of formats. Research data sets are treated and managed as part of the intellectual output of Purdue University and its scholarly record. In the same spirit as this policy and the land-grant mission, our research office, libraries, and office of information technology have collaborated to launch an institutional data repository service, the Purdue University Research Repository (PURR, http://purr.purdue.edu), that provides campus researchers and their collaborators with a platform for online scientific collaboration using the HUBzero cyberinfrastructure. Funded as a university research core facility, PURR supports the entire research data lifecycle with resources for data management planning, a virtual research environment (VRE) for collaborating and conducting research online, the ability to publish data sets with Digital Object Identifiers (DOI), and a secure, trustworthy data archive. Since 2011, PURR has been cited as a component of the data management plans of 702 grant proposals that have been submitted to federal funding agencies. Librarians and research office staff train and consult with investigators to develop effective data management plans and support the use of the PURR service. They employ tools that Purdue helped to create such as the Data Curation Profiles , Databib , DMPTool , and DataCite. These kinds of institutional efforts are complementary to the OSTP policy, and institutions such as Purdue can play a role in helping the agencies in implementing it. A model for supporting and ensuring public access to data should balance flexibility â€" to accommodate variation among the needs and communities of practice supported by the different agencies â€" and consistency, to establish a common expectation to support and meet the goals of the policy. The model should address the full research lifecycle (e.g., data management planning, proposal review, compliance, and reporting) as well as the training and support. A variety of approaches for data stewardship should be encouraged including solutions from institutions

(e.g., institutional repositories such as PURR), professional societies, publishers, consortia, and the funding agencies themselves. Over time, these solutions may converge as best of breeds emerge. Purdue offers the following additional recommendations: 1. Each research project funded by the federal government should be required to deposit its resulting data in a suitable repository and be made publicly accessible. 2. Data should be made available in a timely manner, ideally upon completion of the grant. 3. A suitable repository should be defined as one that is trustworthy and meets all requirements for ensuring full public accessibility, productive reuse (including downloading, machine analysis, and computation), interoperability with other repositories housing federally funded scientific data sets and publications, metadata based on open standards, long-term stewardship and preservation, and appropriate confidentiality safeguards, without charge to the author or end-users of the data. 4. Agencies are encouraged to leverage the public investment in central repositories such as those provisioned by the National Institutes of Health, when a government-sponsored option exists. 5. Some existing university repositories, disciplinary repositories, and other data collection platforms may also be suitable. Allowing researchers to deposit data in the repository of their choice, when an existing central repository is not available, will increase compliance. 6. Agencies should require the use of persistent, globally-resolvable, unique identifiers for data sets (such as DataCite DOIs) in order to facilitate the precise location and identification of data sets as well as encourage their reuse with proper citation and attribution. 7. In addition, agencies should require the citation of data in scholarly publications resulting from federal sponsorship as well as all agency reports. Manuscripts and final published papers should be linked to their source data to allow for reuse and replication of results. 8. To track the effectiveness of agency policies, a variety of metrics and identifiers should be supported to provide information on access, use, and impact of the availability of data sets resulting from federal research. Various metrics have been implemented in university repositories. Agencies should also develop plans to assess the broader economic and societal impact of their policies. We believe that the development of consistent federal agency policies to ensure access to research data will benefit our nation, our economy, and our future, and that it will accelerate scientific discovery, improve education, and empower entrepreneurs to translate research into commercial ventures and jobs. To realize this potential, we encourage agencies to be as consistent as possible in their policies and compliance requirements to minimize the cost and complexity of compliance with grant requirements for both principal investigators and research administration. Unlike commercial interests, the enduring mission of universities is to generate new knowledge, and the mission of the libraries is to preserve and make accessible that knowledge for future generations. We strongly believe that, just as U.S. Copyright Law does not protect data sets of facts facilitating re-use of data, implementation of this policy must ensure that no one single entity or group has exclusive rights to the use or re-use of federally funded

data. Public access policies can stimulate the development of new tools and services that generate opportunities for the public, industry, and the scientific community.

Ann Wolpert awolpert@mit.edu Massachusetts Institute of Technology

The Massachusetts Institute of Technology (MIT) welcomes the opportunity to comment on the digital data aspects of the February 22, 2013, White House Memorandum on "Increasing Access to the Results of Federally Funded Scientific Research." MIT's mission includes a commitment to generate, disseminate, and preserve knowledge, and public access to scientific data resulting from research funded by federal science and technology agencies is a topic of substantial significance to this institution. There is no universal model for data deposit; therefore, this welcome policy directive must balance economy with the potential to accelerate the generation of new knowledge through ease and assurance of sustainable and appropriate access to digital data. MIT supports the primary objective of the digital data aspects of the Memorandum, which is to ensure that the direct results of federally funded research are made available to and useful for many stakeholders. In the development of federal agency digital data plans, MIT additionally urges consideration for the interests of researchers, students, and research universities, as well as for industry, the public, and the broader scientific and educational community. MIT has a long history of consistent commitment to the ideals and objectives of open access to research outputs, and a record of innovation in support of the principles of open access. MIT's support for the OSTP's extension of a public access policy for digital data includes the following recommendations and considerations. 1. To generate the greatest compliance and benefit, agencies should adopt common practices to avoid unnecessary proliferation of idiosyncratic requirements. 2. A suitable repository should be defined within data management plans prescribed in the Memorandum, and such plans should be subject to evaluation. Suitable repositories will vary by discipline, but ideally, a suitable data repository will be one that: a) meets all requirements for ensuring full public accessibility, productive reuse (including downloading, text mining, machine analysis, and computation), b) interoperates with other repositories, c) uses metadata based on open standards, and d) provides long-term stewardship and preservation. Data repositories should be able to demonstrate, rather than simply assert, their capability for long-term stewardship of data. 3. Since most digital data is not subject to copyright, agencies should develop a common usage agreement or rights waiver that makes clear the rights of the data producer and consumer. Copyright or IP rights should be only be assigned, if necessary, in a non-exclusive manner that would ensure preservation, discovery, mining, and sharing of digital data. Licensing arrangements, if any, should ensure that no one single entity or group secures an exclusive right over digital data or new business opportunities. 4. Data that supports research results published in a peer reviewed article should deposited and be available at the same time that an article referring to the data is published. If an embargo period for data is necessary, that embargo period should be as short as practicable, but ideally no longer than six months after the publication in a peer-reviewed journal of the last scholarly article resulting from the grant. 5. Proposals for Federal funding for scientific research should allow for provision for

appropriate costs for data management, preservation and access. 6. Agencies should work with disparate stakeholders to develop a minimum set of core metadata for all datasets, along with an API for standards-based data exchange, to help ensure a level of interoperability and discovery across all disciplines. Persistent identifiers for data sets are critical. 7. Stakeholders should adopt an agreed upon standard for citing data.(1) This will enable the easy reuse and verification of data, allow the impact of data to be tracked, and create a scholarly structure that recognizes and rewards data producers. 8. Absent the need for necessary restrictions such as the protection of privacy and confidentiality, data should be made available for open reuse as a default. Reports by the National Research Council (2) argue that data produced or funded by government agencies should continue to be made available for research through a variety of modes, including full access to original data under appropriate license and security restrictions, mediated access to confidential data through interactive systems, and open access to data altered to maintain confidentiality. This approach will avoid the problems inherent in inconsistent or simplistic treatment of information confidentiality and security, which are a barrier to efficient access to and reuse of research data. The development of consistent federal agency policies to ensure open access to scientific data in digital formats will provide economic and social benefits to our nation, by accelerating discovery and science, supporting education, democratizing access to information, and fueling economic growth, entrepreneurship and job creation. MIT offers these recommendations from the perspective of an educational and research institution which, in concert with other U.S. universities, is a regular creator and consumer of scientific data (as well as scholarly publications). MIT also has demonstrated experience and expertise with digital preservation and the long-term stewardship of a range of digital outputs. Research universities have a primary and enduring mission to generate new knowledge, to preserve it, and to share it, and we are uniquely positioned to appreciate the benefits and challenges of the goals of the Memorandum. MIT commends the OSTP on the Memorandum and stands ready to provide additional input at any stage in the evolution of the implementation plans. NOTES 1. See for example DataCite's efforts described at http://www.datacite.org/whatdowedo 2. See: National Research Council. 2005. Expanding access to research data: Reconciling risks and opportunities. Washington, DC: The National Academies Press. And: National Research Council. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: The National Academies Press. Respectfully submitted by Ann Wolpert, Director of Libraries, Massachusetts Institute of Technology

