Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

This is a response from the UCLA Social Science Data Archive http://dataarchives.ss.ucla.edu/
Prepared by Libbie Stephenson, Director 310-825-0716 libbie@ucla.edu

The Data Archive has been in operation since 1977 and serves the entire UCLA campus of faculty and students engaged in quantitative research, including archiving original collections of data through surveys, and providing access for the re-use of de-identified (public use) data by faculty and students for research and instruction. Faculty members who are engaged in survey research use the Data Archive as a support to data collection processes and life cycle management, and in making the data publicly accessible. The Data Archive is a member of the Inter-university Consortium for Political and Social Research, an organization of over 700 members, worldwide, engaged in the use and preservation of data used in quantitative research. "ICPSR maintains a **data archive** of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields." UCLA is one of the heaviest users of ICPSR public use data, well above (approximately 3x heavier usage) the median usage by other institutions with the same Carnegie Classification. The re-use of public data at UCLA is vital to the generation of scholarship and social science pedagogy.

Thank you for the opportunity to respond to this RFI. We support efforts of the Working Group on Digital Data "to encourage and coordinate the development of [federal] agency policies and standards to promote long-term preservation of and access to digital data resulting from federally funded scientific research." We agree that all such policies should "follow best practices for protecting confidentiality, personal privacy, proprietary interests, intellectual property rights, [and] author attribution…"

Summary of comments:

In this response we suggest that Federal agencies should develop policies reflecting the nature of open access as it is understood today, but also make provision for an evolution of what open access will be given continued changes in technology and the need to be able to address open access from a global perspective. We provide some considerations about the nature of open access followed by comments on the questions posed. Some key points include:

- Open access has associated costs; data are not free; Federal agencies should allocate funds for long term data management.

- Open access data use requires skills and expertise; Federal policies need to acknowledge that open access does not preclude the need for the user to have the necessary technical, statistical and analytical skills necessary to work with research data effectively.

- Open access has boundaries when it comes to issues such as protection of privacy and confidentiality, national security, data embargos, copyright and intellectual property. Policies established by federal agencies should reflect these boundaries in ways that promote the widest possible access.

- Open access goals assume the existence of a robust infrastructure; however, there is no data stewardship infrastructure that currently exists anywhere in the world with the capacity to manage the amount of data generated. Furthermore, there are not enough trained members of the workforce capable of managing the masses and varieties of data being produced. There has been no scientific study of the strengths and weaknesses of current data stewardship organizations, operations, technological approaches, or repository software. The Trusted Repository Audit Framework, the Data Seal of Approval and other assessment tools should be widely applied.

- Policy development aimed at achieving open access to research data can best be accomplished with community involvement of all stakeholders: researchers, archival professional groups and scholarly societies. Evolution of policy should flow between stakeholders and agencies based on scientific study of research methods among disciplines; archival operations operating at international, national, regional, state and local levels; and of the variety of practices and standards used to manage data for the long term.

- Research proposals containing data management plans should be evaluated by experts in long term data stewardship, and by those with knowledge of best practices for organizing and documenting data. Compliance with data management plans can be verified through the use of registries of unique identifiers.

Some considerations about "open access"

Open access has associated costs

For many the idea of openness suggests no monetary cost. Digital data are not free. Their collection is funded through contracts and grants paid by taxpayers. The computing facilities (software and hardware), technology experts, administration, and data collection itself (such as through survey research centers) is funded partially by grants, but largely by academic institutions. The long term maintenance of digital research data is funded by universities and organizations; for example, the Inter-university Consortium for Political and Social Research (ICPSR) http://www.icpsr.umich.edu is a member-based organization where member dues contribute to data management. These costs are rarely addressed in a research grant and where funds are budgeted, they do not contribute to the long term sustainability of digital data. Therefore, we argue that funds for research data collection do not necessarily provide for free access to digital data. In order for this to be true, federal agencies that fund collection of research data must also allocate monies for the long term management of the data. Experts suggest that in addition to providing support for data preparation, agencies should provide sufficient funding (1-5% of an award) to guarantee long term management of open access data.

