

**RESOLVING THE TENSION BETWEEN ACCESS AND CONFIDENTIALITY:
PAST EXPERIENCE AND FUTURE PLANS AT THE U.S. CENSUS BUREAU**

by

Lucia Foster *
U.S. Bureau of the Census

Ron Jarmin *
U.S. Bureau of the Census

and

Lynn Riggs *
U.S. Bureau of the Census

CES 09-33 September, 2009

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.ces.census.gov or contact Cheryl Grim, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K130B, 4600 Silver Hill Road, Washington, DC 20233, Cheryl.Ann.Grim@census.gov.

Abstract

This paper provides an historical context for access to U.S. Federal statistical data with a primary focus on the U.S. Census Bureau. We review the various modes used by the Census Bureau to make data available to users, and highlight the costs and benefits associated with each. We highlight some of the specific improvements underway or under consideration at the Census Bureau to better serve its data users, as well as discuss the broad strategies employed by statistical agencies to respond to the challenges of data access.

* Foster: Center for Economic Studies, U.S. Census Bureau, Lucia.S.Foster@census.gov; Jarmin: Center for Economic Studies, U.S. Census Bureau, Ron.S.Jarmin@census.gov; Riggs: Center for Economic Studies, U.S. Census Bureau, Tammy.Lynn.Riggs@census.gov. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

1. Introduction

As economies and societies become more complex, policymakers, businesses and citizens increasingly need access to high quality and timely information for decision-making. National statistical offices (NSOs), such as the U.S. Census Bureau, play an important role in providing such information. The quality of statistical information produced by NSOs is correctly perceived to be of the highest quality due to their objectivity and strict adherence to professional and quality standards. A common and important contributor to the quality of data products released by NSOs is their ability to elicit high quality, original source information from survey respondents and administrative records providers under a pledge of confidentiality. The promise to maintain the confidentiality of data on households and businesses provided to them allows NSOs to achieve high response rates and to elicit truthful responses. For this reason, maintaining the confidence of the public is critical for NSOs. If NSOs are perceived, by the public or businesses, to be disclosing too much information, participation in data collection activities could be adversely affected. Since it is expensive to collect the information – the life cycle cost for the 2010 U.S. Decennial Census alone has been estimated at \$13.7 to \$14.5 billion¹ – obtaining inaccurate information is an inefficient use of resources. Moreover, even small changes in response rates to programs like the Decennial Census imply significantly greater collection costs. Yet, given the costs, not using these data to their fullest extent is also inefficient.

Hence, an unfortunate reality of NSO data is that one of the primary contributing factors to their quality, comprehensiveness, and cost effectiveness – the pledge of confidentiality – also significantly restricts how and by whom the data can be used. In a democratic society, such as the United States, data collected at public expense by Federal NSOs are valued only to the extent they are used. Most U.S. Federal NSOs are required by law to maintain the confidentiality of the data they collect. Doing this requires restricting the amount of information NSOs can disclose to the public. Thus, NSOs face a tension between their goal to provide data users complete, accurate and timely data and their duty to protect the confidentiality of the data entrusted to them by households and businesses.

This tension has increased in recent years as computing costs fall, making it easier for NSOs to make large amounts of data publicly available to users, especially through the Internet, and making it easier for users to access and use large amounts of data. This, however, creates a double-edged sword since it is now easier for people to bring together large amounts of information from disparate sources. Employing advanced, but widely available, analytic methods, data users are increasingly able to re-identify individual units from combined data sources that do not by themselves reveal any confidential information. Therefore, NSOs are increasingly constrained in making data more available while still protecting the confidentiality of respondents. In response to these constraints, many NSOs have moved to multiple modes of data access, where in some cases, an intentional strategy is pursued to deliver data via different modes of access in order to satisfy the needs of diverse users. For example, the Census Bureau provides not only the public-use tabulations that are broadly available, but also the option for users to

¹ U.S. Census Bureau (2008b).

request special tabulations² and to use on-line public-use microdata files for some data sets. These modes of access typically satisfy the needs of users looking for specific pieces of information on a given topic as well as many research needs. For researchers who need more information than is provided publicly, the Census Bureau can provide non-employee access to internal data at secure facilities (e.g., Census Bureau Headquarters, Census Bureau Research Data Centers).³ The primary requirement for non-employee access, by law, is that this research must benefit the Census Bureau data programs in some way.

