A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs

John M. Abowd[1,⋆], Lars Vilhuber[1], and William Block[2]

[1] Department of Economics and Labor Dynamics Institute, Cornell University, Ithaca, NY, USA
john.abowd@cornell.edu
[2] Cornell Institute for Social and Economic Research, Cornell University, Ithaca, NY, USA

Abstract. We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols. It is based on extensible tools and can be easily incorporated into existing instructional materials.

Keywords: Data Archive, Data Curation, Statistical Disclosure Limitation, Privacy-preserving Datamining.

1 Introduction

The era of public-use micro-data as a cornerstone of empirical research in the social sciences is coming to an end—not because it is no longer feasible to create such data without breaching confidentiality. It still is, and statistical agencies like the Census Bureau will continue to do so. Rather, the death knell is being sounded by young scholars pursuing research programs that mandate inherently identifiable data: geospatial relations, exact genome data, networks of all sorts, linked administrative records, and so on. These researchers acquire authorized restricted access to the confidential, identifiable data and perform their analyses in secure environments. And their research is challenging fundamental scientific principles because the restricted access cannot be extended arbitrarily to the whole user community [11].

The Census Research Data Centers are a leading paradigm for such research, but other modalities are proliferating rapidly. The researcher is allowed to publish results that have been filtered through a statistical disclosure limitation protocol. Scientific scrutiny is hampered because the researcher cannot effectively implement a data management plan that permits sharing these restricted-access data with other scholars. In the case of Census RDCs the relevant statute has been interpreted to prohibit granting long-term data custody outside of the Bureau except for copies held by the National Archives, which does not permit

⋆ Corresponding author.

J. Domingo-Ferrer and I. Tinnirello (Eds.): PSD 2012, LNCS 7556, pp. 216–225, 2012.

public access to these holdings. University-operated archives like ICPSR may take custody of non-Census Bureau restricted-access data under some conditions, but they still cannot freely grant access to the confidential micro-data in their repositories. The data custody problem is impeding the "acquire, archive and curate" model that dominated social science data preservation in the era of public-use micro-data.

2 Statement of the Problem

2.1 The Curation of Confidential Data

In the United States, the National Science Foundation (NSF) has required since January 18, 2011 that all scientific research proposals include a detailed, viable data management plan, thus recognizing that the acquisition, archival and curation of scientific data is vital to the integrity of the entire process.[1] The relevant test is not "can the next researcher reproduce current results," rather it is "can a researcher working 50 or 100 years from now recover and correctly re-use the original data." This standard will be met when "sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." [12] Libraries have performed the curation (or preservation) function for millennia. Social scientists recognized the importance of data management decades ago when the Inter-university Consortium for Political and Social Research (ICPSR) was formed, and again a few decades later when NSF funded major social science data initiatives like Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota and the Research Data Centers (RDCs) at the U.S. Census Bureau.

ICPSR is now the largest social science data repository in the world with over 500,000 data sets in its collection, including a growing inventory of restricted-access datasets.[2] IPUMS and IPUMS-International are the definitive sources for household micro-data originating from population censuses around the world, including projects for which IPUMS-International is the long-term custodian of a foreign nation's confidential micro-data.[3] Similar archives, such as the UK Data Archive[4] and the Australian National Data Service,[5] perform similar functions in other countries. Within statistical agencies, researchers working at the U.S. Census Bureau and in Census RDCs have acquired and archived a very substantial collection of micro-data that are now used routinely for scientific research in economics, sociology, demographics, environmental science, health,

[1] http://www.nsf.gov/eng/general/dmp.jsp cited on May 20, 2012.
[2] See http://www.icpsr.umich.edu/icpsrweb/ICPSR/org/index.jsp, cited on May 20, 2012
[3] See https://international.ipums.org/international/about.shtml, cited on May 20, 2012.
[4] http://www.data-archive.ac.uk/about/archive, accessed May 20, 2012
[5] http://www.ands.org.au/, accessed May 20, 2012

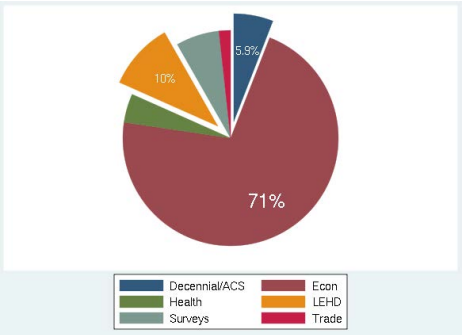

Fig. 1. Data sets used in U.S. Census Bureau RDC projects

and other fields. Other NSF-funded efforts to make data available have also been very successful.

