# Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

March 5 2020

The American Statistical Association (ASA) is pleased to provide comments in response to OSTP's Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research , as invited in the *Federal Register* of January 17, 2020 (85 FR 3085).

ASA's comments were written by members of the ASA Committee on Privacy and Confidentiality and are found on the following pages.

Thank you for your consideration.

*Questions on this document can be directed to the ASA Director of Science Policy Steve Pierson, pierson@amstat.org.*

Authors (from the American Statistical Association, Committee on Privacy and Confidentiality):
Lars Vilhuber, Cornell University, Member
Stefan Bender, Research Data and Service Center of the Deutsche Bundesbank, Member
Frauke Kreuter, University of Maryland, Member
Stephanie Shipp, Biocomplexity Institute, University of Virginia, Member
Aleksandra Slavkovic, Penn State University, Member
Tom Krenzke, Westat, Chair

The authors are responding in their capacity as members of the American Statistical Association's Committee on Privacy and Confidentiality, and are not representing their respective home institutions. They serve voluntarily and without remuneration on this Committee. The Committee's role is described on its website https://community.amstat.org/cpc/home. Relevant for our response to the RFC, the Committee has the charge:

- To monitor and encourage new technical developments related to privacy and confidentiality of data collected or used for statistical purposes.
- To develop appropriate liaison with Congressional Committees and Federal agencies on matters relating to privacy and confidentiality.

The authors come from a variety of disciplines in addition to statistics. They have degrees in sociology, economics, and in their various positions, have experience in creating, managing, and expanding research data centers holding confidential research data, and providing secure, unbiased, controlled access to these research data.

In our response, we will focus on the privacy and confidentiality aspects of the proposed repository characteristics. We draw on examples from the United States, Canada, Germany, and the United Kingdom.

In particular, we will respond primarily to questions of access (I.E.) ease of access (I.F.), fidelity to consent (II.A.). We consider that II.B-F. are not fundamentally different from the overarching question of access (I.E.), and that II.I. (request review) is a variant of I.F. We have additional comments on documentation of privacy (I.I.), and on the availability of metadata (I.C.).

*I.E. Access.* The suggested criteria require "broad, equitable, and maximally open access to datasets," moderated by privacy and confidentiality considerations. We note that there are many considerations why privacy and confidentiality considerations might apply, not just fidelity to consent for human data (II.A.) and compliance with restricted use conditions for human data (II.B.). Additional confidentiality considerations include financial, company, biogenetic, and national security considerations in the domains of biology, nuclear physics, engineering, to name a few. When federal funds are used to support research that use, analyze, generate, or produce such products, safeguards and access restrictions also need to be imposed. These are not fundamentally different from those for human data. To reprise (Desai, Ritchie, and Welpton 2016)[1], in all cases, repositories must need to assess whether access satisfies appropriate criteria

[1] Desai, T., Ritchie, F., and Welpton, R. (2016). "Five Safes: designing data access for research". *Bristol Business School Working Papers in Economics.* All URLs in this document were last consulted on March 4, 2020.

along five dimensions (the "Five Safes"): *Safe projects* (Is this use of the data appropriate?), *safe people* (Can the researchers be trusted to use it in an appropriate manner?), *safe data* (is the disclosure risk in the data appropriate for the purpose?), *safe settings* (from where and how is the researcher accessing the data?) and *safe outputs* (are the published outputs appropriately protected?). These five dimensions can be usefully applied to data ranging from full public use data (freely downloadable without need for any controls) via medium-security data (released to researchers under enforceable data use agreements) to highly classified data. They should thus be criteria applied by and for all federal funded repositories.

# I.F. Ease of access

Where necessary, access restrictions must be imposed. At the same time, repositories should leverage and implement the broadest possible set of tools to make access as easy as possible. The gold standard in terms of ease of use remains public-use data in the public domain, available for direct download, and with few if any use restrictions.

Clearly, when access is subject to some level of control, ease of use must necessarily be reduced. For instance, in the simple case where registration is required to ensure that users agree to terms of use, various access mechanisms can be implemented. Repositories should strive to allow for seamless access using both human and machine-initiated tools. The UK Digital Economy Act of 2017 enshrines a principle of proportionality.[2]

For instance, users could register once, agree to terms of use, and then obtain an access token which allows them to initiate future downloads from the same provider via an API using machine-initiated (automatic) downloads, while still complying with all terms of use. This is standard in many other common situations in the private industry, but is less frequent amongst current repositories.

Similarly, current restrict-access research data centers – a form of repository with access controls – require users to go through user vetting ("safe users") for every repository afresh, without reference to prior vetting at other repositories with similar or identical criteria. For a given repository, project vetting ("safe projects") for a user's multiple projects happens independently every time, without reference to prior projects. Furthermore, current repositories are often separated into distinct "data silos", where data sits in distinct repositories, and data that is primarily hosted at one repository cannot be also accessed at a separate repository. This is still generically true at the federal level, despite progress under CIPSEA (Title V of the E-Government Act of 2002, PL 107–347[3] and Title III of the Evidence Act of 2018, PL 115-435[4]). Impediments are also the norm for federal-state data sharing, and for government-private or government-academic data sharing. Though such data sharing across repositories occurs on a regular basis, each one is subject to laborious ad-hoc re-negotiations.