Ginny Steel vsteel@ucsc.edu University of California Libraries

To: The National Academies From: University of California Libraries Re: Statement for the Public Comment Meeting on Public Access to Federally Supported R&D Data May 8, 2013 The University of California's Council of University Librarians (CoUL) applauds the federal departments and agencies that are developing plans to support increased public access to the results of research funded by the Federal Government, including data. Research data play a fundamentally important role in scientific inquiry and policy discussion, with both direct and indirect societal impacts. Datasets must be properly curated (managed for the long term) in order to ensure that they remain available for use, sharing, and re-purposing by current and future scholars. We encourage the National Academies of Science to emphasize these actions: 1. Encourage data sharing. One of the most direct positive impacts that agencies can have in encouraging long-term management, preservation, and sharing of research data is via instituting appropriate requirements as a pre-condition, and an auditable post-condition, for funding. Data sharing should be an expectation and a default, wherever possible. Although preservation and sharing will initially be driven by external requirements, over time they will come to be accepted merely as normative patterns of scientific activity. Agencies should couch the intent underlying their requirements in terms of the benefits from following these practices, such as increased opportunities for collaboration, directed or serendipitous discovery, publication, and citation. 2. Emphasize thorough data documentation. The objects of stewardship must be clearly documented in terms of significant form or structure, scientific meaning, and desired behavior in order to facilitate successful preservation, discovery, and use. Thorough documentation is most useful if discipline-relevant metadata standards and controlled vocabularies are used. Minimal metadata is rarely adequate for data reuse; agencies should place as much value on thorough, high-quality metadata as on the data itself. 3. Encourage workflow preservation. Although well-organized datasets and high-quality, thorough metadata are often sufficient for facilitating data reuse, reproducibility requires that the workflow itself be preserved alongside the data and metadata. This includes any software or scripts required to read and analyze the data, plus details about any significant hardware dependencies. 4. Enforce standards and interoperability. Funding agencies should recommend the widest possible use of the most common data formats and analytical tools. The use of research data is predicated on three factors: knowing that the data exist, knowing from where the data are available, and having it made available in a form that is easily integrated into local workflows. These suggest the need for common standards for data description, publication, discovery, and representation formats. Data description must be supported at sufficiently fine grain to enable direct and, ideally, automated determinations of the suitability of a given dataset for a particular local purpose. Conformance to standards and best practices will be greatest when those practices are perceived as arising from within the community of concern and practice, rather than being imposed externally. Funders and agencies can play

an important role in encouraging and funding working groups with cross-disciplinary members working towards common standardization or standardized crosswalks. 5. Support preservation and access. Both coarse and fine grained replication and redundancy in all aspects of the stewardship infrastructure â€“ technical, curatorial, and procedural â€“ are some of the most powerful means to minimize the potential for debilitating single points of systemic or correlated failures. However, the costs of proper preservation and provision of access are not inconsequential. Both funders and institutions must recognize and plan for these longer-term costs. The allocation of scarce curatorial resources, whether financial or otherwise, is always based on evaluations of the current and future value proposition for the curated data. Evaluation criteria should include the scientific value, scope of applicability, and degree of uniqueness and reproducibility of the data. Since any assessment of future value can be problematic, it is important that all plausibly useful research outputs are subject to minimally sufficient baseline practices, and that there is some level of ongoing curatorial assessment to select data deserving of added value attention in light of evolving circumstances. Fundamental to preservation and access is a strategy for keeping identifiers persistent and unique. The public-facing dataset identifier strings “whether URLs, DOIs, or ARKs“ should be chosen with care to avoid suggestions of brands, technologies, and other transient properties that over time will hinder fulfillment of the preservation promise. That promise itself must be backed up by commitment to dataset stewardship, to organizational succession planning, and to redirecting identifiers as datasets move from one archive to the next. Opaque identifier strings that are globally unique and never re-assigned work best. Identifiers that can be embedded in URLs today benefit not only from “one-click” access and easy lookup via global search engines, but also from whatever transition will be realized in a post-URL, post-web Internet. 6. Provide governance structures. Facts are not copyrightable under United States law. However, because not all data users realize this, because facts may copyrightable in other jurisdictions, and because it is not always clear whether data are purely factual or contain copyrightable expression, copyright concerns can inhibit productive reuse of shared data. Agencies should enable the widest possible reuse of data by recommending clear and permissible reuse terms, such as the CC0 mark. Agencies can encourage and fund working groups to create frameworks for standardized data use agreements. The University of California recognizes that some data, such as patient records, are sensitive or classified and that immediate sharing and reuse of this data is not practical. Agencies can carve out specific exceptions and limited time embargoes for those cases while maintaining an overall standard of openness. 7. Emphasize incentives, especially data citation. Providing scientists with assurance of appropriate attribution and credit for making available their research output can be facilitated through support for formal data publication and citation. While the historical practice has been to provide public visibility to only one of the many outputs of a research program - the summarizing paper or conference presentation - there is no reason why the other data products could not be similarly

treated, wrapping those products in the familiar facade of academic publication. Providing datasets with persistent identifiers and descriptive citations enables the entire scholarly publication infrastructure to come into play to provide sophisticated aggregation, indexing and abstracting, enhanced discovery, and attribution, all of which should combine to encourage more widespread use and repurposing. Publishers and data repositories should be encouraged to add and maintain bi-directional links between traditional academic publications and the data that underlies them, with all of the attendant mechanisms and incentives for citation, attribution, and impact analysis.

David Wojick Dwojick@craigellachie.us David Wojick Consulting

It would be better if NRC accepted written comments after the hearings so people can respond to the issues raised therein.

Mark Newton mnewton@columbia.edu Center for Digital Research and Scholarship, Columbia University Libraries/Information Services

The Center for Digital Research and Scholarship at Columbia University Libraries/Information Services welcomes the opportunity to respond to the February 22, 2013, White House Memorandum on â€œIncreasing Access to the Results of Federally Funded Scientific Research.â€• As an institution dedicated to advancing knowledge and learning at the highest level and to conveying the products of its efforts to the world, we support the Memorandumâ€™s objectives of making the results of federally funded research available to and useful for the public, industry, and the scientific community. The key to encouraging public access to and preservation of digital data is the creation and funding of open data repositories that adhere to common standards for the identification, description, and storage of data, coupled with a policy framework for funding agencies that requires, and provides monetary support for, the deposit of data from funded research in such repositories wherever possible, as well as a clear framework for communicating the usage rights for that data. Individual government agencies increasingly have an important role to play in encouraging the benefits of publicly available data, both by creating data-aggregating portals that provide a unified point of access to disparately archived data and by promoting and incentivizing best-practice solutions for data archiving and preservation. These solutions may require discrete, focused attention and fostering by federal agencies, as researchers continue to direct limited resources on dissemination goals and problems of access, when they address data management issues at all. 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 individual academic institutions, consortia, and/or scholarly societies in partnership with government, rather than by any individual government agency alone. The first step is to formally build the funding and tracking of research data costs into the data management plan (DMP) requirements that already exist. Only by integrating the costs of long-term stewardship and dissemination of data into the granting process will it be possible to gather enough information to allow for a proper consideration of the relative costs of various data types, which will be a prerequisite for an evaluation of the full benefits to other researchers and to the public of such data. Of course, formally integrating these costs into the granting process, while an absolutely vital first step, will not be the end of the story. In addition to setting the stage for further evaluation of the costs and benefits of different data types, agencies must pay attention to the ongoing, unanticipated costs of data stewardship, such as data migration, and create mechanisms for meeting those emergent needs that cannot be integrated into and accounted for in the existing grant funding workflows. Mechanisms for making research data available should be closely tied to existing workflows for grant management so that important stakeholders are minimally burdened by these new requirements, thereby reducing both administrative costs and obstacles to compliance. Automation will be essential to

minimizing the burden of compliance. Such automation requires that compliance be integrated into existing workflows (for granting, research, and publication/dissemination) and that it be based on clearly communicated standards. These mechanisms will also need to be tied to shifts in the workflows for publication and dissemination of research more broadly speaking. Since access to research data is absolutely necessary to ensure reproducibility, there will need to be identification and description standards built into the compliance process that ensure that data are clearly associated with the publications that cite them and the code used to process them. The work done by DataCite (http://datacite.org/), using widely accepted standards such as Digital Object Identifiers (DOIs) to facilitate the discovery, reuse, and impact tracking of data, and by NIH, integrating compliance for published articles based on NIH-funded research into standard publication and grant workflows, both offer models for similarly handling data tracking and compliance. The multiplicity of standards poses a challenge to any widespread attempt at data sharing, although we do join in the endorsement for a centralized index of such standards, believing that such a resource could foster adherence to community practice and reduce barriers to interoperability. A successful plan to address data archiving and preservation will tackle questions of governance, adoption or development of standards and conventions among disciplinary communities, and necessary new investments in technological infrastructure that make data management possible. The establishment of baseline metadata requirements for interoperability is a key area where agencies can provide leadership, working closely with discipline-specific groups such as professional and scholarly societies, information technology specialists, librarians, and research administrators to ensure that the data in these repositories are stored and described in ways that enhance their discoverability, as well as providing for machine-readability. It also means that researchers and other stakeholders must agree on basic standards for the identification and description of data, though those baseline standards should remain minimal, with specific disciplines having room to establish their own, more granular standards to meet discipline-specific data and metadata needs. Standards must also address issues in the area of intellectual property, focusing on encouraging the clear labeling of data so that all stakeholders are aware of the use conditions of a given dataset. That labeling should be done in a human- and a machine-readable format with an awareness of the global context in which data live today, as well as of the ever-greater speed with which novel uses for them emerge. It is also vital that government agencies encourage the use of such labeling at all steps of the data lifecycle, since raw data may go through many transformations before they find their way into publications and other end-uses, but the ability to trace those data end-to-end will be an essential part of the verification process. Perhaps more significantly, however, we once more urge the involvement of the scholarly and professional societies in a direct way in the identification and development of these 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. Empowering these organizations (again, through incentives articulated centrally through individual agencies) thus both strengthens their positions as arbiters of authority and respects individually established contexts, initiatives, and standards.

