American Economic Association
Office of the Data Editor, Email: dataeditor@aeapubs.org

The Data Editor of the American Economic Association (AEA) is pleased to respond 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).

Thank you for your consideration.
Questions on this document can be directed to the Data Editor of the AEA, Lars Vilhuber at dataeditor@aeapubs.org.

Primary discipline and roles

The American Economic Association (AEA), was founded as a professional society in 1885. Current membership is comprised of over 20,000 economists in academia, business, and government service. The AEA publishes eight journals, including the most prestigious academic journals in economics, as well as an electronic bibliography that serves as a comprehensive index to peer-reviewed journal articles, books, book reviews, collective volume articles, working papers, and dissertations.

In January 2018, I was appointed as the first Data Editor of the American Economic Association, with the mission to "design and oversee the AEA journals' strategy for archiving and curating research data and promoting reproducible research."

Comment

The importance of sharing data (and computational instructions, "code") for the purpose of transparency and reproducibility of science is paramount to AEA and for science in general. Repositories used by scientists to deposit the inputs, tools, code, and outputs of research, whether funded through federal funds or other, play a key role.

We in the AEA emphasize that the scope of these considerations should include research created by scientists in the direct employ of the federal government, data created for public and research use with federal funds as part of the business of the 13 federal principal statistical agencies, as well as any data created for research and evaluation under H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018. All of the above are federally funded, and are frequently used to validate research findings. It is as important to include the preservation of such data in the considerations of the SOS, and to ensure consistency of application of any guidelines issued across all these different domains.

We support the reference embodied in the cited standards (ISO16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements). In what follows, I comment on specific aspects of the characteristics as outlined in the RFC.

I. Desirable Characteristics of Data Repositories

A. Persistent Unique Identifiers

We agree that persistent identifiers are an important attribute of data in repositories. However, we also suggest that the federal government set aside funds specifically to support the registration of persistent unique identifiers in central registries. While the individual price seems low (as of February 2020, CrossRef charges $0.06 to assign digital object identifiers (DOI) for datasets or components, and the lowest tier at DataCite another registrar, is 500€), the associated cost of implementing robust integrated systems to perform the initial registration and maintain the associated landing pages is probably non-trivial. Assignment of DOI to specific (reproducible) queries or data extracts in interactive systems can quickly escalate. Costs for maintaining such systems typically extends beyond initial funding periods, but must in principle be supported "permanently".

> *Recommendation 1: Allow for funding in grants and research contracts for the maintenance of persistent identifiers.*

B. Long-term sustainability

Maintaining data assets for a sufficient long time is *critical* to ensure reproducibility. Two aspects are worthy of consideration here. First, most federal funding does not provide clear guidance that would allow for the expenditure of funds beyond the funding period. For instance, most research grants allow for expenses for the 2-5 years of the grant period, but are unclear about the use of funds to pay for storage or maintenance costs beyond the end of the grant period. In Europe, recent funding guidance clearly identifies data management costs as eligible costs, and explicitly allows for the costs of deposit of research data in an open access data repository (run by an external organization).

> *Recommendation 2: Explicitly allow for deposit costs as line items in federal funding vehicles, clarify usage of such funds when benefits accrue beyond the funding period.*

Second, we also note that not all data needs to be preserved into perpetuity. The question of how to identify when data can be de-accessioned or even destroyed is one where very little guidance exists in practice. Proper tracking of re-use (I.G) can provide some guidance, but is inherently a backward looking metric, whereas de-accessioning requires forward-looking analysis. We would encourage providing research funding to better understand how and when de-accessioning of data should be considered.

> *Recommendation 3: Fund research into the measurement of the long-term value of data.*

Finally, we recommend that whatever the preservation or retention policy may be, repositories should clearly state both a general policy as well as an object specific policy. Such policies can be recorded with central registries (e.g., Registry of Research Repositories, re3data) and within object-specific metadata, for instance the DOI record (DataCite Metadata Working Group 2017). Having this information easily available allows researchers to immediately assess the utility and robustness of a particular data item for their research, contributing to its reproducibility.

> *Recommendation 4: Require that information about dataset persistence be easily available in human and machine-readable form.*

C. Metadata

We strongly endorse the requirement of sufficient metadata. Much of economic research uses datasets which for a variety of reasons (ethical, commercial interests, security concerns) cannot be made available as public use data, and yet may be accessible through a variety of tiered access mechanisms (Federal Statistical Research Data Centers, licensing agreements, non-disclosure agreements, etc.). In order to make such access mechanisms more efficient, and to allow for re-use (I.G.), metadata is critical. Metadata allows researchers to prepare analysis code prior to accessing the restricted-access data (examples from Norway and Germany (Müller and Möller 2019) illustrate such procedures), making such procedures much less costly to researchers, and supporting ease of access (I.F).