Figure 1 shows the overall distribution of data sets used in current and historical RDC projects. It summarizes 1,505 project-dataset pairs.[6] Fully 71% of all project-datasets use economic (business or establishment) micro-data. Such data are primarily the establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD). With the exception of the recently-released Synthetic LBD [3,13], there are no public-use micro-data for these establishment-based products. Yet, they form the core of the modern industrial organization studies [7,16] as well as modern gross job creation and destruction in macroeconomics [6,9].

The next most frequently used data come from the Longitudinal Employer-Household Dynamics (LEHD) program, a longitudinally integrated employer-employee database that was created following a joint Census Bureau-NSF investment in 1999 [2]. New confidentiality protection methodologies [1,15] have unlocked large amounts of data for public-use but the structured metadata has not kept pace. While highly detailed local area tabulations exist based on the LEHD data, no public-use micro-data exist for this longitudinal job frame or any of its derivative files.

Somewhat surprisingly, only about 6% of the project-dataset pairs involve confidential Decennial/American Community Survey (ACS) data. Public-use decennial files from both the long and short forms have existed for decades. These lacked geographical detail when they were based on the old long form. However, geographically detailed historical census and ACS files are now part of the Census RDC-accessible micro-data collection. Thus, one can reasonably speculate that the fraction of projects that use confidential American Community Survey (ACS) will rise in the coming years.

[6] Many projects use multiple datasets.

Over the course of the last decade a framework for providing access to the confidential micro-data that form the basis for the Census Bureau's major data products has emerged. This framework is consistent with the statutory obligations of the Bureau's co-custodians; namely, that research use of the micro-data be consistent with the enabling legislation for each constituent data source and that the appropriate administrative review occur prior to the onset of new research. This framework is currently the best available political compromise in the United States, but it can be considered neither permanent nor durable.

A similar spectrum of data access protocols has emerged in Europe. They range from relatively easy research access to confidential micro-data to remote processing of firm or person micro-data[7] to simple online tabulators at most statistical agencies. As of 2012, efforts are underway to harmonize European [5] or international [14] regulations, facilitating a standardized approach to cross-national data access. However, it appears that most efforts have concentrated on technical and legal questions.

To the extent that the next generations of social scientists build their careers on the basis of original discoveries emanating from these confidential data in the United States and elsewhere, a regulatory consensus must emerge that treats the underlying confidential data as a vital scientific asset, including its curation procedures.

When this consensus emerges, it will be too late to begin the curation process. In contrast to printed data (otherwise known as books and journals), which have unique handles (International Standard Book Number (ISBN) and International Standard Serial Number (ISSN) are almost universally applied), data files generally have not yet been managed in a similar fashion.[8] Part of the problem, of course, is that while the origin and version of printed matter used to be easily identifiable (expensive print runs and distribution paths ensured that no book ever got to its 500th edition), data have become more and more variable and extensible. Thus, most data currently lack a unique handle that can be used to trace their design, provenance and vintage.

2.2 Current Archive Model Fails

Big data archives such as ICPSR, IPUMS, the UK Data Archive, or the International Data Service Center at IZA have done an extraordinary job of preserving public-use data–often rescuing them from oblivion–and provide some idiosyncratic way to refer to specific samples. But there is a fundamental, and critical, difference between the approach taken by the data archives as compared to the approach taken by the U.S. Census Bureau, other governmental agencies and most private organizations that use confidential micro-data as the basis for original research or provide research access to such data. The curation function is

[7] See for instance http://www.bancaditalia.it/statistiche/indcamp/sondaggio/bird and http://www.lisproject.org/data-access/lissy.htm accessed May 20, 2012.