Susan King s_king@acs.org

American Chemical Society, Professional & Scholarly Publishing Division, Association of American Publishers

On behalf of the Professional and Scholarly Publishing Division of the Association of American Publishers (AAP/PSP), I appreciate the opportunity to comment on the Office of Science and Technology Policy's (OSTP) memorandum on â€œIncreasing Access to the Results of Federally Funded Scientific Researchâ€• and to offer the support of publishers as federal agencies craft plans to efficiently and effectively promote access to data. We welcome the opportunity to work together to address the needs of the federal agencies and the scholarly communities we both serve. As organizations deeply engaged in the dissemination and discovery of information, AAP/PSP members' innovative products and services enhance and add value to taxpayer-funded research activities. We understand the opportunities and challenges inherent in the OSTP memorandum and are committed to helping the scientific community address the goal of expanding access to data. For AAP/PSP members that publish scientific journals and other peer-reviewed scholarly publications, the primary goal of their publishing activity is to disseminate information and provide access by providing a high quality and user-friendly digital environment in which to discover, analyze and link to the latest breakthroughs and developments in scientific and other scholarly research. In particular, publishers of scientific journals have, for more than 100 years, played an integral role in building and documenting the unrivalled US scientific research enterprise. In addition to their efforts to disseminate publications that report and analyze the latest research, publishers also have considerable experience and investment in digital technology, metadata standards and tools to help users discover, understand and manipulate data. This makes publishers uniquely positioned to help the Federal Government expand public access to digital data, ensure the long-term stewardship and discoverability of data and support the innovation and economic development that is derived from scholarly advancements. Publishers support better discoverability and reuse of scholarly data and are pleased that OSTP has recognized the distinction between data and peer-reviewed publications in its policy memo. Dissemination and discovery of information is an area of publishersâ€™ professional expertise, and data access policies potentially impact AAP/PSP members. There are challenges in increasing access to data that are distinct from those in publications, however the fundamental principles remain the same: any policy should be collaborative and flexible; build on existing infrastructure and investments where possible; and recognize and account for the costs of the validation, curation, dissemination, discovery and preservation of research outputs and their derivatives. In particular, publisher investments have created digital platforms with the latest and continually evolving Web discovery tools, providing researchers with faster and more robust delivery of scholarly information, new ways to present data and research findings, and links that enable information to be found and navigated with ease. Publishers have improved

interoperability through new metadata standards and pilot projects, which are driving innovation and providing for better information discovery and expanded use of research results. Government should not diminish the incentives for creative publication that allow publishers to provide tools that enhance innovative reuse and discovery of research information. While incentives to share research through publication are embedded in the academic and research processes, incentives for sharing data and ensuring complete tagging of individual datasets are limited. There is no consistent approach to presentation, no widely adopted standards for which data should be preserved or its overall management, and no consistent approach for data storage, tagging and dissemination. In addition, the significant costs of storage, distribution bandwidth and overall management and curation must be addressed. These issues should be addressed through collaboration among all stakeholders to promote the development of robust, sustainable and flexible standards that meet the needs of users at all levels. It is critical that the federal government continue to distinguish between information products created, often at considerable cost, for the specific display and retrieval of data (â€œdatabasesâ€ •) and sets or collections of raw relevant data captured in the course of research or other efforts (â€œdata setsâ€ •). The research interest and value of raw research data sets and individual data points is entirely different, and serves different purposes, from that of specific databases that have been organized and compiled by publishers for particular research needs and to which intellectual property or copyright protection may apply . The ODE Data Publications Pyramid provides a useful model for understanding how research data can be presented in a variety of ways with increasing levels of curation and analysis. In addition to the technical and structural challenges to promoting data access, there are also cultural issues to be addressed. Incentives, rather than mandates, should be used to encourage the deposit of data in appropriate repositories, linking with publications and between data-sets, and the curation of datasets across projects and fields. Incentives could also play a key role in the development and deployment of technical standards for transparent description, identification, management, retrieval, and reuse of datasets and the integration of distributed data, text and tools. A federal role in expanding access to and the preservation of digital data could include partnering with the scholarly community for the identification of standards and best practices for the interoperability of data repositories; creating clear rules for citation, modification and privacy; improving links between data, research grant reports and peer-reviewed publications; and advancing policies and funding to ensure the long-term sustainability of data archives. For example, all data could be assigned persistent, unique Digital Object Identifiers which would aid their discovery, use and citation. Appropriate metadata â€“ with agreed-upon standards â€“ could be generated with the data to foster understanding and reuse. OSTP could also learn from initiatives already underway to standardize metadata and provide links between sources of research information. Collaborative approaches - such as CrossRef (www.crossref.org), DataCite (www.datacite.org), Opportunities for Data Exchange (www.ode-

project.edu), APARSEN (www.alliancepermanentaccess.org/index.php/current-projects/aparsen/) and the NISO/NFAIS Working Group on Supplementary Journal Information (www.niso.org), among others - provide the best way forward towards broad access to and preservation of digital data. A pilot project on linking data and publications is already underway at NSF; similar research and experimentation should be undertaken to ensure the best policies move forward. Publishers seek to continue to advance science by providing high-quality services to researchers, and data is already a part of that mission. We want to work in partnership with other stakeholders to achieve the vision where data are more broadly managed, preserved and reused to enhance science and innovation in the US.We welcome collaboration with agencies and other stakeholders to establish a sustainable framework for the discovery and use of data.

Micah Altman escience@mit.edu Massachusetts Institute of Technology

Comments on Public Access to Federally-Supported Research and Development Data The Data Preservation Alliance for the Social Sciences (Data-PASS) welcomes the February 22, 2013, White House Memorandum on â€œIncreasing Access to the Results of Federally Funded Scientific Research.â€• Data-PASS (http://Data-PASS.org) is a broad- based voluntary partnership of data archives dedicated to acquiring, cataloging, and preserving social science data, and to developing and advocating best practices in digital preservation. Collectively, the founding partners have over 200 years of combined experience in social science data sharing Data sharing needs to be built into the research and publication workflow â€" and not treated as a supplemental activity to be performed after the research project has been largely completed. Furthermore, ensuring long-term access to data requires a multi-institutional approach. As many threats to long-term access can be effectively ameliorated only when collections are replicated, geographically distributed, and audited by independent institutions. [Rosenthal 2005] Data-PASS, as stewards of established non-profit data repositories, and as stakeholders, respectfully offers the following recommendations: 1. In 1985, the NRC [Fienberg, et. al 1985] issued recommendations for access to research data. The core recommendations of this report should guide the development of policies requiring data management plans and the creation of individual data management plans. In particular: (a) Sharing data should be a regular practice. (b) Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances. (c) Data relevant to public policy should be shared as quickly and widely as possible. (d) Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible. 2. Any data that is essential for the full understanding of a published work should be recognized as an essential part of the scholarly record [Altman 2013]. Such data should (a) be included for public distribution in a data management plan; and (b) cited in any publications which rely on that data. 3. Robust infrastructure is now available for data citation. [Brase 2012] Data citation should include at least the following elements: author (or authoring entity), title (possibly a generic title), a date (or formal database version, if available), a persistent identifier (such as a DOI), and some form of fixity information (that can be used to validate data retrieved later). [Altman & King 2007] 4. At each stage of a research lifecycle, from project design through data collection, analysis and publication, knowledge about the research and data is created. When information about instruments, methods, context, and meaning from across the stages of research are shared, data are more trustworthy and linkages among disparate data can be formed. Standards for capturing metadata (â€œdata about dataâ€•) should be supported and encouraged. 5. Inconsistent and simplistic treatment of information confidentiality and security are a barrier to efficient access to and reuse of research data. A series of reports by the National Research Council [2005, 2009], have reinforced that data produced or funded by government agencies should continue to be made available for research