Open access data use requires skills and expertise

The term "open access" suggests the idea that data can be used by anyone, even by those who have no knowledge about the original data collection. In order for this to be possible, the use of federal funds for research should support best practices for organizing and documenting data. It is important to note that even with good documentation and publications, data are not necessarily usable by just anyone. Federal policies need to acknowledge that open access does not preclude the need for the user to have the necessary technical, statistical and analytical skills necessary to work with research data effectively. And, while it may be outside the scope of this RFI, it is worth noting that the American public is woefully statistically illiterate. We

encourage federal agencies to provide support to training and education programs so that users beyond the research community can make effective, informed use of digital data.

Open access boundaries

It is also important to consider that not all research data can be categorized as open access, either because of the sheer volume or complexity of the data, or for reasons of confidentiality, national security, embargo period and so forth. We concur with the views expressed by the American Association for the Advancement of Science "that the discussion surrounding public access must clearly distinguish between providing access to research results in support of scientific progress and access to scientific information as a crucial element of public engagement." That is, for some research, open access to underlying data may not be practical or possible; (e.g., the general public does not have the facilities required to personally analyze petabytes of data collected by astronomers), therefore access to information about the research, the analysis of the data, the conclusions reached, and so forth, through publications, may be the best option for providing the general public with usable information about research conducted using tax dollars. We encourage development of policies where these distinctions can be recognized.

Copyright and intellectual property rights boundaries can be addressed within an open access environment. The Creative Commons organization exists to provide a simple way for research data and results to be shared openly with all. http://creativecommons.org/ Creative Commons provides a set of copyright tools and options so that researchers can receive credit for their work and still share it as widely as possible. The result is a public arena where content can be shared, copied, edited, or reformatted, while still acknowledging the work of the original investigator. Depending on the nature of the data and the kinds of rights the investigator desires, a variety of boundaries of access can be established. We would encourage federal agencies to coordinate with or adopt processes to support a Creative Commons approach to issues of intellectual property and copyright parameters.

Open access goals assume the existence of a robust infrastructure

Currently only a very small amount of data collected with funds from Federal agencies is ever deposited for long term stewardship. In a white paper prepared by the International Data Corporation, IDC the authors noted that "In 2006, the amount of digital information created, captured, and replicated was 1,288 x 1018 bits. In computer parlance, that's 161 exabytes or 161 billion gigabytes… This is about 3 million times the information in all the books ever written." Further, the authors projected that "[b]etween 2006 and 2010, the information added annually to the digital universe will increase more than six fold from 161 exabytes to 988 exabytes." **There is no data stewardship infrastructure that currently exists anywhere in the world with the capacity to manage this much data.** Furthermore, **there are not enough trained members of the workforce capable of managing the masses and varieties of data being produced.** And, the repositories and archives that do exist operate in very different ways, employing widely varying definitions of what it means to provide data stewardship and what technologies are needed to achieve desired levels of curation. **There has been no scientific study of the strengths and weaknesses of current data stewardship organizations or operations.**

In developing policies and framework for stewardship of research data in an open access environment, agencies will need to invest in expanding and building upon the archival infrastructure that currently operates. Funding is needed for:

- Physical facilities, such as data centers for storage of large amounts of data
- Viable technological solutions for carrying out data stewardship processes for versions control, evolving file formats, discovery, and access

- Programs to assess the validity, competency, and track record of existing archives and repositories
- Programs to educate and train a skilled workforce

Physical data storage facilities should be funded to provide capabilities on a global level. Research is not only conducted from multi-disciplinary perspectives, it is also an international phenomenon; facilities for data generation, storage and access will need to take advantage of cooperative alliances with all stakeholders internationally. For example, investment in secure satellite storage and retrieval facilities is essential.

