The International Journal of Digital Curation

Issue 1, Volume 2 | 2007

Using a Remote Access Data Enclave for Data Dissemination

Julia Lane,

National Opinion Research Center (NORC),

University of Chicago

Stephanie Shipp,

Advanced Technology Program,

National Institute of Standards and Technology

July 2007

Summary

This article describes the approach taken by NORC and NIST to provide remote researcher access to confidential business micro data. We have combined technical, legal and organizational approaches to protect respondent confidential. We have also instituted a number of technical approaches to encourage researchers to provide metadata documentation.

"Many believe that the problem of finding ways to meet the legitimate privacy and confidentiality concerns of human subjects is the Achilles heel of the current data explosion." (Berman & Brady, 2005, p. 22)

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The creation and analysis of high-quality information are core elements of the scientific endeavor. No less fundamental is the ability to replicate scientific analysis. Yet the individual level data on human behavior which is the basis for empirical research in a wide variety of disciplines – such as the biological, social and some computer sciences – is often not accessible to others for replication and validation. One reason, as indicated by the opening quotation, is that access to data on human subjects is limited because of both legal and ethical protections. Another, as noted in a recent National Science Board report (2005), is that researchers typically do not archive and curate their data sufficiently to provide accurate replication of their work, making it difficult to document the benefit of researcher access. In particular the Board states "to make data usable, it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called metadata. Ideally, the metadata are a record of everything that might be of interest to another researcher".

Ensuring that high-quality research on microdata is possible requires solving a series of technical and social challenges, namely:

1. The human beings who are the sources of the data must be convinced that their information is protected from access and use by unauthorized individuals and for unauthorized purposes.
2. Researchers must be provided with a research environment which facilitates high-quality research.
3. The benefits of researcher access to microdata must be clearly demonstrable to the producer to justify both the risk and the cost of providing that access.

The NORC Data enclave, which has been developed by NORC in conjunction with the National Institute of Standards and Technology (NIST) Advanced Technology Program[1] has begun to provide remote researcher access to microdata as an initial attempt to address these issues. It combines elements from the computing and social sciences to develop secure remote data access protocols that not only provide technical security, but also create an environment whereby researchers can do high-quality research. In particular, it creates an environment which facilitates documentation and research dissemination, and hence helps demonstrate the benefits of researcher access to the producer.

Background

The view expressed by a recent panel of the National Commission on Health Statistics[2] about the best way to promote high-quality academic research is shared by many: panel members explicitly stated that researchers needed to access and view original microdata directly in their offices, conduct their analyses, share their results, and engage in discourse about different aspects of the data, the analysis, and the interpretation of the results. Yet there are legal and ethical reasons that have prevented such access occurring – with substantial negative consequences.

[1] http://dataenclave.norc.org

[2] http://www.ncvhs.hhs.gov/060918tr.htm

The legal framework for the protection and dissemination of the administrative, clinical and survey data that underpins much empirical research is complex. The legal requirement is not typically defined, but is left to the discretion of the agencies. By contrast, although the authorizing legislation for different agencies typically requires them to produce information for decision makers, research access to microdata is not an explicit part of their mandate[3].

The ethical framework is similarly complex. Statistical agencies, like most data collectors and custodians, provide respondents with a guarantee that their identity will be protected. Safeguarding this guarantee is critical to maintaining their reputations and, not coincidentally, their future response rates. Protecting confidentiality necessitates perturbing the data in some fashion so that individual respondents can not be identified. While statistical agencies go to great lengths to collect high-quality data, the necessity of protecting the data results in some data quality compromises.

Not surprisingly, the complex ethical framework and the severe adverse consequences associated with breaches of confidentiality, leads to what Madsen (2003)[4] refers to as the "Privacy Paradox". As he points out, data custodians who interpret the right to privacy as a near-absolute ethical standard, might have a much more extreme understanding of the nature of the responsibility of confidentiality than is socially optimal. In other words, data custodians who operate within a myopic framework, and establish new and more restrictive controls on data access, act to reduce the scientific value of data, and hence substantially reduce the social benefits of the data collection.