through a variety of modes, including full access to original data under appropriate license and security restrictions, mediated access to confidential data through interactive systems, and open access to data altered to maintain confidentiality. 6. Like treatment of other risks to subjects, treatment of data privacy risks should be based on scientifically informed analysis that includes the likelihood of p risks being realized, the extent and type of the harms that would result from realization of those risks, the availability and efficacy of technical, computational/statistical, and legal methods to mitigate risks. [Vadhan, et al. 2010] 7. There exists diversity in approaches for data management within various scientific communities, which is healthy. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure. However reliability cannot be assumed. In support of the stipulation (OSTP memo, section 2.c) that agencies develop â€œstrateg[ies] for measuring and, as necessary, enforcing compliance with its plan." We recommend that (a) providers of infrastructure for access to research data regularly demonstrate rather than simply assert capability for long term stewardship [NDSA, 2011]; (b) the effectiveness of agency data availability policies be regularly assessed; (c) individual data management plans be systematically evaluated for compliance. 8. We strongly support section 4.c of the OSTP memo which allows the inclusion of appropriate costs for data management, preservation and access in proposals for Federal funding for scientific research. The costs for these activities and their infrastructure over time are non-trivial. We believe that the development of consistent federal agency policies to ensure access to data will accelerate scientific discovery, improve education, and empower entrepreneurs to translate research into commercial ventures and jobs. Our organizations, unlike commercial entities, have a primary and enduring mission to generate new knowledge, to preserve it, and to share it. We are uniquely positioned to support the goals of the Memorandum. We commend the OSTP on the Memorandum, and we stand ready to provide additional input at any stage in the evolution of the implementation plans. Respectfully submitted on behalf of the Data-PASS by itâ€™s steering committee. George Alter, ICPSR; Micah Altman, MIT; Mark Abrahamson, Roper Center; Merce Crosas, Harvard U. Jon Crabtree, Odum Institute; Gary King, Harvard; William LeFurgy, Library of Congress; Amy Pienta, ICPSR; Libbie Stephenson, UCLA References: Altman, M., & King, G. (2007). â€œA Proposed Standard for the Scholarly Citation of Quantitative Dataâ€□. DLib Magazine, 13(3/4), Altman, M. 2012. â€œData Citation in the Dataverse Networkâ€• , in For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop, National Academies Press. Brase, Jan, 2012, The DataCite Consortirum in Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop, National Academies Press. Fienberg, et al. (eds). 1985. Sharing Research data. Washington, DC: The National Academies Press. National Research Council. 2005. Expanding access to research data: Reconciling risks and opportunities. Washington, DC: The National Academies Press.

National Research Council. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: The National Academies Press. NDSA 2011. â€œResponse to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Researchâ€• . Available from: http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf Vadhan, S. , et al. 2010. â€œRe: Advance Notice of Proposed Rulemaking: Human Subjects Research Protectionsâ€• . Available from: http://dataprivacylab.org/projects/irb/Vadhan.pdf David S. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. â€œRequirements for Digital Preservation: A Bottom-Up Approachâ€• , D-Lib Magazine 11 no. 11 (2005)

Anthony DeCrappeo tdecrappeo@cogr.edu Council on Governmental Relations (COGR)
Council on Governmental Relations (COGR) The Council on Governmental Relations (COGR) is an association of 190 research universities concerned with the impact of federal regulations and practices on the conduct of research. Our goal is to ensure that policy goals are met without burdensome administrative structures that may hinder compliance. Observations on Objectives for Scientific Data: We appreciate that the definition of research data is â€œconsistent withâ€• the definition of data in OMB Circular A-110 noting, however, the distinction between the definitions. A-110 defines data as â€œthe recorded factual material commonly accepted in the scientific community as necessary to validate research findings.â€• The OSTP definition addresses digital data needed for validation and includes a reference to â€œdata sets that support scholarly publications.â€• This is more than a distinction without a difference because it is important to understand what must be stored and made available to search, retrieve and analyze. There is a difference between research data and research materials. For much of the research community, research materials are those from which data can be extracted. Materials are tangible or physical objects, e.g., a database, cells, designs, forms, flow charts, planets, and/or plants. Research data, on the other hand, is information that provides a quantitative and/or qualitative description or characterization of the materials holding the data. Thus, it is important to distinguish between the entities containing the data and the data themselves with regard to preservation and access. The definition appropriately excludes, for example, lab notebooks, recordings of interviews, or an insect which are not data but contain information from which data (as characterization) can be created. Harmonized Approach: A common definition is critical as a foundation for a common approach in agency plans. If the Federal government is serious about addressing the administrative burden then the plans approved by OSTP and OMB must be harmonized to ensure compliance. The policies governing data generated under grants must be harmonized with data developed under contracts which are under the Federal Acquisition Regulations which define data differently. The worst case would be a multi-agency project funded through differing mechanisms with competing definitions and requirements for preservation and access. In such a case, there will be a single set of data that would have to be stored in varying formats or repositories with differing access requirements. Processes for establishing the primary or prevailing requirement must be set by the agencies. Confidentiality and Intellectual Property: OSTP cautions that access must be weighed against competing concerns including protection of confidentiality and personal privacy; preserving proprietary and/or business interests and intellectual property rights; and the value of preserving and access balanced against the costs. There may be additional restrictions on access to certain data like export controls or â€œsensitive but unclassifiedâ€• status which should be exempt from agency plans as well. We share OSTPâ€™s concerns about providing appropriate protections and a careful assessment of the costs weighed against the administrative burden. We are very concerned about

the protection of intellectual property rights in a rush to provide access. The person most capable of using the data is the person who created it and it would be a mistake to jeopardize their ability to exploit the potential of the data. Circular A-110 provides mechanisms for access to research data supporting regulations. Access through a FOIA request may not be judged to maximize access, but these mechanisms establish a useful and, we would argue, necessary controlled access to data. With recent changes in the laws governing patent protections, delays in the release of data are critical to meet the new first-to-file standards. Unlike patentable inventions falling under Bayh-Dole provisions, there is no single source of authority on ownership and protection of data. Generally the agencies that provide Federal financial assistance allow recipients to copyright and own data developed under the award, subject to the right of the agency to use the work for Federal purposes and/or specific programmatic requirements. Under contracts which include software as data, the government may allow copyright but normally retains the ability to exercise all the rights of the owner, e.g., distributing copies to the public, further complicating property issues. Data from Federally supported research can be of interest to potential private investors but without further investigations and directed proofs of the concepts, the usefulness is limited. The ability to protect the intellectual property is what attracts businesses to make the investment in time and resources to license the technology and bring products to the market. Institutions have been aggressive in identifying research with the potential to stimulate economic development and work to patent and license the intellectual property to the benefit of the business partners and, ultimately, the public. Public access may deter pursuit of these endeavors. Security: We caution the agencies to avoid invoking or imposing significant security standards on data â€“ standards that can pose a costly challenge for institutions. Data can and should be secured for its stability but there is and must be maintained an important distinction between Federal government data bases that are subject to FISMA and those maintained on campuses or through other extramural arrangements. Of course security will be a concern in an environment of broad public access, but requirements to deposit data in Federal data sites that trigger compliance with Federal security requirements will be a significant disincentive. Controlled Access: We believe that the agencies should consider a controlled access approach to research data preservation and access. Using something akin to a virtual data enclave approach would, at the outset, help to provide and yet control access and ensure data stability and security. The enclave, depending on the nature of the data, may have a more or less restrictive access but necessary information on appropriate attribution and tutorials or training on access and use can be provided to the public to ensure that that access provides value. Data that holds sensitive personal information or potential proprietary information could be housed with restricted access until such time as limited data sets could be created or patent applications and/or licenses can be secured to afford necessary protections. Costs: The enduring problem with allowing direct costs related to management, storage and access on proposals for or an award of

Federal funds is that the actual storage, maintenance and access costs will continue after the funded activity is complete. There will be no Federal award open to allocate those costs to and limited funds to ensure long-term preservation and access. Without an acknowledgement of these increased costs in the Facilities and Administrative (indirect cost) recovery rate, institutions will be forced to implement another unfunded mandate from the Federal government. Summary: Plans must be harmonized across agencies and mechanisms; access must be organized and indexed to provide value; and must be controlled to provide necessary stability, security and value to the public.

