Making confidential data part of reproducible research

Pre-publication at Chance

Carl Lagoze (University of Michigan) and Lars Vilhuber (Cornell University)

This version: 2017-08-21

Disclaimer and acknowledgements: While this column mentions the Census Bureau several times, any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau or the other statistical agencies mentioned herein. The authors' work was supported by NSF grant #1131848 (NCRN), and by a grant from the Alfred P. Sloan Foundation.

Introduction

The rise of data-centric research practices has uncovered shortcomings in the traditional scholarly communication system. The foundation of that system, the peer-reviewed publication, "[the] selective distribution of ink on paper, or … electronic facsimiles of the same" (Bourne, et al., 2011), does not adequately support what has become an essential element of scholarship; the *reproducibility* of research results. That is, duplicating a reported result using the data, tools, techniques, etc. used in previous research. The notion of reproducible research is appealing for a number of reasons including facilitating novel research that "builds on the shoulders of giants", allowing the testing of veracity of existing research, and educating new scholars in common research practices. Reproducibility depends on disaggregating and exposing the multiple components of the research - data, software, workflows, and provenance - to other researchers and providing adequate metadata to make these components usable. The belief in the importance of reproducibility is shared by a large number of scientists, in many disciplines, who have pushed ever stronger for a modernization of the current system of scholarly communication, including support for reproducibility. In particular, we refer to the set of recommendations articulated by Stodden and co-authors in 2016, and which they call "Reproducibility Enhancement Principles" (REP, see Further Readings).

This column focuses on issues of confidentiality, which is intimately linked to reproducibility. There has been considerable concern in academic circles with respect to perceived lack of reproducibility of studies that are based on "proprietary" or "confidential" or "administrative" data, terms that are often conflated but that do not, in fact, describe the same type of data. The key worry is access: the authors of a study that uses confidential data cannot themselves deposit the data with the journal, thereby impairing easy access to those data and consequently impeding reproducibility. The conclusion, often heard at conferences, is that confidential data cannot be part of scientific process. We beg to disagree with that blanket statement.

In this column, we will address issues surrounding the reproducibility of confidential data held by national statistical offices (NSOs). In the United States, these might be the U.S. Census Bureau, the Bureau of Labor Statistics, or the Energy Information Administration, but similar research data access is possible in Canada, Germany, France, Norway, etc. We will not address issues surrounding confidential data stemming from sub-national administrations (states, counties, cities, school districts) or from private companies (individual company data, or data on other entities provided by privately held companies, such as social media data). The key distinction between NSOs and these other entities is how well the data owner curates the data, and manages access to the data. There are real issues with sub-national administrations and private companies that cannot easily be handled. We will argue here, though, that data held by NSOs do have attributes that lend themselves to reproducibility exercises, though this may, at present, not always be communicated correctly. How significant is this body of research? Based on our analysis of one particular journal (American Economic Journal: Applied Economics from 2009-2013), studies that use data held by NSOs accounts account for about half of all studies using confidential data.

While some groups propose ambitious, long-term transformations in scholarly communication to achieve large-scale reproducibility, our goal here is to describe some "low hanging fruit" with which we can achieve a measure of scientific reproducibility for this large body of scholarly work based on confidential data in the custody of NSOs. We will argue that such data can, in fact, be effectively part of a reproducible scientific endeavor, although as we will point out, there are potholes and bumps on the path to achieving the goal of reproducibility. Our modest "proposal" (a series of suggestions) leverages existing tools and practices. It promotes reproducibility through a number of components:

- *Descriptions of methodology*, which already exist in the traditional publication framework.
- *Naming systems*, which make it possible to assign unique identifiers to the components of research; data, publications, software artifacts, etc.

- *Archived data,* which in all cases is citable by other researchers and, in cases where the data is properly anonymized, is retrievable.
- *Metadata*, which describes the structure (variables) of data used in the research. In cases of restricted micro-data this metadata may be partially cloaked to hide protected variables and values.
- *Template language*, which can be used by researchers in their publications to highlight availability of data, proper data citation, and caveats to the data

An important component to these methods are an institutional commitment to maintain such practices, and easy-to-use tools and templates for researchers to leverage.