NSOs, in general, do benefit from granting access to their internal data in terms of improvements in the quality and utility of their data products. First, access to microdata can provide substantial improvements to a statistical agency's data products, methodologies, and underlying sampling frames. As succinctly stated by the Panel on Data Access for Research Purposes, “[r]esearchers’ use of government data creates an effective feedback loop by revealing data quality and processing problems, as well as new data needs, which can spur NSOs to improve their operations and make their data more relevant” (NRC, 2005). One of the best means by which the Census Bureau can check on the quality of the data it collects, edits, and tabulates is to make its micro records available in a controlled, secure environment to sophisticated users who, by employing the micro records in the course of rigorous analysis, will uncover the strengths and weaknesses of those micro records. Each set of observations is the end result of dozens upon dozens of decision rules covering definitions, classifications, coding procedures, processing rules, editing rules, disclosure avoidance rules, and so on. Therefore, the validity and consequences of all these decision rules only become evident when the Census Bureau's microdata are tested in the course of analysis. Exposing to the light of research the conceptual and processing assumptions that are embedded in the Census Bureau's microdata constitutes a core element in the Census Bureau's commitment to quality and improves the publicly-available microdata and tabulations that are available to the broader community of users.

By providing qualified researchers access to confidential microdata, NSOs gain access to knowledge, skills and experience not available within the agency while also enabling research projects that would not be possible without access to respondent-level information. Hence, increasing the amount of research conducted increases the value of data that has already been collected and reduces inefficiencies of having multiple collections of the same information. Access to the microdata also allows for data linking not possible with aggregates – both cross-survey linkages and longitudinal linkages – which also leverage the value of preexisting data. Creative use of microdata can address important policy questions without the need for additional data collections. For example, the research done by Davis, Haltiwanger, and Schuh (1996) on job creation and job destruction in U.S. manufacturing provided the first comprehensive look at the demand

² These special tabulations are typically reviewed by the Census Bureau’s Disclosure Review Board to ensure that there is no primary or secondary disclosure risk from the data being released.

³ The research being conducted is thoroughly vetted by the Census Bureau to ensure that the research meets the legal mandates that protect the data. Moreover, all researchers must go through a complete background check just like employees and are subject to the same penalties. This will be discussed further in Section 2.

side of labor and the reallocation of workers in a market economy. This research would not have been possible without access to the microdata nor without the expertise of the authors.

The rest of this paper is organized as follows. Section 2 briefly describes the history of data access from the Census Bureau. Section 3 uses the Longitudinal Business Database as an example to illuminate current Census Bureau methods for providing data access. Section 4 describes future plans to further increase data access using the Longitudinal Business Database as example. Section 5 concludes.

2. History of Data Access from the U.S. Census Bureau

In the early population censuses (1790-1840), the law required the public posting of individual responses in order for respondents to ensure that their information had been reported accurately. Access to data collected by the Census Bureau and the laws governing this access have changed dramatically since then, with data access currently governed by legal protections restricting access only to individuals sworn to protect the confidentiality of respondents and only for statistical purposes. The changes in these laws are primarily attributed to 1) the increasingly sensitive nature of the information being collected and 2) the use of these data for purposes other than the stated statistical purpose under which the data were collected (U.S. Census Bureau, 2001). In the early censuses (1790-1840), the data collected during the population census was very basic – the county, parish, township, town, or city where the family resided; the name of the head of the family; a statement for each family of the number of free white males and females of various ages; the number of all other free persons (except Indians not taxed); and the number of slaves. In 1850, additional questions were added to the population census to collect more sensitive information (e.g., literacy within the household, disabilities of household members, value of real estate, occupation). The U.S. marshals, who conducted the interviews and collected the data in 1850, were also instructed to compile supplemental information to produce social statistics, which included information on taxes, schools, crime, and mortality (U.S. Census Bureau, 2002; Wright and Hunt, 1900).