Indiana University - Indiana University

Robert McDonald rhmcdona@indiana.edu Libraries & Data to Insight Center

Crowd-Sourced Infrastructure: Universities as Partners in Provisioning Public Access to Federally Supported Research Robert H. McDonald, Inna Kouper, Beth Plale Data to Insight Center (http://d2i.indiana.edu), Indiana University I. Introduction The Office of Science and Technology (OSTP) recognizes that the discovery and exploitation of the results of federally supported research (FSR) can be fully realized only when those results are widely available to researchers, corporations and the public (Holdren, White House Office of Science and Technology Policy. February 22, 2013. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf). In this position paper we look at universities as key partners in provisioning public access to FSR and argue that a decentralized solution that â€œcrowd-sourcesâ€• rich cyberinfrastructure and personnel resources from many universities will enable and enhance public access to federally supported research products. As a significant portion of federal funding goes into research universities, their activities and infrastructure, including technological capacity and library and administrative resources, offer immense capabilities in implementing national and global public access that is efficient and at scale-able costs. Our work with an NSF-funded project for data stewardship for sustainability science known as Sustainable Environment Actionable Data, or SEAD (SEAD: An Integrated Infrastructure to Support Data Stewardship in Sustainability Science. Available at: http://dx.doi.org/10.6084/m9.figshare.651719), demonstrates that a decentralized crowd-sourced cyberinfrastructure supports the OSTP goals of leveraging existing archives, fostering public-private partnerships, optimizing search and discovery, and enabling compliance with federal policy mandates. Initially, public access infrastructure will be scoped to publications and data, but in the future we see a need to think beyond this initial strategy to find options to include a wider diversity of research products, such as software, workflows, specimens, instruments and so on. II. Leveraging Existing Infrastructure via a Decentralized Federation Many universities are tapping into their own resources in supporting access to research publications and data. The richness of university resources that can be used to support public access to FSR can be seen within each of our institutions of higher education. Universities are bringing the expertise and existing cyberinfrastructure together with the appropriate policy organizations to drive long-term preservation of research output and permanence of the research record including the ability to deliver enhanced public access to research publications and data. These partnerships have built capabilities for linking publications and data, capturing data provenance, and re-using data through computational modeling and synthesis. The time has come for universities and federal agencies to shift from an isolated individual agency or institutional approach to a collective effort in public access that relies on local governance and loose standards-based infrastructure, evolves organically and leverages existing institutional resources. A decentralized system

that uses research universities as anchors or nodes in facilitating access to FSR will leverage the following components of existing infrastructure: â€¢ Storage Systems, including systems of immediate storage and access, such as institutional repositories and digital libraries that exist at many universities and archiving partnerships such as the Digital Preservation Network (DPN) (http://www.dpn.org) â€¢ Networking services as developed through partnerships such as Internet2 (http://www.internet2.edu/) â€¢ Data Curation and Management Services (IU Data Management Task Force (2011) https://scholarworks.iu.edu/dspace/handle/2022/13221) â€¢ Computational Expertise to automate metadata harvesting, search federation and component integration â€¢ Administrative Workflows to leverage existing research administration systems III. Challenges and Benefits of Decentralized Approach The decentralized university-based approach raises a number of challenges. How can data stores effectively support two forms of data - observatory data, i.e., data that are collected over time and sampled by various researchers for the purposes of their own research and focused data, i.e., data that are collected for the purposes of a particular research? How can the underlying infrastructure tap into semantic linked data approaches to support linkages within and across universities, government agencies and their resources? How diverse organizationally and technically can the infrastructure nodes be? How can local governance and practices be harmonized at both a macro and micro level? Does decentralized sharing need to support flexible â€œpluggingâ€□ and â€œunpluggingâ€□ from the global structure? Among the barriers that the decentralized approach would also need to address are the issues of policy and integration between multiple federal agencies and state and private institutions, lack of integration between various stakeholders, for example possibly competing interests of commercial publishers and disciplinary-based institutional repositories, and the challenges of supporting standardized scholarly communication workflows that are part open and part closed. At the same time, the benefits of relying on university-centered decentralized infrastructure include: â€¢ Leveraging resources and capabilities across the entire research lifecycle and creating opportunities to intervene at the earliest stages of research. â€¢ Decreasing the gap between data creation and preservation by embedding data curators within the research teams. â€¢ Minimizing costs by sharing existing infrastructure and personnel and by providing local storage and support. â€¢ Fostering partnerships between data producers and data managers and thereby increasing efficiency of data production and dissemination. â€¢ Customizing solutions that address local researchersâ€™ needs. â€¢ Increasing the efficiency of infrastructure use by utilizing sophisticated algorithms that match user needs with system requirements for access and preservation matchmaking. â€¢ Diversifying the system of knowledge production and open access to it by integrating journal publications in their pre/post-print form and related datasets. IV. Conclusion To conclude, a decentralized system of access to federally supported research that is based on current agency and university infrastructure and expertise and that is aligned with the policy outcomes of the federal research

agenda will enhance access to FSR. This will be accomplished by supporting both management and analytics of research products and harmonization of multiple localized access and storage solutions while fostering a community of active proponents that enables long-term access and reuse by future users of FSR products.

Alfred Spector azs@google.com Google Inc.

To the Executive Office of the President, Office of Science and Technology Policy: We commend the OSTPâ€™s goal of ensuring public access to federally funded research and encouraging federal agencies to work together to do so. As the federal agencies develop plans to support increased public access to research, we would like to offer a few comments based on the years we at Google have spent working on our mission â€œto organize the worldâ€™s information and make it universally accessible and useful.â€ 1. Enhance the publicâ€™s ability to locate and process information In the Agency Public Access Plan, each federal agency is instructed to offer a strategy for leveraging existing archives (a) and to focus on improving the publicâ€™s ability to locate and access digital data (b). Among the Objectives for Public Access to Scientific Publications, each agency is asked to facilitate easy public search, analysis of, and access to publications (c). As one example, we hope that agency representatives will consider how search indexing and search engines can help to leverage the data archives that are already available. Google is open to conversations with government agencies about the best way to ensure data can be located and processed to derive scientific value. 2. Store data â€œin the cloudâ€• to help optimize, connect and share In the service of â€œoptimizing search, archival, and dissemination featuresâ€• (Agency Public Access Plan, c) in the areas of accessibility and interoperability, one area that may be important for large datasets is the physical hosting of the data itself. As cloud-based data storage services become more common, thereâ€™s an opportunity to provide cost-effective storage thatâ€™s widely accessible. By consolidating data storage, federal agencies can increase the utility of data by reducing the need for moving data among multiple separate silos. Also, users can access and analyze the data on servers near the data, without needing to make copies. 3. Address cost appropriately With limited budgets, cost is clearly an important aspect of the plan, especially existing budgets (Agency Access Plan, f). In the Objectives for Public Access to Scientific Data in Digital Formats, the OSTP asks that agencies allow â€œinclusion of appropriate costs for data management and accessâ€• (c). The issue of â€œappropriate costsâ€• is a prescient one. Currently many researchers host data locally, at far lower reliability and risk of data loss, because computing hardware can be purchased without paying grant overhead to their institutions, whereas cloud computing resources may incur the overhead charges, despite efficiency from greater scale. Local hosting also incurs many indirect or amorphous costs, even though they many not reach the level of explicit line items in budgets. Also, larger scale cloud computing is typically more energy efficient and environmentally friendly. Current financial incentives discourage hosting data in the public domain (in the â€œpublic cloudâ€•); federal agencies have an opportunity to alter the incentives in a way that can help achieve the OSTPâ€™s goals of increased public access. 4. Public-private partnerships can help to host digital data Among the Objectives for Public Access to Scientific Publications, each agency is asked to facilitate easy public search, analysis of, and access to

digital data (c) and (d) to encourage public-private collaboration that avoids unnecessary duplication (ii). Hosting data in the public cloud can reduce the need for copying data among many local servers, and can help achieve this goal by allowing analysis on servers co-located near the data itself. The OSTP is working towards a vital goal of sharing results of federally funded research. We at Google are open to conversations with government agencies to help ensure the OSTPâ€™s goals are met. Sincerely, Alfred Spector, Vice President, Google Research Jonathan Bingham, Product Manager, Scientific Computing