Data Access is Key

The key to reproducible confidential data is mechanisms to facilitate non-exclusive access. Most NSOs already have mechanisms in place that ensure that data access is not exclusive to the original researcher. It is not, however, available to anybody. We might complain that it is not feasible for the authors of this column (both of whom are Americans) to reproduce a study that uses Norwegian data, because data access there might require citizenship as a condition. But it is also true that in order to access U.S. Census Bureau data, a minimum amount of in-country residence is required. However, in both cases, there are many hundreds if not thousands of other researchers who are, in fact, qualified and enabled to access the data.

All is well then? Not necessarily. Application protocols vary across countries, and within countries, across agencies, sometimes across data sets. Often, information about the application process is obscure or complicated, and sometimes, requesting access is limited to certain “call for proposals” or other infrequent and limited time periods.

The Environment

A second key component is that most access to NSO-managed confidential data occurs within carefully-controlled, restricted environments. These lend themselves particularly well to implementing reproducible research. We illustrate this using the process in the FSRDCs, but similar processes are implemented in most restricted-access environments in the US or abroad.

Upon completion of the FSRDC-resident research, researchers must submit a Request for Clearance of Research Output. This request provides to the legal curator of the data the basis with which they can review the outputs of the research and determine whether they can be safely released publicly. This review verifies that the results are sufficiently anonymized to conform to the confidentiality mandates that the data are subject to. Information on this form specifies three key elements: the *input data* used for the research, the nature of the *output files*, and the analysis programs used to *transform the input data into the output files*, all of which are scrutinized to ensure that the outputs protect confidentiality. Disclosure protection is specified down to the variable level.

We draw the reader's attention to the fact that those same elements are the minimal elements required to "enable independent regeneration of computation results […] data, computational steps that produced the findings, and the workflow describing how to generate the results using the data and code," as recommended by Stodden and co-authors in the REP. Thus, by its very nature, the restricted access environment obliges the researcher to comply with reproducibility requirements – and in fact, in the case of the FSRDC, has been doing so for over 20 years!

So again, all is well? We don't think so. While all the key elements are present, they are, at present, hard to leverage for the average researchers. Even without the use of unique identifiers (which we will get to in a moment), most researchers do not communicate these elements to peers and journals, or if they do, do so in a highly inconsistent fashion. One of us has informally surveyed over one hundred authors of published articles about access to the data used in their study, and of those that responded (less than

half), few could adequately describe the access protocols to the data they had used. While part of the blame must reside with the researchers themselves, the NSOs granting access to the data must shoulder a part of the blame as well, by providing either no or inconsistent citable documentation on those key elements. Even when researchers post pre-publication working papers in NSO-managed archives (a frequent practice in economics, where publication lags are long), the data description is idiosyncratic, and non-compliant with any of several modern data citation standards. In part, this is due to the fact that, with rare exceptions, *no systematic, referenceable catalog* for the confidential data exists. To the best of our knowledge, no restricted-access data center network provides a way to reference the workflow (programs) and its outputs (disclosable results).

Identifiers: a requirement of a reproducible research environment

To facilitate and encourage the reproducibility of scholarly results, all the entities involved in the process of producing those results - people, publications, data, and computational artifacts – should be exposed, and the relationships among them expressed. *Identification* is the foundation to making this possible, enabling the citation and, hopefully, the retrieval of information (metadata) about an information object. Attributes should have *global uniqueness*, should be *machine actionable* and *human usable* as well as *time sensitive* for dynamic data. In particular, identifiers should be *persistent.* Finally, identifiers should be "*metadata aware"* - a data identifier should not only resolve to a particular data set, but also to the metadata associated with it – a particularly important point for confidential data. In human terms, this may be effected by the identifier resolving to a readable "splash page". In general, the commonly used URLs of the World Wide Web are not adequate, and the most frequently used identifier schema for publications, data, and other artifacts is the Digital Object Identifier (DOI).