[8] To the best of our knowledge, only ICPSR and the UK Data Archive assign unique Digital Object Identifiers (DOIs), but only to data that they physically control.

either absent or woefully neglected. Consequently, there is a substantial risk of breach of the scientific integrity of the research process itself because the findings that are reported in the peer-reviewed journals are based on analyses of the confidential restricted-access data, but only public-use data are released for open scrutiny. It is the confidential data themselves that must be curated, not just the disclosure-limited public-use products that this research produces, in order to afford future generations of scientists the same ability to scrutinize this work as many generations have had for work based on the major public-use data products developed in the last 50 years.[9] The statutory custodians of the restricted-access data, in most cases government agencies but also private-sector entities, need substantial help from the scientific community in order to ensure that vital research data they have now acquired are properly curated.

The problem has been caused by a subtle but pervasive barrier to effective application of current best-practice long-term data management systems. When conventional repositories like ICPSR, IPUMS-International and the IZA Data Enclave have attempted to apply the acquisition, archive and curation processes developed for public-use data directly to restricted-access data, the management of restricted-access data adds an additional layer, sometimes called stewardship, to the accepted practices. The data archive takes physical custody of a certified-true copy of the confidential data under the terms of a restricted-access data provider agreement with the statutory custodian. This agreement establishes the statutory custodian's legal authority to grant physical data custody to the archive and delineates the terms and conditions of future use, including any disclosure limitation protocols that must be used. At the same time, the archive acquires or creates the metadata that are essential to the curation process. From this point forward, management of the restricted-access data is very similar to management of public-use data. In particular, many resources from the data archive and the research community can be used to enhance the curation process.

But if the conventional archive cannot take long-term custody of the original data, this model fails because it does not have a mechanism for synchronizing the provenance and metadata histories applicable to the confidential data that can be audited and verified by future data users. The U.S. Census Bureau and many other American government agencies are prohibited by statute from granting an archive like ICPSR or IPUMS long-term physical custody of their confidential data. Private-sector entities may also have legal barriers emanating from data privacy promises, or may simply hesitate to provide potential competitors access to detailed micro-data. Both micro-data and metadata are locked up and inaccessible.

Because private entities like Microsoft or Google and government agencies like the U.S. Census Bureau retain custody of both the confidential data and critical metadata, a substantially modified curation protocol is required to ensure that the actual inputs to published research are preserved. Some requirements for this protocol are discussed here.

[9] The 1960 U.S. Census of Population and Housing Public Use Micro Sample, released in 1963, was the first such product released by a national statistical agency [17].

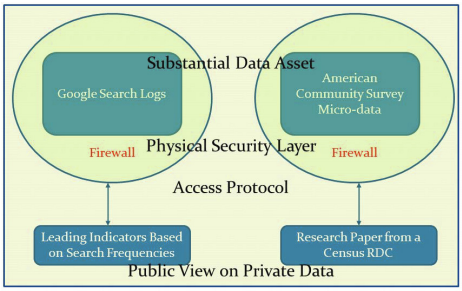

Fig. 2. The Parallel Problems of Public and Private Data Stewards

3 Principles for Solution

3.1 The Commitment of Primary Custodians

Figure 2 shows the problem faced by public or private data custodians who grant research access to their data. The primary data asset is protected by both a physical security layer and an access protocol, both of which stand between the ultimate user of the scientific output and the confidential data. The physical security layer ensures that other potential users do not gain unauthorized access. The access protocol limits what may be released and published using privacy-preserving or statistical disclosure limitation methods.

Unless the primary custodian commits to long-term archival and curation of both the data and their metadata, the integrity of the process is corrupted. In the private domain, future users of the published indicator cannot rely upon the continued scrutiny of other users to expose and correct defects in the inputs and methodology of the published indicators. In the public domain, users of the research output cannot properly review the original work nor reliably build on it in future work. Both failures result from the effective denial of access to both the curated data and metadata.

Once a private or public provider commits to the long-term obligations of scientific data custodian, the problem becomes how to integrate the archival and curation process with their physical security layer and access protocols. This integration is an unsolved problem although tools from both statistical disclosure limitation and data curation are useful.

