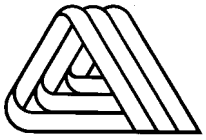

American Educational
Research Association

**Response to Request for Information (RFI): "Public Access to Digital Data Resulting from Federally Funded Scientific Research," Office of Science and Technology Policy (OSTP)**

**76 Federal Register 218, pp. 70176-70178, November 10, 2011**

**American Educational Research Association**
**Felice J. Levine, Executive Director (flevine@aera.net)**

**January 12, 2012**

## About AERA

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.

For more than 20 years, AERA under its Grants Program 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 the National Science Foundation (NSF). This long-term project has led to important scientific discoveries and methodological advances and has contributed to a culture of building scientific knowledge cumulatively through analyses of such data. In 2009, with continued support from the National Science Foundation, AERA expanded its efforts and is now working with principal 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 in data archiving to projects with potential for multi-investigator use and will be holding a small grants competition to stimulate use of these data. Through this initiative, AERA is actively engaged as a leader and partner with a federal agency (NSF), the world's largest archive of social science data (ICPSR), NSF research investigators, and potential scientific users on a model project directed to nurturing and promoting the advantages of data sharing and respectful, responsible use.

Because of its special interest in data sharing, AERA has been at the forefront among research organizations in promoting and exploring ways to promote data sharing. AERA engenders this culture within its own policies, procedures, and practices. The revised AERA *Code of Ethics* adopted in February 2011 mandates data sharing and appropriate acknowledgement of data use and takes account of the potential for data use under restricted access provisions that may be necessary to protect privacy rights and the confidentiality of information. (See Appendix A to this response.) Authors in AERA journals and other publications are guided to cite data in their reference lists so as to acknowledge data as contributions in their own right. And, in AERA's 2008 NSF-funded study of education research doctorate programs in U.S. universities (being undertaking in collaboration with the National Academy of Education), a data archiving and data management plan were an integral part of the submission.

In word and deed, AERA strongly supports the goals that led to this RFI and is pleased to share its perspective on the questions posed by OSTP. We shall do so by responding in question order.

**Response to RFI Questions**

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

Federal policy must promote not just the production, sharing, and preservation of digital data, but also the sharing of data collection instruments, the production of transparent and comprehensible metadata, the health of data repositories and similar institutions, and the development of software that will foster or enable data sharing. In addition, federal policy on digital data sharing should not be limited to federally fund "scientific research" but should extend to federally funded "scientific data" of all sorts, with "scientific data" being broadly defined. This is particularly important in the area of education where administrative and other operational data are often routinely collected and can be of great value in advancing learning and education science. We suggest that the following policies would contribute to the RFI's data preservation and access goals:

- There should be a strong presumption that scientific data collected with federal support along with relevant instruments and related metadata will be preserved as specified in a data management plan and made accessible to others. Strong consideration should be given to data archiving requirements as the most effective and efficient way of promoting data use, ensuring data preservation, and ultimately creating a culture of inquiry that values and acknowledges data products and their use. While a presumption of data sharing should not be absolute, any limitation should be very narrowly defined and carefully scrutinized in advance as part of an agreed-upon data management plan, and every effort should be given to the

feasibility of data sharing under restricted conditions or after passage of time even when there may appear to be substantial privacy and confidentiality concerns.

- A data management plan should be a part of federal grants and contracts that fund data collection, and evaluation of that plan should be part of the review process. The cost of data management should be regarded as an essential cost of the research and be evaluated for adequacy and reasonableness along with other proposal costs. Special justification and guarantees of future accessibility should be required if the data management plan does not include data archiving requirements in a publicly accessible research data repository.

- Data repositories, like the ICPSR, that archive and disseminate data from multiple sources greatly facilitate data preservation and sharing. As data sets have become more numerous, larger and more complex, such repositories are likely to prove necessary for any data sharing system to work. Federal policy should foster the development of data repositories, working to improve and sustain them in fields where they now exist and to create repositories for fields that currently lack them. Federal funding should be available to these ends as institutional start-up assistance and as grants or contracts to support innovations in data acquisition and dissemination technologies and procedures, including research on issues that affect the data sharing enterprise such as protecting subject privacy and ensuring that data uses are in accord with informed consent. In addition, support of ongoing operations is desirable, perhaps as a function of the amount of archived federally funded data and its usage rates and/or as add ons to grants to be used to pay fees for data archiving and dissemination services. Archiving data with approved repositories can be further encouraged by providing that depositing data is sufficient to meet a data provider's responsibility for ensuring that subject privacy, confidentiality, and informed consent interests will be sufficiently protected as the data are stored and disseminated.