The DataCite initiative (www.datacite.org) has taken the lead in data identification, organizing services to mint DOIs for datasets, associate basic metadata with these name data, providing search services for distributed data, tracking data use, and other functions.

Recommendations

Institutional commitment

A key to any of the proposed suggestions, whether they involve simple procedures or more complex mechanisms and infrastructure, is institutional commitment. Policies and procedures must be implemented, committed to, and managed in a persistent and transparent fashion. For instance, the institutional infrastructure that supports DOI landing pages may change radically within even a short time frame, breaking the relationship between data set and the DOI registered for it in an earlier period. There needs to be institutional commitment to ensure that these changes are propagated to the DOI registrar in a timely fashion, guaranteeing that the object is mapped into the DOI contemporaneously. Commitment does not necessarily imply monetary expense, but a high-level promise to engage with these mechanisms as a matter of policy.

> ***SUGGESTION****: Research centers of NSOs should commit to maintaining policies supporting reproducible research consistently and persistently.*

Transparency of Application Process

We noted earlier that access protocols may be ill-defined, may depend on time-sensitive application windows, and may provide little public guidance on the expected duration of application procedures.

> ***SUGGESTION****: Applications for access to confidential data should be allowed continuously, or at regular and high frequencies, for instance, with monthly or quarterly deadlines. The process should be transparent and predictable.*

Predictability does not imply homogeneity. When access requests are reviewed by k multiple agencies, the recurring joke is that the time for approval is ae^{bk}. As long as review periods are reasonable, predictability is key. Over the long-term, a transparent, centralized, multi-stakeholder application tracking process should be built, as is being considered in France. The process could be managed by the

NSO, by a grant agency, or by a third party, using open-access APIs to track a uniquely-identified proposal through well-defined stages of approval and review. This might provide confidence to prospective researchers as well as journal editors that the process is *transparent* and reasonably efficient in moving from proposal to approval stage.

Provide authoritative citations for existing objects and access procedures

Almost all NSO-managed systems share the feature that a researcher will receive at some point in the process an object from the secure system – typically model-based statistics (regression results) – typically by way of a disclosure review board or a privacy officer. We suggest that each such release should contain machine- and human-readable metadata on all the relevant objects: a standard citation for the input data in publications, both in the form of standardized language as well as a full data citation. The German Institute for Employment Research, for instance, provides examples for the former, for instance

> *"This study uses the weakly anonymous Establishment History Panel (Years YYYY - YYYY). Data access was provided via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) and/or remote data access."*

An example of a data citation is

> *U.S. Census Bureau. 2014. Geo-coded Address List (GAL) in LEHD Infrastructure, S2011 Version. [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor]*

In addition to citing the data, the method of access should also be described. A standard statement should indicate a brief summary of what conditions need to be met (if any) in order to qualify for access

to the data, with pointers to more complete descriptions. For example, the following statement might be useful:

> *The data in this article is confidential, and only accessible within the Federal Statistical Research Data Center network to qualified researchers on approved projects. Qualified researchers include most researchers affiliated with a U.S. academic institution. Approved projects are legally required, among other things, provide benefits to programs of the [data–providing agency], require non–public data, and pose no risk of disclosure. More information is available in www.census.gov/ces/pdf/Research_Proposal_Guidelines.pdf.*

These statements can be provided by authors to journals, and can be cited by authors in footnotes, data access descriptions, etc., as appropriate for each publication outlet.

Finally, all managed computer systems are associated with an archive facility. By providing researchers with sufficient information to recover programs from archives, NSOs can make intermediate data, programs, and workflow documentation traceable. Ideally, of course, programs and workflow documentation are themselves not confidential, and should be provided to researchers as part of the release of results. Nevertheless, by providing a citable location, NSOs can cover those scenarios where intermediate programs are too complex and costly to analyze for disclosure risks, and provide additional credibility that the programs provided to journals are, in fact, the programs that were used to produce the results.