3.2 Transparency among Users

All of the data processing for the scientific research referenced in Figure 2 is done in a controlled environment that lacks the tools needed to conform to emerging standards for data documentation. "[T]he metadata of data files are crucial for browsing and searching" because data files generally do not lend themselves to the same indexing techniques as text files [10]. The consequence is data that are difficult to discover, and, when found, only sparsely documented. Researchers

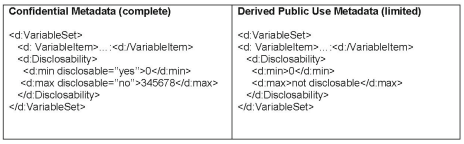

Fig. 3. Example of Confidential and Derived Public-use Metadata

waste valuable time trying to determine the content and structure of confidential datasets in sufficient detail to support their proposed secondary analysis. Some confidential datasets even contain variables whose names themselves are masked.[10] When confronted with difficult problems such as these, researchers resort to time-consuming alternative search strategies like email queries.

A better solution is needed, one that allows researchers to efficiently learn about and work with the confidential data without violating existing access protocols, and one that ensures that the exact historical research inputs and their provenance are curated for a long time. Inefficiencies that current users might be prepared to tolerate discourage potential users from ever starting. The absence of reliable curation may effectively orphan the research done in this early era of restricted-access data use.

3.3 Conformance to Standards

The Royal Society [18] has recently called for metadata that goes beyond basic, generic contextual information and meets four fundamental characteristics. Metadata must be: accessible (a researcher can easily find it); intelligible (to various audiences); assessable (are researchers able make judgments about or assess the quality of the data); and usable (at miniumum, by other scientists).

Leading metadata standards such as the Data Documentation Initiative (DDI) and Statistical Data and Metadata eXchange (SDMX) are flexibly designed to ingest documentation from a variety of source files. Using these tools to standardize the curation of confidential research data permits the exercise to benefit from the same technological innovations that open-access data archives already use. [8,4]

But the benefits go in both directions. These tools need to be extended so that they can naturally accommodate metadata items that respect privacy-preserving and statistical disclosure limitation procedures. In a model based on Extensible Markup Language (XML), for example, this might be done through the addition of machine-actionable attributes to elements describing variables. An example of a possible template, assuming an XML-like structure, is shown in Figure 3. The example could be applied, for instance, to a variable containing data on

[10] For example, the U.S. Census Bureau's establishment micro-data contain data elements from the Internal Revenue Service whose confidentiality stewards have designated the names of certain fields as "official use only," which implies that these metadata are confidential too.

income or sales. The element "Disclosability" is not currently present in the DDI specification, but could be defined in a future release.

The full-information metadata can be presented through a restricted-access website available only within the secure environment itself, running the same web frontend used for the public interface. Such a development itself would provide a major advance in the ability of confidential data researchers to conduct their work because in many environments, including those supported in U.S. Census Bureau RDCs, the public metadata interface cannot be viewed inside the secure layer and the confidential data have not been curated to the same level of specificity.

3.4 Training of Future Users

Graduate social science programs and their faculties haven't worried about how future users would gain adequate instruction in the major public-use micro-datasets for decades. The body of discipline-specific capital is sufficiently extensive and the data curation tools sufficiently advanced, that doctoral programs and social science faculty members can rely on course assignments, specialized workshops, and existing archives and repositories to disseminate such methods. That doesn't happen with confidential data because the potential user must usually already have a specific approved project and be allowed access inside the security protocol layer before any study of the metadata or analysis of the actual data can be done.

These costs are sometimes mitigated by virtual enclaves like the Cornell VirtualRDC[11], the NORC Data Enclave[12], or the International Data Service Center (IDSC) of the Institute for the Study of Labor (IZA).[13] But usually the fixed costs are simply too high to incorporate this kind of hands-on experience in regular doctoral courses or short-term research projects. The existence of coordinated metadata curation, as described above, mitigates this difficulty by providing a layer of access outside of the secure protocol for the metadata that supports the research outputs.