Ross Mounce ross.mounce@okfn.org Open Knowledge Foundation

Each year, the Federal Government spends over $100 billion on research. This investment, in part is used to gather new data. But all too often the new data gathered isn't made publicly available and thus can't generate maximum return on investment through later re-use by other researchers, policy-makers, clinicians and everyday taxpaying citizens. A shining example of the value and legacy of research data is the Human Genome Project. This project and its associated public research data are estimated to have generated $796 billion in economic impact, created 310,000 jobs, and launched a scientific revolution. All from an investment of just $3.8 billion. With the budget sequestration of 2013 and onwards it's vitally important to get maximum value for money on research spending. By ensuring public access to most Federally funded research data it'll help researchers do more with less. If researchers have greater access to data that's already been gathered they can focus more acutely on accumulating just the new data they need, and nothing more. It's not uncommon for Federally funded researchers to perform duplicate research and gather duplicate data. The competitive and often secretive nature of research means that duplicative research and data hoarding are probably rife, but hard to evidence. Enforcing a public data policy on researchers would thus help them to make the overall system more efficient. This tallies with the conclusions of the JISC report (2011) on data centres: â€œThe most widely-agreed benefit of data centres is research efficiency. Data centres make research quicker, easier and cheaper, and ensure that work is not repeated unnecessarily.â€• Another more subtle benefit of making Federal-funded data more public is that it would increase the overall importance and profile of US research in the world. Recent research by Piwowar & Vision (2013) robustly demonstrates that research that releases public data gets cited more than research that does not publicly release its underlying data. The as yet untapped value of research data: I believe most research data has immense untapped re-use value. We're only just beginning to realise the value of data mining techniques on 'Big Data' and small data alike. In the 21st century, now more than ever, we have immensely powerful tools and techniques to make sense of the data deluge. The potential scientific and economic benefits of such text and data mining analyses are consistently rated very highly. The McKinsey Global Institute report on 'Big Data' (2011) estimated a $300 billion value on data mining US health care data alone. I would finish by imploring you to read and implement the recommendations of the 'Science as an Open Enterprise' report from the Royal Society (2012): * Scientists need to be more open among themselves and with the public and media * Greater recognition needs to be given to the value of data gathering, analysis and communication * Common standards for sharing information are required to make it widely usable * Publishing data in a reusable form to support findings must be mandatory * More experts in managing and supporting the use of digital data are required * New software tools need to be developed to analyse the growing amount of data being gathered Further Reading: Science as an open enterprise (2012)

http://royalsociety.org/policy/projects/science-public-enterprise/report/ Tripp, S & Grueber, M (2011) Economic Impact of the Human Genome Project. Battelle Memorial Institute, Technology. Partnership Practice www.labresultsforlife.org/news/Battelle_Impact_Report.pdf Piwowar, H & Vision T J (2013) Data reuse and the open data citation advantage. PeerJ PrePrint https://peerj.com/preprints/1/ JISC (2011) Data centres: their use, value and impact http://www.jisc.ac.uk/publications/generalpublications/2011/09/datacentres.aspx Manyika et al (2011) Big data: The next frontier for innovation, competition, and productivity http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Francis McManamon fpmcmanamon@asu.edu The Center for Digital Antiquity

Public Access to Federally Supported R&D Data We suggest policies and procedures to encourage preservation of and access to data resulting from federally supported scientific research. First, it is important to have a broad definition of â€œfederally supported research.â€• Many federal agencies support research as part of environmental impact studies, for example, and this should be included under the term. Research about archaeological and historical resources should be included, as well as more commonly recognized subjects like geology, biology, medicine, climate, etc. For example, most of the archaeological research in the US is funded as part of environmental impact and historic preservation reviews required by NEPA, the National Historic Preservation Act, or the Archaeological Resources Protection Act. Federal agencies report that the number of archaeological investigations they undertake or require exceeds 50,000 annually. The data from such studies should be considered as â€œfederally supported scientific research.â€• Data generated from such research should be archived in an appropriate trusted digital repository or archive dedicated to the preservation of and access to data and its supporting documentation. Part of this requirement should include the creation of appropriate and sufficient metadata for discovery, so that data are not simply preserved, but also readily accessed and available for future uses. Regarding the kinds of digital repositories most appropriate for data archiving, we suggest that discipline-specific repositories provide a rich context of similar materials so that users, as they search, are provided with search results tailored to their expectations and needs. Metadata in disciplinary repositories contains phrases, keywords, and categories that match subject matter domains, making search results much more targeted to information resources that are especially useful. Disciplinary repositories, as opposed to institutions or organizations with generalized missions, provide efficiency and productivity of tasks related to digital archiving and minimize cost while enabling better chance for discovery and access to data. Policies and procedures should encourage data repositories to include in their archive any documentation relevant to the original data sets. For example, repositories should include reports related to the data. Repositories also should ensure appropriate linkage to metadata, which would ease searches among related data and make background research more efficient. Federal agencies should work with organizations that have established disciplinary repositories to develop policies and procedures relevant to their area of expertise. Professional societies, publishers, and other organizations in some disciplines are addressing issues of data preservation and access specific to their research. Working with such groups on developing policies and procedures is sensible. For example, within the discipline of archaeology, the Center for Digital Antiquity (http://digitalantiquity.org) and the Archaeology Data Service in the UK have cooperated on various topics related to the digital archiving of, providing access to, and the use of archaeological digital data. Lost or inaccessible data essentially means that federal funds spent on a research project were expended for limited results. There is no potential for

future benefit to the public or the American scientific enterprise through reanalysis or alternative uses of the data. Preservation and access may require marginally more funding to ensure that the research results are accessible and preserved, but such costs are amortized over a very long time period and will have a broad range of economic, scientific, and educational benefits. The cost for long-term digital archiving is more complex than simple data storage. It requires active curation of digital files and their ultimate conversion as software becomes obsolete and hardware and software advances are made. A â€œpay once, store foreverâ€• model (e.g., Goldstein and Ratliff, 2010 [http://arks.princeton.edu/ark:/88435/dsp01w6634361k]) is a reasonable option. Federally supported projects should include an appropriate direct cost line item for long-term curation and preservation of digital research data. As a condition of grants, agencies should require thorough citation details as part of the metadata for archived resources. Appropriate attribution is a fundamental part of the responsible conduct of scientific research. Disclaimers and terms-of-use could be required by the digital repository laying out expectations of proper attribution if the data or supporting documentation is used in any fashion. Repositories also should be encouraged to offer services that provide, for example, a standard format for citing the research data set (similar to the formats used for citing published works). Included in this citation should be permanent identifiers that are associated with a particular data set (e.g., Digital Object Identifiers [DOIs]). Best practices for the long term preservation of digital data should be compiled, published, and promoted, such as the Guides to Good Practice (http://guides.archaeologydataservice.ac.uk/) developed for archaeological archiving and data files by the Archaeology Data Service in the UK and the Center for Digital Antiquity in the US.

Brian Athey brian.athey@transmartfoundation.org tranSMART Foundation

The tranSMART Open Data Sharing and Analytics Platform to Enable International Biomedical Research at Scale The broad dissemination of digital information and data from biomedical research in a manner that makes it generally available and specifically useful to the research community is a key goal of the Obama administration[1], and of the newly formed tranSMART Foundation[2]. Our common vision is to realize the promise of translational biomedical research through an open-data framework. The tranSMART Foundation was formed using a Public-Private Partnership (PPP) model to organize and foster the development of an open-source / open-data community around the tranSMART data warehouse and analytics platform, enabling â€œgenotype-to-phenotypeâ€• studies. This platform lowers barriers to entry for translational research, enabling biotechnology, pharmaceutical, academic, government and non-profit research efforts through ready access to state-of-the-art platforms to access, manage and utilize digital information and data from publicly sponsored biomedical research. The tranSMART Foundation strongly believes that the experience of early and open sharing of data from the Human Genome Project should be adopted by other segments of the biomedical research community, including proteomics, metabolomics, and other biomedical molecular measurement â€œâ€™omicsâ€• fields. The tranSMART Foundation is a Delaware not for profit membership corporation that was formed to enable effective sharing, integration, standardization, and analysis of heterogeneous data from collaborative translational research by mobilizing an open-source and open-data community around the tranSMART platform. Such an eco-system creates the opportunity for innovation by lowering the barriers to access to these enabling digital resources by using cost-effective, cloud-based resources and engaging a vibrant â€œopenâ€• community. The tranSMART platform was initially developed by the pharmaceutical company Johnson & Johnson (JnJ). Open source code from the NIH National Centers for Biomedical Computing (NCBC) program was used as the basis for this effort. tranSMART was released into the public domain under an open source license to facilitate precompetitive pharmaceutical research[3,4]. The platform has since been deployed by more than 20 different organizations including academic medical centers, non-profit foundations and leading pharmaceutical companies. Test-bed instances are also undergoing evaluation at the Food and Drug Administration (FDA), overseen by its Office of the CIO. tranSMART is cloud-based and enables broad collaboration that will include the ability to implement privacy, de-identification, and security as required by HIPAA. The tranSMART Foundation is in active discussions with the â€œBionimbus Protected Data Cloudâ€• project (led by the University of Chicago) which already provides access to key NIH-funded data resources. Such secure cloud platforms will ensure appropriate IT security and privacy, facilitating secure and rapid access across global high-bandwidth optical research networks. tranSMART enables global collaboration and is the foundation for two large, international biomedical research initiatives. TraIT, a four-year â‚¬16M project, will facilitate the collection, storage,