> *<u>SUGGESTION</u>: object citations and access descriptions should be provided as part of **every release of results** by NSOs, customized to the project that is requesting release of such results. They are cheap to provide, and will go a long way to improving the perceived reproducibility of the research.*

By doing so, the NSO makes it easy for researchers to give proper credit for shared digital objects, as noted in the third recommendation in the REP.

Consistent use of persistent identifiers for all the components of the research process

We argued that unique identification of information resources is a necessary precursor for their reusability. Unique identification hinges on having "naming policies" in place, and procedures to implement them. In order to be useful, these identifiers need to be public, but they do not have to use DOIs – they can rely on existing identifier systems. Converting idiosyncratic identifier systems to DOIs at a later stage is easy, and should certainly be included in any long-term plan, but much mileage can be had out of implementing *some standardized* identifier system right away. Critically, a "landing page" for each object – a dedicated, referenceable page of an online catalog – needs to exist, with suggested citations.

Using the FSRDCs as an example, all datasets and projects are tracked by a formal internal project management system, called "CMS." A dataset might be referenced as "cmsd00035", and projects as "cmsp000538." Where appropriate, versioning should be handled, for instance "cmsd00035v2" would indicate a second released version of dataset 35.

Once a naming policy is implemented, it becomes relatively straightforward to provide landing pages for all such objects, e.g., http://rdc.nso.gov/cmsd00035. While most data archives (e.g., ICPSR) use such standardized URLs, most NSOs that we are aware of do not. In particular, project-related splash pages do not exist. Nevertheless, it would seem straightforward to publish some details for projects, such as titles and abstracts, as some FSRDCs already do in newsletters and the like.

> *SUGGESTION: NSOs should **implement a naming schema** covering all objects related to the research workflow, **publish** the details for the naming schema, and **create landing pages** for all such objects.*

The suggestion actually entails implementing all the conditions for assigning DOIs. The long-term goal should be assigning DOIs to all these objects, and thus achieving the second recommendation in the REP.

Benefits

Once global identifiers permeate the system, it is trivial to use the existing facilities of CrossRef, DataCite, and other ongoing projects to construct citation impacts for the data used, and for the papers created based on the data. Researchers can obtain credit for the digital objects they created. NSOs and funding agencies can measure the impact on research and policy in an objective manner. Infrastructure to that extent already exists in Europe, see OpenAIRE. Researchers who receive funding from NSF, NIH, etc. obtain citable objects for grant reporting mechanisms, and can prove compliance with data management plans. With properly identified data, replicable processes, and predictable access mechanisms, it becomes possible to conceive of replication challenges.

Conclusion

We have outlined a number of steps that national statistical offices and their associated research centers could undertake to improve actual and perceived reproducibility of research that leverages data under their control. Some of these steps are very easy and could be implemented quite quickly. We take no credit for coming up with the examples – for each process we suggest, we are aware of at least one NSO that is actively following that process. However, no NSO, to our knowledge, implements all of the suggestions. Additional steps are required beyond these suggestions, such as comprehensive documentation of all digital objects (the fourth recommendation in the REP), but taking these initial steps is critical for starting down that path.

Further Reading

Abowd, J. M., Vilhuber, L., & Block, W. (2012). A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs. In J. Domingo-Ferrer & I. Tinnirello (Eds.), *Privacy in Statistical Databases* (Vol. 7556, pp. 216–225). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/978-3-642-33627-0_17

Altman, M., Arnaud, E., Borgman, C., Callaghan, S., Brase, J., Carpenter, T., … Socha, Y. (Editor). (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy and Technology for Data Citation. *Data Science Journal*, *12*, 1–75. https://doi.org/10.2481/dsj.OSOM13-043

Bourne, P. E., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E., & Shotton, D. (2011). *FORCE11 MANIFESTO*. FORCE11: The Future of Research Communications and e-Scholarship. Retrieved from www.force11.org/group/joint-declaration-data-citation-principles-final

ICPSR. (2016). Citing Data. Retrieved from www.icpsr.umich.edu/icpsrweb/content/datamanagement/citations.html

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., … Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