Given the volumes of data being produced, technological solutions to long term data stewardship are going to have a vital role in whether or not research data can be saved. Different kinds of data have different requirements for ensuring usability over the long term. In the social sciences, best practices for managing versions and file formats focus on techniques such as format migration and media refreshing. Ways of carrying out these processes mechanically are being explored by such groups as the ICPSR, the California Digital Library, and the Chronopolis/iRODS program, along with numerous academic institutions where smaller archives and repository software tools (such as Islandora or Stanford University's Hydrangea project) have been built. However, none of the new approaches thus far proposed or promulgated have been tested. There is no empirical evidence demonstrating that any one technical approach is any better than another. It is unclear whether micro-service or rule-based solutions are best and for which kinds of data are such systems most suited. There have been no assessments evaluating and comparing repository systems on a side-by-side basis to establish the strengths and weaknesses of each.

While there have been many claims made by many institutions that research data preservation and curation services are provided, none of them have been assessed by any standard criteria. Such standards have been developed, including the Trusted Repository Audit Checklist and the Data Seal of Approval. Open access cannot be accomplished until there is a careful investigation of the existing infrastructure, its capabilities, and whether or not organizations involved are achieving the desired goals for long term stewardship. Further, there is no agreed upon set of principles about which data needs to be maintained, in what way and for how long.

In a robust open access infrastructure, enabling workforce capacity is a serious concern. There are relatively few programs for training personnel to handle the variety of tasks involved in data management. Some look to the training provided through library and information sciences, but such programs are uneven and there is no established formal set of skills that all training programs should provide. None of these programs have ever been evaluated and there is no across the board accreditation process. Much of the time, 'training" involves reading numerous articles and reports and discussing issues from a philosophical or theoretical perspective. There is very little effort to ensure that new data archivists have the statistical, computer programming, or data science practical skill sets needed to perform in a variety of academic, private, and public agencies and firms. Current workforce capacity in commercial firms, banks, industrial occupations, medicine and physical and life sciences is lacking in the same skill sets. For open access to be a path toward innovation and economic strength a trained workforce is essential to such an infrastructure.

Responses to selected questions

Question 1
What specific Federal policies would encourage public access to and preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

Stewardship of data includes the development of detailed metadata, the long term curation of data, and technical mechanisms to enable discovery or and access to data. Ensuring that researchers and the public alike are able to find and use existing data will contribute to increased productivity and will expand the opportunity to increase knowledge. We suggest, along with the National Digital Stewardship Alliance that policies of funding agencies need to "go beyond the data management plan, and should explicitly recognize "data under stewardship" as a core indicator of scientific effort and include this information in standard reporting mechanisms." Broadly stated, researchers who document and ensure the long term management of their data should be rewarded when future applications for funding are submitted.

Federal policy should also focus on the data stewardship infrastructure to ensure that data are maintained by trusted digital repositories (For example, see Beagerie, et al. Trusted Digital Repositories http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf). Not all archival operations function equally. Federal funding programs should support existing archives and repositories to become certified as trusted repositories or to meet the disciplinary data stewardship requirements of the data being managed. Further, the existing archival infrastructure should be studied at the national, state, local and institutional levels to ascertain strengths and weaknesses and to establish funding streams to upgrade and re-engineer older archival operations to make use of and/or develop new data management tools, software, equipment and data management facilities. Finally a study of the skills and training possessed and/or needed by current data management professionals should be made and a program to develop and train new professionals should be established.

Question 2
What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

This response pertains to data gathered in social sciences research. It is currently the custom that data collected by social scientists belong to the universities at which they are employed. However, sharing of data and making it publicly accessible is expected. Projects funded by National Science Foundation (NSF) and National Institutes of Health (NIH) now require data management plans with a view toward open access and sharing of data. In order to protect intellectual property interests, researchers have a right to expect that users of their data will be properly cited and that the original researcher will be given due credit.

