ICPSR Comments on Public Access to Federally-Supported Research and Development Data

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.

5. Confidential (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.