NOT TO BE INCLUDED

References

Abowd, J. M., Vilhuber, L., & Block, W. (2012). A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs. In J. Domingo-Ferrer & I. Tinnirello (Eds.), *Privacy in Statistical Databases* (Vol. 7556, pp. 216–225). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/978-3-642-33627-0_17

Altman, M., Arnaud, E., Borgman, C., Callaghan, S., Brase, J., Carpenter, T., … Socha, Y. (Editor). (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy and Technology for Data Citation. *Data Science Journal*, *12*, 1–75. https://doi.org/10.2481/dsj.OSOM13-043

An Audit Checklist for the Certification of Trusted Digital Repositories. (2005). Retrieved from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Big Data Senior Steering Group, Subcommittee on Networking and Information Technology Research and Development. (2016). *The Federal Big Data Research and Development Plan*. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/bigdatardstrategicplan-nitrd_final-051916.pdf

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923.

Bourne, P. E., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E., & Shotton, D. (2011). *FORCE11 MANIFESTO*. FROCE11: The Future of Research Communications and e-

Scholarship. Retrieved from https://www.force11.org/group/joint-declaration-data-citation-principles-final

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, aaf0918. https://doi.org/10.1126/science.aaf0918

Chan, L., & Zeng, M. (2006). Metadata Interoperability and Standardization–A Study of Methodology Part I. *D-Lib Magazine*. Retrieved from http://dlib.org/dlib/june06/chan/06chan.html

Chetty, R. (2012). The Transformative Potential of Administrative Data for Microeconometric Research. Retrieved from http://conference.nber.org/confer/2012/SI2012/LS/ChettySlides.pdf

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, *4*(3), 139–160. https://doi.org/10.1007/s12145-011-0083-6

Garfield, E., & Welljams-Dorof, A. (1992). Citation data: Their use as quantitative indicators for science and technology evaluation and policy-making. *Public Policy*, *19*(5), 321–327.

ICPSR. (2016). Citing Data. Retrieved from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/citations.html

Jantz, R., & Giarlo, M. J. (2005). Digital Preservation: Architecture and Technology for Trusted Digital Repositories, *11*(6). Retrieved from http://www.dlib.org/dlib/june05/jantz/06jantz.html

Lagoze, C., Block, W., Williams, J., Abowd, J. M., & Vilhuber, L. (2013). Data Management of Confidential Data. Presented at the International Data Curation Conference, San Francisco, CA.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, *6*(2), 4–37. https://doi.org/10.2218/ijdc.v6i2.205

Pollard, T. J., & Wilkinson, J. M. (2010). Making Datasets Visible and Accessible: DataCite's First Summer Meeting. *Ariadne*, (64). Retrieved from http:/sss/www.mendeley.com/research/making-datasets-visible-and-accessible-datacites-first-summer-meeting-ariadne-issue-64-1/

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's "Retroactive Facilitation of Recall" Effect. *PLOS ONE*, *7*(3), e33423. https://doi.org/10.1371/journal.pone.0033423

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, *1*, e1. https://doi.org/10.7717/peerj-cs.1

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., … Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Stodden, V., & Miguez, S. (2013). *Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research* (SSRN

Scholarly Paper No. ID 2322276). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2322276

U.S. Census Bureau. (2015). *Census RDC Research Proposal Guidelines*. Retrieved from http://www.census.gov/ces/pdf/Research_Proposal_Guidelines.pdf

Velden, T., al Haque, A., & Lagoze, C. (2011). Resolving Author Name Homonymy to Improve Resolution of Structures in Co-author Networks. In *Joint Conference on Digital Libraries*. ACM/IEEE.

Vilhuber, L., & McKinney, K. (2014). *LEHD Infrastructure files in the Census RDC - Overview* (Working Papers No. 14–26). Center for Economic Studies, U.S. Census Bureau. Retrieved from https://ideas.repec.org/p/cen/wpaper/14-26.html

Zeng, M., & Chan, L. (2006). Metadata interoperability and standardization-A study of methodology, Part II. *D-Lib Magazine*. Retrieved from http://mirror.dlib.org/dlib/june06/zeng/06zeng.html