While public concern about confidentiality of responses to the early population censuses seemed to be low, confidentiality concerns by respondents in the early censuses of manufactures, which began in 1810, was much higher. In fact, questions about manufacturing were not included in the 1830 census due to the poor response rates and low quality of responses collected in the 1820 census. In 1840, the manufacturing questions were once again included, but the marshals collecting the data were instructed to consider the information related to the business of people as confidential and to assure respondents that no names (whether of an individual or of a business) would appear in any statistical tables (U.S. Census Bureau, 2001).

While the individual results were no longer posted after 1840, access to the population censuses was unrestricted until the Permanent Census Act in 1902, which left access to census records at the discretion of the Director of the Census Bureau. In 1904, the Census Bureau instituted a policy stopping all public access to the original records, not due to confidentiality concerns but primarily due to concerns about the deteriorating

condition of the records caused by handling.⁴ It was not until the Fifteenth Decennial Census Act in 1929 that the use of the data collected by the Census Bureau was limited specifically to statistical purposes, with a provision for the release of records to a state/territory or to a court left to the discretion of the Director under certain conditions, including “That in no case shall information furnished under the authority of this Act be used to the detriment of the person or persons to whom such information relates” (U.S. Census Bureau, 2003a, p. 4).

The Fifteenth Decennial Census Act was the precursor to the current Census Act, Title 13, United States Code⁵ that was passed in 1954 and became effective on January 1, 1955. Title 13 provides detailed laws concerning all aspects of the Census Bureau including authorization for the collection and maintenance of data as well as the conditions and limitations placed on these activities. Hence, Title 13 provides legal protections for the data collected by the Census Bureau, with penalties for unauthorized disclosure of Title 13 information (fines up to \$250,000 and/or jail terms of up to five years). This unauthorized disclosure includes disclosing or publishing any private information that identifies a business or individual.

Using Title 13 as a basis, the Census Bureau has developed its privacy principles as guidelines to help ensure that the Census Bureau maintains the trust of its respondents and follows the legal mandates that protect respondents’ information. These guidelines include the following:

- necessity (only collecting data that are needed),
- openness (informing respondents about how their data will be used),
- respect for respondents (minimizing respondent burden, using only legal, ethical, and professional practices, and adhering to federal protections for sensitive populations),
- and confidentiality (using technology, statistical methodology, and physical security procedures to protect data).

The statistical safeguards that the Census Bureau uses include review and analysis of all products including the use of disclosure avoidance techniques such as cell suppression and noise infusion.

Title 13 is only one statute that protects the confidentiality of data collected by Federal NSOs, just as the Census Bureau is only one agency in the highly decentralized U.S. statistical system. In the U.S., there are over 100 NSOs producing information.⁶ Other Federal NSOs in the U.S. are also legally required to protect identifiable, individual-level information based on a number of different statutes, though the rules concerning data access vary across the agencies. Moreover, much of the legislation to increase protections on these data is recent. The Privacy Act, which among other things, limits and restricts

⁴ It was still possible to obtain transcripts of these records for a fee, though.

⁵ The United States Code is the codification of the general and permanent laws of the United States and is divided into 50 titles.

⁶ FEDSATS links to more than 100 agencies in the U.S. government see <http://www.fedstats.gov/agencies/>.

the disclosure of personally identifiable individual-level data that is maintained by various Federal government agencies was not passed until 1974. The Health Insurance Portability and Accountability Act (HIPAA) was passed in 1996 to regulate the confidentiality of medical records.⁷ Most recently, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) was passed in 2002 to establish minimum standards for confidentiality protections for information collected by federal NSOs, since very few of the existing penalties were as severe as those established by Title 13. “CIPSEA thereby provides statutory protection to the many NSOs that previously had only custom or other nonstatutory authority to back up pledges of confidentiality” (National Research Council, 2005, p. 23). While this is only a brief summary, the legal and policy environment for data access in the U.S. is described in more detail in the Appendix.