Professional organizations such as the International Association for Social Science Information Services and Technology (IASSIST) promote the proper citation of data. Several organizations have developed methods for providing unique identifiers for data. The plethora of possible uniquely identifying systems

is confusing to researchers who do not know the difference between a DOI, URN, universal handle or some other kind of registry. Federal policy could support the establishment and use of a single unique identifier and system for storing and verifying such identifiers, or for building interoperability among domain or discipline specific systems. Research into what the best system and mechanisms to accomplish this should be should be carried out.

Question 3
How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on management of data?

Agencies can work with discipline-specific professional archival and data management organizations for ascertaining best practices in each discipline. Many practitioners have established best practices within their disciplines for managing data. There are also best practices for different file formats, including text, image, video, audio, simulation, game, and so forth. Data management plans call for researchers to specify the kind and format of data they will be collecting. At the same time, research at academic institutions is increasingly inter-disciplinary. Federal policies and data management plans must take into account the possible hybridization of data produced in such activities. There is no one solution which will address data in every discipline. Still, policies can focus on some common data management procedures for authentication, for version control, for establishing unique identifiers, and for following standards for establishing authorship, author rights, and verifying trusted repository or archival entities. Any policies established will need to be flexible enough to adapt to changes in research data gathering practices, technological advances, and changes in best data management practices.

Question 4
How could agency policies consider differences in the relative costs and benefits of long term stewardship and dissemination of different types of data resulting from federally funded research?

An assessment of the costs and benefits of long or short term stewardship of data produced with federal funds should be part of the process of review when a grant proposal is first submitted. Each agency that provides funds for research has a set of review criteria. For example, the National Science Board has recently release new criteria for proposals submitted to the National Science Foundation http://nsf.gov/nsb/publications/2011/06_mrtf.jsp. While these criteria do not specifically address short or long term stewardship as review criteria, (though perhaps they should) the perceived intellectual merit and broader impact of the intended research can provide guidance on when and how to allocate resources.

Further, the ability of the investigators (or the specified archive in which the data will be eventually placed), to disseminate data in user-friendly public access modes should be considered; many agencies conducting their own surveys now provide ways to produce simple tables and visualizations while at the same time providing the research community with access to the underlying, complete data used to produce the tables, charts and reports. The same consideration could be applied to federally funded data collections. This provides both a short term solution to data access, and still allows for in depth research to take place.

Cost and benefit of long term stewardship can be considered from other aspects. For example, not all data needs to be kept indefinitely. Not all data needs extensive maintenance. In some cases, merely providing what is termed 'persistent access' will suffice, but in other cases the data may be considered to be so valuable as to need full data curation. This can be file format dependent. Data produced in software dependent formats will need more attention over time to ensure that the data can still be used even if the original software used to produce the data no longer exists. The cost to do so needs to be factored into the grant proposal budget, and should be considered by those who review such proposals. In some cases, the data collection and its management will require that a data archivist is embedded into the project at the very beginning. Doing so can result in a final data product whose long term stewardship is less costly than it would be if data management is handled at the very end of the project or even some years after the completion of the project. All of these factors should be part of an evaluation of a grant proposal, and such an evaluation should be carried out by those informed about long term data stewardship practices and institutions.

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

The kind of detail requested in data management plans asks researchers to think about data outside of an analytical context. Unless the researcher has frequently re-used data from previous studies they are unaware of the kinds of data management activities needed throughout a research project as well as once the project is completed. The preparation of a data management plan generally requires a consultation between the researcher and the data archivist with knowledge about stewardship of the type of data being collected and the research methods being used.

The stakeholders need to be able to advise the researcher about best practices for data and file formats, software and computing technology, metadata development, and on documenting their datasets. Stakeholders can best contribute to implementation of data management plans by having fully trained staff, by having certified data management operations and by possessing knowledge about the research applications of the data they are hoping to manage. Grant proposal reviewers will need to look beyond the data management plan itself to verify that the repository or archive selected is actually able to carry out the plan specifications.