4 Conclusion

In the United States, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) formalized the obligation of every federal statistical agency to take long-term custody of the confidential micro-data used for its work. These agencies all face the same problem as the U.S. Census Bureau, which assumed a comparable obligation when U.S. Code Title 13 was adopted in 1954 and national statistical agencies around the world, which usually operate under legal constraints that forbid granting long-term custody to an entity

[11] See http://www.vrdc.cornell.edu/news/ cited on May 20, 2012.
[12] See http://www.dataenclave.org/index.php/home/welcome cited on May 20, 2012.
[13] See http://idsc.iza.org/ cited on May 20, 2012

that is not part of their government. The acquisition, archival and curation system described here can be generalized to restricted-access research requirements of many statistical agencies and private data stewards. The tools would allow such agencies to harness the efforts of researchers who want to understand the structure and complexity of the confidential data they intend to analyze in order to propose and implement reproducible scientific results. Future generations of scientists can build on those efforts because the long-term data preservation operates on the original scientific inputs, not inputs that have been subjected to statistical disclosure limitation or privacy-preserving filters prior to entering the repository. Such curation provides sponsors like national scientific research organizations with a viable system for enforcing data management plans on projects, ensuring that results can be tested now and replicated many years in the future.

Acknowledgment. We acknowledge NSF grants SES 9978093, ITR 0427889, SES 0922005, SES 1042181, and SES 1131348.

References

1. Abowd, J.M., Gittings, K., McKinney, K.L., Stephens, B.E., Vilhuber, L., Woodcock, S.: Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series. Technical report, Federal Committee on Statistical Methodology (January 2012)
2. Abowd, J.M., Stephens, B.E., Vilhuber, L., Andersson, F., McKinney, K.L., Roemer, M., Woodcock, S.D.: The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In: Dunne, T., Brad Jensen, J., Roberts, M.J. (eds.) University of Chicago Press (2009)
3. Abowd, J.M., Vilhuber, L.: Synthetic data server (2010)
4. Blank, G., Rasmussen, K.: The Data Documentation Initiative: The value and significance of a worldwide standard. Social Science Computer Review 22(3), 307–318 (2004)
5. Bujnowska, A.: European regulations and current changes. Presentation, First Data without Boundaries European Data Access Forum (March 2012)
6. Davis, S.J., Haltiwanger, J.C., Schuh, S.: Job Creation and Destruction. MIT Press, Cambridge (1996)
7. Dunne, T., Roberts, M.J., Samuelson, L.: The Growth and Failure of U.S. Manufacturing Plants 104(4), 671–698 (1989)
8. Gregory, A., Heus, P.: DDI and SDMX: Complementary, not competing, standards. Paper, Open Data Foundation (July 2007)
9. Haltiwanger, J., Jarmin, R.S., Miranda, J.: Who creates jobs? Small vs. large vs. young. Working Papers 10-17, Center for Economic Studies, U.S. Census Bureau (August 2010)
10. Hense, A., Quadt, F.: Acquiring high quality research data. D-Lib Magazine-the Magazine of Digital Library Research 17 (2011)
11. Huberman, B.A.: Sociology of science: Big data deserve a bigger audience. Nature 482(7385), 308–308 (2012)
12. King, G.: Replication, replication. PS: Political Science and Politics 28(3), 444–452 (1995)

13. Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., Abowd, J.M.: Towards unrestricted public use business microdata: The synthetic longitudinal business database. International Statistical Review 79(3), 362–384 (2011)
14. Lunati, M.: Oecd expert group for international collaboration on microdata access. Presentation, First Data without Boundaries European Data Access Forum (March 2012)
15. Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: International Conference on Data Engineering, ICDE (2008) (in press)
16. Steven Olley, G., Pakes, A.: The dynamics of productivity in the telecommunications equipment industry. Econometrica 64(6), 1263–1297 (1996)
17. Ruggles, S.: Comparability of the public use files of the U.S. Census of Population, 1880-1980. Social Science History 15(1), 123–158 (1991)
18. The Royal Society. Science as an open enterprise: Open data for open science. report 02/12, The Royal Society Science Policy Centre (2012)

Acronyms Used

ACS American Community Survey
DDI Data Documentation Initiative, see http://www.ddialliance.org/
DOI Digital Object Identifier
ICPSR Inter-university Consortium for Political and Social Research
IDSC International Data Service Center
IPUMS Integrated Public Use Microdata Series
ISBN International Standard Book Number
ISSN International Standard Serial Number
IZA Institute for the Study of Labor
LBD Longitudinal Business Database
LEHD Longitudinal Employer-Household Dynamics
NSF National Science Foundation
RDC Research Data Center
SDMX Statistical Data and Metadata eXchange, see http://sdmx.org
XML Extensible Markup Language