Two examples serve to illustrate that the full returns to data collection are not being derived, and why it is necessary to develop new access modalities that can be trusted by both data producers and the respondents who provide them data. One is derived from the National Science Foundation supported Census Bureau's Research Data Centers (RDCs). Researchers who want to access microdata on businesses or individuals have to go through a lengthy proposal submission and security clearance process that often takes 6 to 12 months or longer. Once approved, they must then physically go to and work inside the RDC. Yet at least the Census Bureau provides access to researchers. Many other U.S. federal and state agencies, including the Centers for Medicare & Medicaid Services, have either severely restricted or indeed eliminated access altogether, judging it too expensive in the current fiscal environment[5].

Another approach used by statistical agencies to provide microdata access has been to produce public use files. By the measure of researcher take-up, this has been a huge success – not only do thousands of academic social scientists publish papers using datasets such as the Decennial Census Public Use Microdata files, and the Current Population Survey, to mention just two, but undergraduate and graduate students learn to employ analytical tools using such datasets. Yet, in order to protect

[3] Indeed, the 2001 Criteria agreement between the Internal Revenue Service and the Census Bureau is clear that the predominant purpose for microdata access must be to improve Census Bureau data products. Analytical research is secondary. The agreement identifies nine separate ways in which researchers can comply with this requirement (see www.ces.census.gov).

[4] See also http://www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf

[5] The infrastructure costs of the RDC program to the Census Bureau are about $3 million/year, the marginal costs of each RDC, about $200,000 - $300,000, is supported by the RDC host institution.

respondent confidentiality, data quality is routinely compromised (Zayatz, 2005) as sensitive information, such as income, is typically rounded[6] or topcoded[7]. A good example of the resultant difficulties is illustrated in a paper by Stuart Soroka and Chris Wlezien entitled 'How Measures Matter' (2002). The authors ran the same model on three different quality UK budget datasets: the unadjusted data (i.e. what is reported by the UK Government to OECD); data adjusted by simply treating public corporations consistently, and the full adjustments for backward compatibility. The first model yielded insignificant results in the wrong direction. The second yielded insignificant results in the right direction. The third confirmed the model. It is worth noting that, despite the potential consequences, few, if any, statistical agencies inform researchers about the potential consequences of disclosure protection techniques and edits on the quality of their analysis (Kennickell & Lane, 2006).

Each of these access modalities is very far from the ideal described in the first paragraph of this section. Yet advances in the computer sciences could be used to address access issues in a more scientific manner than the two examples outlined above. Indeed, there is no technical reason why researchers could not access confidential data remotely from their offices. Protecting databases against intruders has a long history in computer science. Computer scientists themselves are interested in the protection and the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses).[8] [9] [10]

Despite these computer science advances, neither the national nor the international statistical community has adopted them. Indeed, the Conference on European Statisticians developed a set of guidelines that proposes a set of organizational, statistical and legal guidelines for statistical agencies to follow (Lane, 2003; UNECE, 2006)[11] with no reliance on the new cybertechnologies that could be used to address the data access challenges[12]. Discussions with both U.S and international statistical agency heads lead us to conclude that a major reason for their hesitation is the difficulty of guaranteeing that researchers who access data remotely will protect the data, as well as the difficulty in assuring their respondents of a guarantee of such protection[13].

[6] Ranges of income are provided: e.g. 0-$1,000; $1,000-$4,999; etc.

[7] High income levels are replaced by a top code (e.g. incomes over $100,000 are simply coded as "over $100,000).

[8] http://abilene.internet2.edu/observatory/

[9] Carl Landwehr pointed this application out to us.

[10] SBE/CISE workshop, Match 15-16 2005, http://vis.sdsc.edu/sbe/About

[11] See UNECE "Managing Statistical Confidentiality And Microdata - Guidelines Of Good Practice" and the keynote address "The Uses of Microdata" by Julia Lane to the Conference of European Statisticians, June 12, 2003.

[12] An excellent related discussion is provided by Weber, in Values in a National Information Infrastructure: A Case Study of the U.S. Census (2005).

[13] See "Privacy as Contextual Integrity" (Nissenbaum, 2004).

Approach

The NORC data enclave represents a first attempt to provide remote access to confidential microdata on businesses for federal statistical agencies. It has begun to pilot a portfolio approach to data access, that includes some statistical protection (mainly deleting obvious identifiers), screening of researchers, training researchers in legal and ethical confidentiality requirements, and both secure onsite and remote access.

