

American Statistical Association

Promoting the Practice and Profession of Statistics

732 North Washington Street, Alexandria, Virginia 22314 USA
(703) 684-1221 • Fax: (703) 683-2307 • Email: asainfo@amstat.org
Web site: <http://www.amstat.org/>

January 12, 2012

Interagency Working Group on Digital Data
National Science and Technology Council
Office of Science and Technology Policy
Executive Office of the President
Washington, DC 20502

Dear Working Group Members,

The American Statistical Association and its Committee on Privacy and Confidentiality appreciate the opportunity to offer the attached comments on the request for information, “Public Access to Digital Data Resulting from Federally Funded Scientific Research.”

As background, the American Statistical Association (ASA) is the world’s largest statistical society, with over 18,000 members in some 90 countries (though most are in the United States). One of its core missions is to advise government on matters related to data-centric research and policy-making. The Committee on Privacy and Confidentiality is an appointed group of ASA members with expertise in the technical methods and policy issues related to data access and confidentiality. The Committee members who endorse the comments in this letter includes:

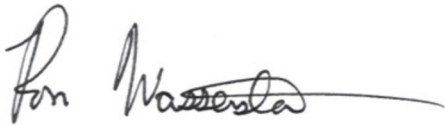
Jacob Bournazian, MA, JD	Energy Information Administration
Julia Lane, PhD	Committee Chair
Krishnamurty Muralidhar, PhD	University of Kentucky
Aleksandra Slavkovic, PhD	Pennsylvania State University
Lance Waller, PhD	Emory University
Simon Woodcock, PhD	Simon Fraser University

The Committee, and the ASA more broadly, would be delighted to share our expertise with the Interagency Working Group on Digital Data.

Sincerely,

A handwritten signature in cursive script, appearing to read "Julia Lane".

Julia Lane
Chair, ASA Committee on Privacy and Confidentiality

A handwritten signature in cursive script, appearing to read "Ron Wasserstein".

Ron Wasserstein
Executive Director, ASA

We applaud the OSTP initiative to enhance the availability and preservation of digital data. Such access is consistent with the principle of reproducible research, which we believe is vitally important to the scientific and policy-making process. Such access also opens opportunities for new insights based on existing data.

Preservation of data is equally important, since it is impossible to know what currently collected data will be useful for future generations of scientific inquiry.

We know that the OSTP recognizes the importance of preserving data subjects' confidentiality in making digital data access policy. The American Statistical Association's Privacy and Confidentiality subcommittee wishes to offer the following general suggestions on confidentiality issues related to the OSTP initiative.

1. Research has shown that many data subjects are skeptical of government agencies' abilities to protect confidentiality. A policy that greatly increases possibilities for confidentiality breaches could severely weaken the ability of both researchers and of government agencies to collect data. Researchers are subject matter experts in their own area of research, not in data dissemination and confidentiality protection hence the development of sound policy requires input from experts in these areas. The statistical community can contribute to ensuring both that data confidentiality is protected and that research subjects understand what is being done to protect the confidentiality of the data.
2. Researchers in confidentiality protection methods often distinguish two general classes of methods. Restricted access limits who can use the data, for example via secure data enclaves, remote access, or licensing.

Restricted data limits what data are made available, for example via data suppression, aggregation, swapping, or simulation. Each approach has a purpose. Restricted access is arguably the best solution for purposes of reproducible research and data preservation. It is also the best approach for complex data (e.g., relational or high dimensional data), which are difficult to protect adequately without degrading the usefulness of the data.

3. For restricted access, recent cyberinfrastructure advances have led to the development of data repositories that are managed by national data producers -or contracted to non-government parties--which investigators can use for data dissemination and preservation. These repositories can be staffed by data dissemination professionals and advised from committees of experts on data confidentiality practice.

Such remote access systems have been developed at the NORC/University of Chicago data enclave in the US, the Secure Data Service in the UK, the Microdata Online Access (MONA) system in Sweden, and the remote online system at Statistics Netherlands. These enable users to analyze the data at their own computers; however, users cannot save or print data locally.

When coupled with vetting and education of data users, as well as penalties for misuse, such systems can provide access while minimizing risks.

4. For restricted data, the ASA subcommittee has commented on recent HIPAA revisions to recommend against the adoption of safe harbor type standards. Put simply, each dataset has unique disclosure risks that cannot be captured with a list of common prohibited identifiers.

The subcommittee also wants to provide specific feedback on the following questions.

Question 1) What specific federal policies would encourage public access to and the preservation of broadly valuable digital data...

Users need a centralized location for searching available data bases. Websites such as Data.gov are possible choices where federal agencies can place research data that is publicly available from their agency website. The metadata supporting the files would be standardized and there could be a single path through the "Raw Data" for viewing what digital data may be accessed. Also, federal agencies need to reach consensus on standardized coding to use in the front section of the URLs that link to digital data so that researchers can easily identify digital research data files and can direct their search queries appropriately.

Federal agencies should consider reaching consensus on some criteria for setting time limits on the sensitivity of certain categories of information when products or procedures are no longer produced or applied, or structural changes occur within an industry due to changes in regulations or the application of new technology. Not all information that is protected at the time of collection needs to remain protected for decades into the future.

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...

Authors of published research should provide a proprietary notice on all work that is publicly released. Federal agencies should provide researchers with the capability to claim copyright protection in their work that is federal funded. Many times, federal agencies receive FOIA requests from the public for this information as a method for gaining a copy of the research and the releasing agency cannot impose or enforce any protections for the author if the authors do not claim any proprietary protection.