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

There is no easy answer to this question. Many funding agencies allow for proposals to include data preparation costs which are intended, in part, to ensure that the resulting data product will be usable by someone not familiar with the original project. Agencies should continue to provide funds for the preparation of data collected with federal monies, but such support should also reflect the cost of ensuring long term access. When archives receive data in a well-documented form, with datasets built according to disciplinary standards, it is possible to carry out appraisal and ingest with low cost and the long term curation can be straightforward. The higher costs come from trying to manage complex, large volume, multi-format and poorly documented collections. Data produced in proprietary formats,

or software dependent formats also raises costs. It is important that these issues be raised at the grant preparation stage, within the data management plan and that proposal reviews include an assessment of the data management plan and proposed costs.

Question 7
What approaches could agencies take to measure, verify, and improve compliance with federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

The best compliance will come from developing the best plans. When researchers understand what is involved in data management during and after a project and are informed about the kinds of support and archival expertise available they will be better able to carry out their data management plans. When data archivists are involved in the development of plans, or are partners in the research there is greater likelihood of compliance.

One possible approach involves the assignment of a unique identifier to a fully processed set of research data. The unique identifier can be used as a way to *verify* that the data management plan has been carried out, and would be assigned by the archive or repository with responsibility for housing a researcher's data and reported to the funding agency. There are a number of such operations for obtaining a unique identifier, including the use of DOI's, URN', persistent identifiers and there are any number of locally designed registries. Federal agencies could require that researchers much use one certain system in much the way that publishers require bibliographic citations in specific order and format. The unique identifier accompanies the data and is referenced in publications and citations.

Question 8
What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

Enabling innovative re-use of existing data should be one of the core criteria used to evaluate data management plans. In order for this to happen, the data have to be discoverable, the usability needs to be assessable, the data have to be accessible, and the potential of the data has to be made evident. Data discovery mechanisms are searchable catalogs or directories containing complete citations to existing data. Users can assess the utility of data by being able to evaluate data content and format using searchable detailed levels of metadata. Data needs to be stored in platforms that make it easy to obtain copies for analysis on the user's own computing facilities. These features exist for a small number of domain specific archives, such as the Inter-University Consortium for Political and Social Research (ICPSR), the Data Preservation Alliance for Social Sciences (DataPASS) and in some academically based small archives or repositories. The vast majority of data produced is not managed by such facilities and is therefore not easily publicly accessible.

Ability to assess the potential of data for innovation requires a skill set and area of expertise commonly referred to as data science. This is an emerging occupation and there are not enough trained persons who can carry out this kind of work. Numerous commercial enterprises are searching for this kind of expertise; some of these companies spoke about this at the 2011 Web 2.0 Summit. For example, Bluefin Labs http://www.web2summit.com/web2011/public/schedule/detail/21613 has managed to

build what they call a "data genome" using data science technology and expertise. Scholars could advance their research if they were able to partner with such firms to contribute to and take advantage of investments firms make in data science based initiatives.

The need for data science expertise is also required in disciplines that collect or produce large quantities of data. Technologies such as Hadoop, used to pull together and mine big data exist but are not always easily employed by researchers and their use is as yet underemployed in business applications. Companies such as Microsoft are trying to bring these tools to the average researcher (http://www.web2summit.com/web2011/public/schedule/detail/21995).

There needs to be an effort by the Federal government to promote and enable more industry-university partnerships so that the unique research data being collected can be made more accessible, used in more innovative ways and that there is a newly trained set of professionals who can best stimulate the use of research data in innovative ways. Funding to establish data science training programs is needed. Funding to universities to implement and contribute to the development of data mining tools, in partnership with industry is needed as well.

Question 9
What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

There has already been considerable effort which is ongoing to provide standards for citing data and providing attribution. DataCite is an example of an internationally supported organization devoted to just this area of work. http://datacite.org/ Further, there are examples of how to create citations available from many professional organizations, archives and repositories. The issue is in getting researchers to use them. Many publishers now require a citation to data used in a publication as do scholarly societies with active publishing enterprises.