analysis, archiving and securing of the biomedical data throughout Holland. It is a joint initiative involving 26 partners that includes the Center for Translational and Molecular Medicine (CTMM), Dutch Cancer Society, Dutch Heart Foundation, Netherlands Federation of University Medical Centers (NFU), Netherlands Bioinformatics Centre (NBIC), String of Pearls Initiative (PSI) and Netherlands eScience Center (NLeSC) [5]. The European Innovative Medicines Initiative (IMI) funded eTRIKS consortium (â‚¬24M, 5 years) aims to provide an open, sustainable research informatics and analytics platform for use by IMI funded projects[6]. eTRIKS partners provide support, expertise and services so that users gain maximum benefit from these platforms, and includes participation and investment from ten leading European pharmaceutical companies. Additionally, twelve tranSMART instances currently exist in Japan, leading the implementation of this collaborative platform throughout Asia. Efforts in the United States (US) have not benefited from resources and public backing as exists in Europe and Asia. A number of US-based tranSMART projects are being considered; a pilot demonstration project in personalized medicine is being developed by the University of Michigan in cooperation with the Johns Hopkins Medical School Brady Urological Institute. Other key US Academic Medical Centers are evaluating the platform in the context of their NIH NCATS Clinical and Translational Sciences Award (CTSA) activities. The Institute for Systems Biology (ISB) is evaluating the platform for its potential to share proteomics data and systems biology analytics in the public domain. Several major US, multi-national and European corporations are engaged in establishing the tranSMART Foundation and support its use in customer sites. It is envisioned that the US-based tranSMART efforts will scale to complement the European and nascent Asian efforts to enable a true global open data and analytics enterprise. Thus, the tranSMART platform can provide an accessible and appropriate infrastructure for organization, dissemination and utilization of publicly funded research data, enabling the collaboration and innovation that underlies the intent of the Obama administrationâ€™s initiative. Key challenges that â€œopen dataâ€• initiatives face include data quality, curation, integration and aggregation necessary for the development of rich repositories of biomedical information to support translational research and precision medicine. Available funding to support the entire data generation, management and analysis lifecycle is not sufficient to create and sustain these repositories. Federal policies with public funding matched by private contributions through effective public-private partnerships will significantly contribute to â€œopen dataâ€• ecosystems necessary to support future healthcare discovery. PPP initiatives such as the tranSMART Foundation, Bionimbus Protected Data Cloud collaboration and Open Cloud Consortium will benefit from federal policy and public funding that contributes to the â€œopen dataâ€• intent of the Obama Administration initiative. We request, as a part of the ongoing initiative to release digital information and data from publicly-funded research, that the Federal Government also provide financial support for the myriad efforts that are supporting the organization, dissemination and utilization of these

valuable public resources. These efforts are essential resources for increasing the liquidity of data and to enable and accelerate knowledge creation and practical benefits. Without an infrastructure that can be used to support and maintain these resources, they will not truly be available to the public in a manner that facilitates innovation. [1] http://www.whitehouse.gov/the-press-office/2013/05/09/obama-administration-releases-historic-open-data-rules-enhance-governmen [2] http://www.transmartfoundation.org [3] Clin. Pharmacol. Ther. 87: 614-616; advance online publication, April 7, 2010; doi:10.1038/clpt.2010.21 [4] http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-fda-final.pdf [5] http://www.ctmm.nl/pro1/general/start.asp?i=1&j=0&k=0&p=0&itemid=330 [6] http://www.imi.europa.eu/content/etriks

Victoria Stodden vcs@stodden.net Columbia University

Massive computation has begun a transformation of the scientific enterprise that will finish with computation absolutely central to the scientific method. From the ability to capture data, methods, create simulations, and provide dissemination mechanisms, science has gone digital. Convenient access to data and software is a necessary step in enabling reproducibility in computational science, and preservation ensures reproducibility persists. Openly available data and methods will maximize the downstream discoveries that could be made the information contain in the data and the know-how of methods contained in the code. This availability means curious STEM students, for example, can try their hand at replicating published results from the data and software, and learn about the science (and perhaps contribute further discoveries). It is not costless for a researcher to share the data they used in a discovering a published result. Depending on the dataset, there could be curation, reformatting, documentation, and other annotation to be done. There may be barriers to sharing such as privacy, confidentiality, and other legal impediments. It is important that the availability of the data persist over time, which is another expense associated with shared research data. I believe, however, these expenses are necessary to maintain and preserve the integrity of the scholarly record. Computational methods are typically very detailed and too often their communication is necessarily obscured in the published report simply be space limitations. For this reason discussions of data must include discussions of the code that takes the reader from the data to the final published results - ensuring reproducibility in computational science. With data and code unavailable for verification of the published results we are facing a credibility crisis in computational science. Data (and code) that underlies published computational results must be made available with the published paper, at the time of publication, as a default standard, if at all possible (subject to the limitations described previously). Data must be permanently linked to all publications that use it. We need stewards to care for and maintain these links with data, and the datasets over time. Versions for data must be carefully established. Corrections or updates in the dataset must not corrupt the link of the previous dataset with published results. As far as is practicable datasets should be made readily available for search, including descriptors, tags, keywords, and other metadata, and stored in as transparent formats as possible. This is not the place for proprietary or closed formats. This is also not the place for "gatewayed" or other for-profit data publication mechanisms, even if some instantiation of the original data is ostensibly made available for free. The ability to verify and reproduce results is the rationale behind open data in scientific research. It is not enough to rely on independent studies to verify results - what if the findings in the independent studies differ? Without open data we cannot reconcile differences between studies and find errors. Even though the notion of open data in science is based in reproducibility and understanding scientific findings, there are many corollary benefits to open data. These include the ability to use the data for other research which may be very different from the original

research the data were collected for. Other reasons may include further research, entrepreneurship, enhancement of existing proprietary methods, but none of these downstream uses should be able to close off the availability of the original data in any way. Data must be made as openly available as possible, subject only to the restriction that the source must be cited. I believe public funding should be set aside in order to meet these ends. This could enable researchers to complete documentation and other tasks required for data release. It could fund the development of public repositories (not those owned by private companies such as publishers, that I believe should not be stewards of our scientific culture. They have done an irresponsible job with published papers by restricting access). The notion of reproducibility scopes the question of sharing - make available the data and code needed to reproduce the results presented in the published paper.

Felice Levine **American Educational Research Association**

The American Educational Research Association (AERA) is the major 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. Founded in 1916, AERA as a scientific and scholarly society has long been committed to knowledge dissemination, building cumulative knowledge, and promoting data access and data sharing. The AERA *Code of Ethics* mandates data sharing and acknowledgement 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. They are also expected to cite data in their references to acknowledge data as scientific contributions and appropriately credit scientists. For more than 20 years, AERA under its Grants Program funded by the National Science Foundation (NSF) has fostered the use of federally supported data sets, especially those of the U.S. Department of Education's National Center for Education Statistics (NCES) and NSF. 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. Under this project, AERA now works with investigators of NSF-funded research on sharing and archiving data from completed studies on education and learning. In collaboration with the Inter-University Consortium for Political and Social Research (ICPSR), AERA is providing support and technical assistance to projects with potential for multi-investigator use and will be holding a small grants competition to stimulate use of these data. This initiative is directed to promoting data sharing and respectful, responsible use. AERA applauds both the principles and the objectives for public access to scientific data in digital formats. We also applaud the leadership role of the Office of Science and Technology Policy (OSTP) and key science agencies like the National Science Foundation (NSF) in promoting access to data through strengthened policies and data management plans. As emphasized almost 30 years ago by the National Research Council (NRC) in *Sharing Research Data* (1985), secondary analysis of extant data is essential to verification, replication, and new discoveries in science. At that time, the NRC commended various stakeholders, including federal agencies and scientific societies, to devise policies and plans for enhancing data sharing and use. One such early effort, driven in 1987 by the then NSF Division of Social and Economic Science, required grantees to commit to data sharing and archiving plans; in 1989, NSF specified a broad, agency-wide policy on data access and sharing. Since 2000 alone, the NRC has produced more than a dozen reports on expanding access to data and encouraging quality use consonant with protecting privacy and confidentiality. Also, ICPSR, now over 50 years old, has led innovations in access to useful data (including new forms of data), data preservation, appropriate use of confidential