- Data management should be recognized as a scientific profession in fields where it is not now adequately recognized. Federal policy can support this by supporting data manager education and research in ways similar to the support that it provides students pursuing education and careers in other science fields. Federal policy and funds could also support meetings of data repository managers and others involved in data archiving and dissemination to ensure that their treatment of data and metadata is mutually compatible and to maximize the feasibility of working with data drawn from different repositories.

- The demands and challenges of data management, including archiving and dissemination, change regularly as new technologies develop and new policies, like revised privacy rules, are put in place. Federal agencies should be encouraged to contribute funds for periodic National Academy of Science studies such as the National Research Council's 2005 report, *Expanding Access to Research Data:*

*Reconciling Risks and Opportunities,* or the 2009 report on *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age*. Also, investments in initiatives like the NSF-funded AERA/ICPSR project that enables investigators and their teams to implement plans for data archiving and use or address challenging issues can have high payoff and long-term impact for relatively modest cost. State of the art and consensus conferences centering on issues of standards for data and metadata and consistent forms of data citation are also important short-term priorities and could usefully inform further federal policy on data sharing, access, and preservation.

- Where data are collected by a federal agency the data should be available for sharing to the widest extent possible as determined by federal law and administrative rules. Some agencies, in particular some federal statistical agencies like NCES, have been leaders in this effort for quite some time, but this issue is worthy of consideration federal-wide. In some instances, it may be appropriate for federal agencies to form their own plans and systems for access to digital data, but the existence of such plans and systems should not preclude making all or portions of federally collected data available in data repositories even if they can be also acquired from the government. Indeed, multiple systems of availability should be encouraged since this broadens access and different providers may create tools that make working with the data easier and more cost effective.

*(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?*

First, the intellectual property interests of all involved should be specifically defined. The definition should make clear that, if scientific data have been collected with federal funds (unless the conditions of funding provided specifically to the contrary), those data and related metadata are in the public domain. That said, data collected and prepared by researchers are a product that merits and deserves appropriate credit by those engaged in their use. Citation of such data with appropriate attribution should both facilitate tracing advancements enabled by data resources and also offer a vehicle for giving credit and measuring interest and use.

Scientists understandably want to have the opportunity to analyze the data and make contributions to knowledge that follow from their conceiving of the project and the data collection effort. Such use is an appropriate incentive and reward for engaging in data collection. The time period for exclusive use should be specified as part of a data management plan and be approved or modified by the federal funder. Depositing the data in an archive for dissemination need not and ordinarily should not await the end of the exclusive use period so long as the dissemination of the data to others is embargoed until the exclusive use period has expired. Creators and licensees of works that use the

data, such as reports or articles analyzing the data, should have the usual intellectual property protections for the products of their use unless the conditions of funding the research or licensing the work provide otherwise.

Problems may exist when collected data are, like some business data, proprietary or the intellectual property of the person or entity supplying the data. When this situation exists, the rights of the data provider must be recognized and protected by the data collector even if the collector has no property rights in the data. In such circumstances the person or entity collecting the data should attempt to draft an agreement that calls for the widest data sharing arrangements that the data provider will agree to, and should propose confidentiality, data source masking, or other arrangements to facilitate sharing. In making funding decisions, federal funders determining the likely value of the proposed research should take into account the likelihood that the intellectual property or proprietary rights of data sources will limit reuse of the data.

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

The basic scientific principles and value of data sharing may vary less by discipline and field than it might seem, although in practice the levels of experience and exposure to data sharing and archiving vary. Thus, much of the challenge may be less in the policies derived than in the federal plan for investment in implementation and education so that sharing approaches are specified consonant with the different forms of data and methodologies for data acquisition. As social/behavioral scientists, education researchers collect and use the full spectrum of data at the individual, social, and institutional levels employing systematic and rigorous methods for data acquisition and analysis. Increasingly in fields like ours, there is growing use and attention to physiological and biological information, and we can anticipate both greater use of biomarker data in the social/behavioral sciences and greater use of social/behavioral measurements in the biomedical and biological sciences. Since many of these data collections are large-scale and longitudinal, the discussion of data sharing and access has already commenced (see, for example, the 2010 NRC report on *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biadata*).