Question 10
What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Within the social sciences data community, a standard for describing the content of data files has been established through the Data Documentation Initiative (DDI) and this has evolved into an internationally accepted and implemented standard http://www.ddialliance.org/. DDI is not only a metadata describing standard, it also provides for machine actionable metadata enabling analysis of specific data elements across studies. Tools for employing DDI have been developed and continue to evolve; Colectica is an example http://www.colectica.com/ and there are numerous open source resources shared by the DDI community of implementers.

Question 11
What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

DDI was a social science community driven effort and was adopted through the use of training in using the standard, in forming working groups of practitioners, and in having the funds to continue development. There was a recognized need for the standard to enable archivists and researchers alike to organize, manage, promote discovery of and reuse data collected in social sciences research. Further there is an ongoing desire to promote interoperability among metadata standards so that data from many disciplines, not just social sciences, can be mined and used. The standard will be able to evolve to accommodate changes in research strategies and methods; initially DDI focused on a metadata standard for surveys but work is ongoing to accommodate research projects collecting qualitative information, and to accommodate different types of data gathering instruments.

Another example of a standard is the Data Seal of Approval (http://www.datasealofapproval.org/). The Data Seal of Approval is a set of guidelines that archives and repositories can use to provide depositors with some guarantees that the archive follows the best practices for long term stewardship of data. It has been adopted by organizations internationally, including the ICPSR. The Data Seal of Approval was developed by a single organization, but its utility has been recognized by the data community. Professional organizations such as the International Association for Social Science Information Services and Technology (IASSIST) http://www.iassistdata.org/ recognize and support the use of the Data Seal of Approval by its members.

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

Federal agencies can look to professional organizations which are international in scope for coordinating standards. Groups such as IASSIST (http://www.iassistdata.org/) address issues on data management, national policies on data use, data sharing, and data access. Members provide context for emerging policies in different countries; issues on data preservation and stewardship are discussed and standards for best practices have evolved through the organization.

Another internationally based archival effort was launched through the International Household Survey Network (IHSN) working in collaboration with the Accelerated Data Programme (ADP) (http://www.ihsn.org/adp/) This work is designed to help countries develop their data and statistical programs according to standards and best practices and to promote the use of their data. These newly formed archives use the DDI standard for metadata, and as described in their website "provide technical and financial support to survey data documentation and dissemination, and to the improvement of survey methods. Key outputs include the establishment of national survey databanks, and the establishment of national data collection standards to foster comparability of data across sources." Federal agencies could promote and finance similar programs within the U.S. and its territories at the state and local level.

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

Work to enable data access from publications that report on data use and analysis is ongoing; organizations such as the ICPSR are engaged in working with publishers to provide links from publications to data managed at ICPSR. The National Digital Stewardship Alliance has taken a leadership role in promoting collaborative stewardship and standards with respect to linking of data to publications.

The issues to be addressed in making such linkages more universal have to do with being able to maintain version control, to be able to update or correct datasets, and the archive or repository must be able to provide access indefinitely. The use of unique identifiers can aid in version control but publishers will need to be able to accommodate updated material. Requirements to properly cite the use of data provided with publications need to be made; investigators must be given credit not only for their publications but also for the data they produce and share.

The indefinite access to data in a form that is usable despite the passage of time requires that the archive operate according to the best practices for the data format and discipline of the data being managed. Repositories and archives need to be assessed according to criteria for evaluating the ability of the facility to maintain and provide access to the data. Examples could include the use of the Trusted Repository Audit Checklist (TRAC) based on the ISO 16363 Standard for Trusted Digital Repositories and the Data Seal of Approval. Data management plans could specify that publications and data will be linked and that data will reside in a verified archive or repository. Further, such repositories should be able to demonstratethat they have a valid plan for succession in the eventuality that the repository ceases operation or experiences a disaster. For example, the Data Preservation Alliance for the Social Sciences (DataPASS) provides a complete succession plan for the management of all data shared by alliance members.