data, and data citation. We offer our statement in this context to 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 in action requires implementation steps that can deepen and widen appreciation of the scientific value of data sharing, access, and use. We briefly offer comments to facilitate that end. Related guidance was provided in January 2012 in response to an OSTP request for information on Public Access to Digital Data and is available at http://www.aera.net/Portals/38/docs/Publications/2AERAResponsePublicAccessDigitalDataOSTP_FR76, No%20218,70176__1-12-12_.pdf. 1. 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 (e.g., interview protocols, measures, coding guides, and manuals). 2. Data Management and Sharing Plans already required by agencies like NSF are essential. Funds should be provided in awards to support archiving in data repositories to maximize data standards, access, and preservation. Renewal proposals should report on prior implementation. 3. To maximize meaningful access and contain costs, agencies should require use of data archives as the default and investigator- or institution-provided access as the exception. Agencies might offer a certified list of data archives with appropriate capacity and expertise. 4. Funds need to be provided to support repositories to expand their capacity to make accessible an expanded body of federally funded data; prepare for a larger, wider number of users; and innovate in mechanisms of access and use (as well as retiring data from use). The social sciences are fortunate to have a number of such repositories; ICPSR is the largest in holdings, innovation, and use. Fields of science with no or only limited repositories may need support to launch such entities. 5. Educational materials, webinars, or courses should be supported by science agencies, potentially 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. Also, funds for activities like the AERA/ICPSR data sharing project that enable investigators to implement data sharing plans can have high payoff and long-term impact for relatively modest cost. 6. 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. There are excellent materials from the NRC and federal statistical agencies on use of federal data bases and major federally funded data sets. An entity like the NRC might prepare a general guide for data sharing for federally

funded research. The guide could also address sharing proprietary data (and working out agreements for same), big data, or administrative records when access may be affected by privacy acts (e.g., FERPA or HIPPA). 7. 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, 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.

William Stall Histochemical Society

The Histochemical Society (HCS), long standing publisher of the *Journal of Histochemistry & Cytochemistry (JHC)* offers our support for the Office of Science and Technology Policy (OSTP) memorandum on Open Access publishing regarding implementation of a twelve-month post publication embargo period as a guideline for making research papers publicly available. HCS has continuously published over fifty years of JHC, highlighting basic cell and tissue research. In partnerships with librarians, researchers and our membership, we strive to offer the highest quality research at the most reasonable cost. We have long worked with librarians to make sure that their access to journal content was preserved for posterity and as a publisher on Stanford's HighWire, *JHC* has long made its content freely available to all, twelve months post publication. The twelve-month post publication embargo allows HCS enough time to recoup our expenses incurred producing the Journal. In turn, HCS hosts an annual meeting at Experimental Biology, offering travel awards to students and young researchers to attend Experimental Biology and present their research. In addition, HCS heads up the Immunohistochemistry and Microscopy Course, held each March at the Marine Biological Laboratory, in Woods Hole, MA. Without the monies from *JHC,* the Society would not be able to provide these opportunities to up and coming scientists for additional education. At the same time, because of the nature of *JHC,* the research remains applicable following the twelve-month embargo and our records show that usage is still high. We have calculated what would happen if the embargo was shortened to six months and extrapolated that HCS would be unable to continue publishing *JHC* because of the loss in revenue.Please recognize that small publishers and societies are major providers of research articles that benefit the economy, and early development of medical and scientific breakthroughs in the United States. Please help us continue to do so by keeping the twelve-month post publication embargo period for publishing research articles.

Jared Lyle Lyle@umich.edu Inter-university Consortium for Political and Social Research

The Inter-university Consortium for Political and Social Research (ICPSR), a research center and social science data archive in the Institute for Social Research at the University of Michigan, strongly backs the recent Office of Science and Technology Policy (OSTP) memorandum directing federal agencies to "develop a plan to support increased public access" to federally funded research, especially scientific data. For over fifty years, ICPSR has distributed and preserved data, as well as championed data sharing. As we stated in our response to the 2011 Request for Information on Public Access to Digital Data and Scientific Publications: "A general Federal mandate requiring grantees to archive scientific data for secondary analysis would promote re-use of scientific data, maximize the return on investments in data collection, and prevent the loss of thousands of potentially valuable datasets" (http://www.whitehouse.gov/sites/default/files/microsites/ostp/digital-data-(%23043)%20ICPSR%20Response.pdf). Maximizing public access to research data requires significant planning and foresight. Standards and guidelines are available to help, which we synthesize below. Specifically, we encourage federal agencies developing public access plans to make research data: 1. Discoverable -- Finding and accessing data requires metadata ("data about data") in standard, machine-actionable form. Metadata help search engines find and catalog data, as well enable researchers to perform detailed searches across data collections. In the social sciences, the Data Documentation Initiative (DDI) is an international standard for the description of data (see: http://www.ddialliance.org/). In the UK, the Digital Curation Center has recently created an inventory of disciplinary metadata standards at http://www.dcc.ac.uk/resources/metadata-standards. 2. Meaningful & Usable-- Access involves not just finding data, but also knowing how to use and interpret the data. Incomplete, incorrect, or messy data limit use and reuse. Proprietary or obsolete data formats can be unreadable or limit access. Repositories 'curate', or enhance, data to make it complete, self-explanatory, and usable for future researchers. This includes adding descriptive labels, correcting coding errors, gathering documentation, and standardizing the final versions of files. Curation is crucial to maximizing access. 3. Persistent -- Valuable research data deserve safekeeping for future researchers -- for replication and reuse. Preserving digital data requires much more than storing files on a server or desktop. Digital preservation is the proactive and ongoing management of digital content, with an eye toward lengthening the lifespan of the information and mitigating risks. Preservation actions are taken to guard against physical deterioration, accidental loss, and digital obsolescence. We also recognize that not all data are worth preserving indefinitely; less valuable or easily producible data may be preserved for shorter periods -- perhaps five to ten

years depending upon the scientific domain. 4. Trustworthy -- Data producers need to trust that the data they archive will be properly stored and shared, rather than lost, corrupted, or neglected. Data consumers need to trust that the data they receive is the original, unaltered version saved by the producer. The Open Archival Information System (OAIS) Reference Model, the Trusted Repositories Audit & Certification (TRAC) standard, which is now ISO 16363, and the Data Seal of Approval are standards that guide repositories in documenting and verifying that they are organizationally, procedurally, and technologically sound as data custodians. 5Confidential (when applicable) – A growing number of studies include sensitive and confidential data. Stringent protections must be in place to guard and provide access to these data. Robust methods, such as those promoted by the American Statistical Association (http://www.amstat.org/news/statementondataaccess.cfm), are in place for evaluating and treating disclosure risks, and repositories can offer technologies, including virtual data enclaves, for protecting and safely sharing confidential data. 6. Citable -- Properly citing data encourages the replication of scientific results, improves research standards, guarantees persistent reference, and gives proper credit to data producers. Citing data is straightforward. Each citation must include the basic elements that allow a unique dataset to be identified over time: title, author, date, version, and persistent identifier (such as the Digital Object Identifier, Uniform Resource Name URN, or Handle System). Some academic journals, such as the American Sociological Review, have already adopted a set of standards for citing data. An international consortium, DataCite (http://www.datacite.org/), strives to improve and support data citation. Our *Guide to Social Science Data Preparation and Archiving* (http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/) provides more details about many of these guidelines and standards, as do resources from our sister organizations, including the UK Data Archive (http://data-archive.ac.uk/media/2894/managingsharing.pdf). Finally, we note that providing access to and preserving scientific data can be expensive. We are encouraged that the OSTP memo allows for the "inclusion of appropriate costs for data management and access" in proposals, although we also wonder what existing, additional, or 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.

Robert Hauser rhauser@nas.edu National Research Council

I speak primarily as a social scientist who has created and used public research data with federal support for more than 40 years and not as a member of the staff of the National Research Council. I will make six observations, but offer no recommendations:

1. There is nothing new about public access to federally supported research data: a) OCG 1962 and 1973; b) Public Use Microdata Files from the Decennial Census, starting with 40-50; c) Wisconsin Longitudinal Study, almost all data public for 30 years; d) Uniform Edition of the October Current Population Survey; e) Dozens of social and biosocial surveys have been created for public use and provide templates for sound data policy
2. 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.
3. Access to federally supported research data will depend heavily on the responsibilities and actions of investigators and not merely on those of agency personnel. As in publication, that will cost time, effort, money – and require incentives for compliance. The infrastructure for data protection and dissemination is far better developed in some disciplines than others.
4. 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 also can and should be documented. For example, some journals now require authors to submit the code that was used in published analyses. 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 scientific. Look no further than the case of the Reinhardt-Rogoff paper. Will it be necessary to develop standards for documentation of data management as well as data files?
5. Some federal data are already well-covered by legislative or regulatory provisions that are highly protective, and in some cases unreasonably so, and 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. The user as well as the use might be considered. For example, commercial users of some kinds of data are now far less subject to regulation than are researchers.
6. Finally, with regard to data on human research participants, many of you are, no doubt, award of the ANPRM that was issued in the summer of 2011, yielded well over 1000 comments, and has not, as yet, resulted in new regulations 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, and an enlarged committee is now beginning work on a consensus report.