When it comes to establishing policies regarding data sets that integrate various forms of human and organizational data, federal agencies have considerable social and behavioral science resources to draw on. In setting data sharing and management policies, they should be encouraged to utilize and build upon this expertise end experience. First, there is the ICPSR, now 50 years old and a pioneer in gathering and facilitating the further use of data that has informed original research across social science fields. ICPSR has had to confront such issues as obtaining usable data, data security, subject privacy protections, allowable use, legitimate access limitations, and requirements for metadata among other matters. They have considerable wisdom to

share and their experience will be highly instructive. Second, there are in the social and behavioral sciences a number of important longitudinal surveys whose purpose is to provide data to broad user communities, often giving no user priority to the data collectors (e.g., the Panel Study of Income Dynamics, the National Longitudinal Study of Adolescent Health, or the Health and Retirement Survey). Their experience, particularly in providing data in transparent user friendly forms can also inform federal policies and requirements. Third, there are professional associations, like AERA, that have given considerable attention to issues of data sharing as a professional responsibility. Their consideration of these issues can inform federal agencies of the data sharing behavior that is coming to be regarded as normative within our fields and can provide a professionally acceptable starting point for establishing data sharing policies. Fourth, there are federal agencies like the National Science Foundation (NSF) and the National Institutes of Health (NIH) which have established data sharing policies that can serve as models for other agencies. The obligation to share data collected with NSF funds and the commitment to a data management plan, although only recently reaffirmed and formalized as an agency-wide requirement, was initiated as far back as 1987 in the then NSF Social and Economic Science Division and 1989 in an NSF policy statement on data access and data sharing. Finally, there is an extensive body of knowledge through books, articles and reports, like the NRC Reports mentioned above, that deal with issues relating to justifications for, problems posed by, and the mechanics of social and behavioral science data sharing. This learning should be consulted.

We do not know whether similar resources exist to aid agencies in establishing data sharing policies for other kinds of data, but to the extent they do exist we believe federal agencies should take advantage of them. We also believe that federal agencies should eschew "one size fits all" policies when establishing rules and recommendations for data sharing. Instead, the characteristics of the kinds of data collected and used in different sciences should be examined, and policies should reflect the experiences and knowledge specific to different disciplines and fields.

*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?*

We are not clear on whether this question is concerned with the mechanisms for considering the cost-benefit issues it raises or with the standards and policies that should be applied. However, in either case consulting the sources of aid and information identified in our response to Question 3 is advisable. We would add that the cost-benefit issues raised by this question are important. The fact that a data set is created with federal funds does not necessarily mean that it is worth preserving and sharing over the longer run. At a minimum, however, data should be preserved and made available to others until a reasonable period of time after publications and reports drawing on the data have appeared. This allows for verification of results, examination of alternative hypotheses and questions, and consideration of these data to address

other issues or problems. Beyond the overall value of data preservation, judgments regarding data retention and dissemination should be based on the quality of the data, the range of issues that the data address or could address if combined with other data, the importance of the issues the data address, and foreseeable future uses of the data. In addition, when data have been available for some time in a publicly available repository, the demand for the data should be taken into consideration in any data culling decisions. We suggest that the federal agencies that fund data collection will most often be in a poor position to judge these issues, and that these determinations should be delegated to scientific advisory committees on data acquisition and retention attached to different approved data repositories and similar organizations. In addition, we suggest that no data should be removed from an archive and/or made unavailable to researchers if an agency wants the data retained and is willing to pay the costs of retention. We believe that any group determining what data should be kept should begin its evaluation process with a presumption that favors data retention. And even data that appears of little current value, as might be the case if it has been exhaustively analyzed, should be retained if there appears to be a real possibility that future access to the data will prove scientifically important.

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

Representatives from stakeholders like these can serve as participants in conferences and on advisory bodies when issues regarding data retention and dissemination policies and practices are being raised. It should, however, be recognized that in many instances the views of those who represent such organizations will be the views of consumers of data services and not the views of data stewards or management experts. In addition, research communities, universities, and research institutions can contribute to the implementation of data management plans by making clear to their members and/or employees that such plans should be a routine part of any grant application and that there is value in working with data repositories in providing data for further use. In particular, professional associations, research institutions, and universities should support the ethical standards of data sharing and responsible use consonant with human subjects research protections. Publishers may contribute by cooperating with journal editors to establish policies for citing data by a persistent digital identifier, such as a digital object identifier (DOI), thus encouraging data management plans that call for the prompt contribution of data to repositories. Finally, universities and research institutions could do more to recognize the scientific status and role of data managers.

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

Costs of preserving digital data and making the data accessible should be an accepted or even a required element of any grant or contract seeking federal funds which includes a

data management plan. Federal support might also be made available to recognized data repositories, perhaps in relation to the amount of federally funded data stored and demands for the data's use. In addition, special grant programs might be developed to support research aimed at innovations in data management technologies and practices as well as seed money or ongoing support for a consortium of data repositories and for technological linkages and periodic conferences that facilitate communications among different data managers and data management organizations. Finally, support in the form of educational grants and internships would indirectly reduce the costs of preserving data and making them accessible by expanding the pool of those with careers dedicated to these ends.