One of the key features of the enclave is a collaborative environment within which researchers can share knowledge about the data and hence provide information back to the producer that can be used in archiving and curating the data. Getting researchers to participate in archiving and curation has historically been a challenge, primarily due to lack of incentives, particularly academic credit as well as the time cost of documentation, lack of funding, and lack of standards. Our approach addresses three of these issues, by reducing the time cost of documentation, applying standards, and creating incentives to provide metadata. Our approach builds on our experience working with data producers and implementing the DDI-based Microdata Management Toolkit. Developed by the World Bank for the International Household Survey Network (IHSN), the Toolkit is an extremely user-friendly package that facilitates the archiving of microdata in compliance with the DDI specification. Since its initial release in 2006, it has been adopted by several national statistical agencies, as well as by NORC, and will be used for the archiving of the UNICEF MICS 3 survey program (49 countries).

We have also developed incentives for researchers to document their metadata. Although the metadata are initially prepared together with the producer, the major focus has been to provide an appropriate environment for the researchers to transform the metadata into dynamic knowledge that continues to evolve over time. This is done in a number of ways. One way is to create positive incentives, similar to what has proven successful with the Social Science Research Network and other researcher communities. Researchers' contributions to metadata, whether it be final code, data edits, or any other related metadata documentation, will be indexed and attributed to the researcher, through standard citation. These contributions will be listed by author on the enclave website, and the researcher will be provided a monthly count of how many times the metadata are used (when applicable). Other researchers who use the metadata will be asked to cite the contribution in their published research. All associated research will, in turn, become metadata and associated with the contributing researcher. This approach is consistent with the National Science Board's recommendations. We will also create monetary incentives for researchers to work with metadata. Researchers who are particularly active, or most frequently cited, will be provided with discounts on additional storage or on disclosure review. In addition, we have encouraged a user-producer dialog by means of establishing a collaboratory within which both researchers and producers can discuss data issues and problems on both blogs and wikis. The discussion threads, in turn, become metadata.

Summary

Our approach to providing access to microdata is to combine technical, legal and organizational approaches to ensure that the confidential information provided by respondents is protected from access and use by unauthorized individuals and for unauthorized purposes. We have set up a remote access protocol that enables researchers to do their work in a research environment which facilitates high-quality research. We have also set up a collaboratory environment so that researchers enjoy both lower costs and greater benefits by documenting metadata. It remains to be seen whether this approach will be a success: the enclave only began allowing full researcher access in July 2007.

References

Berman, F., & Brady, H. (2005). *Final Report: NSF SBE-CISE workshop on cyberinfrastructure and the social sciences*. Retrieved on July 10, 2007 from http://vis.sdsc.edu/sbe/reports/SBE-CISE-FINAL.pdf

Kennickell, A., & Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the Survey of Consumer Finances. In J. Domingo-Ferrer and V. Torra (Eds.), *Privacy in Statistical Databases 2006.* New York: Springer-Verlag.

Lane, J. (2003). The uses of microdata. In *Statistical Confidentiality and Access to Microdata.* Geneva: United Nations Economic Commission for Europe, United Nations.

Madsen, P. (2003). The ethics of confidentiality: The tension between confidentiality and the integrity of data analysis in social science research. mimeo, Carnegie Mellon University: Center for Advancement of Applied Ethics.

National Science Board. (2005) *Long-lived digital data collections: Enabling research and education in the 21st century.* Retrieved May 3, 2005, from http://www.nsf.gov/nsb/meetings/2005/LLDDC_draftreport.pdf

Nissenbaum, H.F. (2004). Privacy as contextual integrity. *Washington Law Review, Vol. 79,* No. 1, 2004. Retrieved July 9, 2007 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=534622

Soroka, S.N., & Wlezien, C. (2002). Public expenditure in the UK: How measures matter. Retrieved July 9, 2007 from http://www.degreesofdemocracy.mcgill.ca/HowMeasuresMatter.pdf

UNECE. (2006). Managing statistical confidentiality and microdata access: Principles and guidelines of good practice. Geneva: United Nations Economic Commission for Europe. Retrieved July 9, 2007 from http://www.unece.org/stats/documents/tfcm/1.e.pdf

Weber, T.M. (2005). Values in a national information infrastructure: A case study of the U.S. Census. Retrieved July 9, 2007 from http://crypto.stanford.edu/portia/pubs/articles/W2087827446.html

Zayatz, T.A. (2005). Social Security Disability Insurance Program worker experience. Actuarial study No. 114. Office of the Actuary, SSA. Retrieved July 9, 2007 from http://www.ssa.gov/OACT/NOTES/pdf_studies/study114.pdf