---

[2] Principle 5, https://www.gov.uk/government/publications/digital-economy-act-2017-part-5-codes-of-practice/research-code-of-practice-and-accreditation-criteria

[3] https://www.govinfo.gov/link/plaw/107/public/347?link-type=pdf

[4] https://www.congress.gov/bill/115th-congress/house-bill/4174

Repositories for federally funded data should be held to implement efficient mechanisms that allow for user and project vetting to be streamlined, and that repositories be allowed to share data or be accredited by multiple data owners, thus greatly increasing ease of access. In what follows, we illustrate three examples that have taken first steps, or even successfully implemented such streamlined processes.

# Example 1: Researcher accreditation

ICPSR at the University of Michigan has been developing a "researcher passport" (Levenstein, Tyler, and Davidson Bleckman 2018). Key element is "a credential that identifies a trusted researcher to multiple repositories and other data custodians, […] durable and transferable digital identifier issued by a central, community-recognized data steward." One possible steward might be a federally mandated entity. A portable digital credential is being considered by the European Union. In the UK, the "Digital Economy Act of 2017" went further, and implemented a legal status of "accredited researcher," with criteria laid out in the law itself, and a government panel to consider and vet requests for accreditation.[5]

Such a credential or accreditation would allow for efficiencies in the vetting process, and greatly ease access to data subject to access controls. We note that these must be "standard procedures", ideally initiated or controlled by federal government entity. They are unlikely to work if not mandated, as the current situation suggests.

# Example 2: Streamlining of project vetting

One of the costliest steps in providing secure and ethical access to restricted-access data is the per-project vetting process. While efforts are underway in the US to streamline the application process for federal data in support of the Evidence Act of 2018, less emphasis has been put on the approval process for applications. Currently, even where there is a streamlined application process, each application is evaluated individually, an often lengthy process. For other federally funded repositories, no single application process is envisioned that we know of.

Canada may serve as an example of a system that has attempted to streamline and accelerate such a system, reducing the barriers to restricted-access federal data.[6] Since 2019, certain classes of applicants for access are automatically pre-approved, meaning that they no longer have to go through a review process (they must still satisfy all security clearance criteria). Such applicants include any tenured professor at an accredited Canadian university, or recipients of peer-reviewed funding.

[5] https://www.statisticsauthority.gov.uk/about-the-authority/better-useofdata-statistics-and-research/betterdataaccess-research/better-use-of-data/
[6] https://www.statcan.gc.ca/eng/microdata/data-centres/guide

## Example 3: Coordination among networks of research centers

For better transportability and transferability of sensitive research data, coordination or mutual accreditation of secure repositories should be encouraged. The Federal Statistical Research Data Centers are a successful example in the context of data held by federal agencies, but have been slow in expanding the range of agencies and data. Loose coordination among NIH-funded repositories is an issue for the sharing of biomedical data.

Examples of stronger coordination exist in Germany and the UK. Administrative Data Research UK (ADR UK) plays an important role in bridging the gap between government and academia in the realm of administrative data, and in partnership with the Office of National Statistics (ONS).[7] Multiple "hubs" coordinate and implement access. In Germany, the German Data Forum has successfully established a decentralized network of accredited research data centers (RDCs) as a model solution for scientific data access.[8] A total of 31 research data centers are currently accredited and coordinated by the German Data Forum. Research data centers are annually evaluated. This infrastructure enables researchers to gain flexible access to a wide range of data. The UK and German networks also have an important additional component: outreach. The ADR UK Strategic Hub coordinates public engagement activities, helps to gauge public opinion regarding the use of the administrative data. The German Data Forum advises the German federal government and the governments of the Länder (states) on expanding and improving the research data infrastructure. It facilitates a continuous exchange between data producers and the data users in science and research with the aim of improving access to high-quality and scientifically potent data.

While these examples are primarily focused on data held and made available by the federal government, similar examples in the US are emerging. The Administrative Data Research Network (ADRN) is such an example, bringing together research projects that use data provided by various state and local levels. Many university-based secure computing environments exist, serving an important role, but must be authorized by data providers for each new project. A stronger coordination, for instance an accreditation mechanism for secure repositories for any source of data, has yet to emerge.

# I.C. Metadata

Finally, we point out that effective repositories of confidential data urgently need high-quality metadata (I.C.) on their data holdings, so that researchers can find, assess the utility of, and request access to research data that is pertinent for their scientific endeavors. Metadata on confidential data, when available, is currently scattered throughout various disconnected sites, often in disregard of widely available metadata standards. In general, there are few confidentiality concerns regarding the availability of metadata, and where these arise, for instance in the statistical metadata on extreme values, there are well-established measures to

[7] https://www.adruk.org/our-mission/our-mission/
[8] https://www.ratswd.de/en

handle these. We note that a critical element of the metadata needs to be the documentation of privacy-protecting measures applied to the microdata or the outputs (I.I.). Analyses that do not take full account of the statistical properties of the protection mechanisms are at risk of bias and other statistical problems. Analysts need to know exactly how to take into account these legitimate manipulations of the data. This can only be achieved through detailed information on those manipulations as part of the metadata.

Metadata (and the "connected" microdata) need to be findable, accessible, interoperable, and reusable (FAIR). The best implementations emerging in France and Germany are central metadata catalogs. Data.gov and efforts at various US universities (for instance, the Census Bureau data portal at ICPSR[9]) are a step in the right direction. Repositories that are subject to any future rules that may come out of this consultation should be instructed to provide metadata in such standards, and to provide metadata through standard API that can be queried and crawled by aggregating sites.

[9] https://census.icpsr.umich.edu/census/