However, we would also suggest that there are various degrees of metadata. We would strongly suggest that a minimum (and cheap) requirement for such repositories is to provide **data citations**. Data citations enable more consistent tracking of usage (by data providers) and of provenance (for scientific reproducibility), see (Martone 2014). Persistent identifiers (I.A) like DOI are not a requirement for proper data citation and attribution. Much more helpful is for repositories (in the broad sense) to provide suggested citations, and strongly encourage researchers to use them. An excellent example are the data citation practices of IPUMS. Even before the (relatively recent) implementation of DOI, IPUMS had an excellent track record of getting researchers to cite the (federally funded) data that they have prepared. Thus, the much simpler implementation of "suggested data citations" (prior to implementation of DOI) is a critical element to support

> *Recommendation 5: Require provision of a suggested data citation as the required minimum for metadata.*

D. Curation and Quality Assurance

We believe that there are various levels of appropriateness for curation and quality assurance. While heavily re-used data should be professionally curated, it should be possible to improve curation over time. To the best of our knowledge, there is currently no robust mechanism to

allow for continuous improvement in curation over time, in particular of metadata. In part this is technological (most existing repositories do not support such activities) as well as legal (unclear responsibilities and permissions of data owners). For instance, many entities -- IPUMS, FRED, NBER) have, over time, improved the metadata and curation of federally created data (data from Bureau of Labor Statistics and U.S. Census Bureau), but rely on that data being in the public domain. It is much harder to find examples where such data is freely available under open licenses, and yet being improved by entities other than the original data owners.

E. Access, and I. Privacy

Proper access description is key to broad re-use of data. While reasonable safeguards are necessary, they can take many different forms. The “Five Safes” framework (Desai, Ritchie, and Welpton 2016) highlights that many factors contribute to making data access safe, and can be balanced. Combining legal constraints (entering into enforceable confidentiality agreements), statistical data protections (anonymizing data) with physical constraints (accessing data only from safe rooms) allows data repositories to optimize the access protocol for the broadest possible access. It may be desirable for repositories to allow for multiple access protocols. For instance, allowing remote access to data for individuals with high trust, while allowing safe-room access to individuals who are building their trust, can increase the acceptability of stringent safety requirements.

Similar to our earlier point regarding the visibility of sustainability policies, whatever the access protocols for a particular dataset may be, they should be clearly and visible recorded. Access restrictions should be clearly outlined (for instance on dataset landing pages), and any conditions clearly described (e.g. citizenship or physical presence requirements). These should also be recorded as part of the metadata on the repository (aforementioned re3data) and the object (DOI).

F. Free and Easy to Access and Reuse

While there is little doubt that metadata should be free - a key tenet of the FAIR data principles - it is less clear that access itself needs to be free at the point of service. While free access for downloadable data seems to be a standard, it intersects with the (costly) long-term preservation (I.B.). More onerous but necessary access restrictions to enforce ethical or privacy concerns (I.E.) are generally much more costly. Sustainability in the absence of user fees is thus a concern that needs to be balanced with those aspects. A model that is seemingly practiced in the bio-medical community is for repositories to be developed, with federal funding, by third-parties, implementing access mechanisms, protocols, and policies. Once such repositories are stable, federal institutes (NIH) take over the continued maintenance of the repository, internalizing the maintenance cost. However, neither federal institutes nor funding for external activities are immune from the vagaries of the federal budget cycle, and are at risk of short-term funding cuts.

Alternative models see cost-recovery or user fees at the point of service, with such user fees being allowable on federal grants or other funding sources. An example of such a pricing scheme can be found for the French administrative data center (CASD). Such pricing schemes must balance the inequities that could be generated across the research landscape.

G. Reuse

We believe tracking of data reuse is a key metric to incorporate into any repository. And yet, the current, mostly manually curated bibliographies and other metrics are an inefficient mechanism for doing so. Leveraging persistent identifiers (I.A.), encouraging simple metadata (I.C. and our recommendation 5), and using existing registry infrastructure should automate such processes. However, all such mechanisms are ineffective if researchers do not actually cite the data used. We thus suggest that federally funded researchers be required to cite data, and that this requirement be enforced and rewarded.

Recommendation 6: Require data citations.

Positive reinforcement can come from making data citations a measurable metric in federal funding. For instance, when grant outcomes are reported, automatic mechanisms, fed by data citations in researchers' publications, can populate reports automatically. Use of data citations in grant evaluations and "prior outcomes" would incentivize researchers to adopt and use data citations.

Recommendation 7: Measure data citations in reporting mechanisms

References

DataCite Metadata Working Group. 2017. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.1." Edited by Jan Ashton, Amy Barton, Noris Birt, Stefanie Dietiker, Jannean Elliot, Martin Fenner, Wim Hugo, et al. https://doi.org/10.5438/0014.

Desai, Tanvi, Felix Ritchie, and Richard Welpton. 2016. "Five Safes: Designing Data Access for Research." University of the West of England. http://eprints.uwe.ac.uk/28124.

Martone, Maryann. 2014. "Joint Declaration of Data Citation Principles." Force11. https://doi.org/10.25490/a97f-egyk.

Müller, Dana, and Joachim Möller. 2019. "Giving the International Scientific Community Access to German Labor Market Data: A Success Story." In *Data-Driven Policy Impact Evaluation: How Access to Microdata Is Transforming Policy Design*, edited by Nuno Crato and Paolo Paruolo, 101–17. Cham: Springer International Publishing.