*(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?*

Agencies could require that final reports on grants and contracts demonstrate compliance with federal data stewardship and access policies, and could provide that compliance is demonstrated by providing a persistent digital identifier showing that the data together with relevant metadata have been deposited in a recognized data repository. Problems might be expected, however, because final reports in many instances are likely to be due before it would be expected to deposit data. In such cases, final reports could be accepted conditional on the filing of an addendum demonstrating data policy compliance by a set date.

It might also be appropriate to recognize data repositories available and appropriate for data archiving and preservation. One possibility would be establishing broad federal-wide standards for recognized repositories coupled with a requirement that any repository that wished to be so recognized file a statement with the agency that the standards are met. We suggest that the determination of standards of adequacy and revisions in them be made in the first instance by a group of those organizations that can today be recognized as well-functioning data repositories, with the federal role limited to endorsing or rejecting the standards. As new repositories are recognized, the group setting standards for recognition could expand accordingly.

*(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?*

Anything that increases the availability of data and ease of access and use can be expected to stimulate the use of publicly accessible research data in new and existing markets. To this end federal agencies could (a) support the development of clear and consistent standards for metadata and require those data collectors it funds to meet these standards as part of their data management plans; (b) support for certain important and complex data sets the development of software tailored to the data set

to make it easier for users, including the relatively unsophisticated, to find relations of interest to them, and (c) create or support the creation of a partially catalogued and searchable library of all data preserved in data repositories that have associated persistent digital identifiers.

*(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?*

Some mechanisms, already discussed, for example, in our response to Question 5 above, must be left to the private sector, including universities and professional associations, but federal agencies could call relevant issues to the attention of non-federal organizations and encourage them to consider steps that promote appropriate attribution and credit. Federal agencies could, in addition, set a valuable example by attaching citation information and persistent digital identifiers to data they create and make available and by requiring those who submit grant or contract proposals or who report on their scientific activities to cite sources of data referenced in them and provide persistent digital identifiers if available.

*Questions 10-12 require technical knowledge that others are better positioned to provide than we are.*

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

We have discussed many of the policies, practices, and standards that are needed to support this linkage in our responses to the questions posed above. To summarize, policies and incentives are needed so that data sources will be cited and, when cited, will be accessible to those interested in examining the data more closely. This means that data must be available, preferably in publicly accessible data repositories, and that the data must have associated with it citation information and a persistent digital identifier that encompasses relevant metadata along with the data. The federal government can lead both in its own data management practices and in the standards it sets for its grant and contract seekers. In addition, through the sponsorship of conferences or other means, the federal government can encourage professional associations, publishers, universities and research institutions to adopt as a matter of policy or professional ethics those standards that will promote the linking of publications and associated data.

# Appendix A

**Excerpt on Data Sharing**
***Code of Ethics***
**American Educational Research Association**
**(Code of Ethics Adopted by AERA Council, January 2011;**

*14.06 Data Sharing*

(a) Education researchers share data and pertinent documentation as a regular practice. Education researchers make their data available after completion of the project or its major publications for verification or other analyses by other researchers, except where proprietary agreements with employers, contractors, or clients preclude such accessibility or when it is impossible to share data in any useful form.

(b) In sharing data, education researchers take appropriate steps to protect the confidentiality of the data and the identity of research participants. When appropriate future use necessitates access to identifiable data, researchers take steps to ensure that the data are accessible under appropriate restrictions where the confidentiality of research participants can be secured. See also 12.04(b) and 12.08(c).

(c) Education researchers anticipate data sharing as an integral part of a research plan whenever data sharing is feasible.

(d) Education researchers share data in a form that is consonant with research participants' interests and protect the confidentiality of the information they have been given. They maintain the confidentiality of data, whether legally required or not; remove personal identifiers before data are shared; and, if necessary, use other disclosure-avoidance techniques. When data are shared with personally-identifiable information, education researchers take steps to ensure that access is provided only under restricted conditions where users agree to protect the confidentiality of the data consonant with prior commitments.

(e) Education researchers who do not otherwise place data in public archives keep data available and retain documentation relating to the research for a reasonable period of time after publication or dissemination of results and share data consonant with 14.06(a).

(f) Education researchers who use data from others for further analyses explicitly acknowledge the contribution of the initial researchers.