The following sections describe the various modes of data access employed by the Census Bureau to meet the needs of diverse users.

2.1 Public-Use Tabulations

Data users often do not need access to confidential micro data files to meet their needs, instead their needs can be met by standard aggregations of the micro data. In these cases, these users can use Census Bureau public-use tabulations. The Census Bureau publishes numerous reports summarizing its micro data from households and businesses. In addition, the Census Bureau also provides online tabulations from these micro data. The American FactFinder (AFF) provides access to tabulations from Decennial Census, American Community Survey, Economic Census, and Annual Economic Surveys. In addition to tables, AFF also provides tools allowing users to create thematic maps. The DataFerrett, a data mining and extraction tool that searches across publicly available datasets, allows researchers to create their own customized tables.⁸

Users interested in tabulations concerning business dynamics (including measures of establishment entry and exit, job creation and destruction, and job expansions and contractions) can access these data from the Census Bureau’s Business Dynamics Statistics (BDS) website. The BDS tabulations are created from micro-level longitudinal data that permit tracking of establishments and firms over time thus allowing for a detailed portrait of the dynamics of businesses.⁹ These tabulations are available for 1976-2005 and over a variety of establishment and firm characteristics (including firm age, firm size, state, and industry classification). Data users interested in quarterly variation in business statistics can access the Quarterly Workforce Indicators (QWI). The QWI provides data users with information on thirty labor force indicators using information from the state-federal partnership known as Local Employment Dynamics.

⁷ In general, the HIPAA Privacy Rule applies to health plans, health care providers, and health care clearinghouses, yet it is pertinent since it covers the transmission of medical records from these organizations to NSOs to be used for public health and research purposes.

⁸ DataFerrett is available at <http://dataferrett.census.gov/>.

⁹ The dataset underlying the BDS is the Longitudinal Business Database (LBD) which is described in detail below. The BDS website is located at: <http://www.ces.census.gov/index.php/bds>.

More generally, the current administration has set-up a website, www.data.gov, that will bring together databases from across federal agencies. Currently, Census Bureau data hosted on this website is limited in scope, covering only information derived from the American Community Survey.

2.2 Household Public Use Microdata

The first public-use microdata were from the 1950 Censuses of Population and Housing, released by the Census Bureau on punchcards (U.S. Census Bureau, 2001). Yet, the release of the 1960 Public Use Sample and summary tape files was the first release of today's public-use microdata files. There have also been efforts, like the Minnesota Historical Census Projects, to create public use microdata files from earlier censuses for which public use files were not originally released (Ruggles and Menard, 1995). To prevent identification of individuals in the sample in the years still protected¹⁰, however, certain precautions must be taken which can limit the usefulness of the data for research purposes.

The Census Bureau uses a number of methods to protect the identity of individual respondents when releasing household public-use microdata, and these methods have become more rigorous over time. "Because of the rapid advances in computer technology since 1990 and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through enhanced disclosure limitation techniques" (U.S. Census Bureau, 2003b, p. 2-1). One of the primary protections is to provide only a sample of records to users, rather than the full file. In both 1990 and in 2000, the Census Bureau made 1% and 5% Public Use Microdata Samples (PUMS) available. However, in 2000, the minimum geographic threshold was raised from 100,000 to 400,000 for the 1% PUMS in order to provide similar detail as was provided in 1990. In the 2000 5% PUMS data, the geographic threshold was held at 100,000 (as it was in 1990), but the trade-off was increasing the degree of variable collapsing in the data. Other protections include data-swapping, topcoding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables (U.S. Census Bureau, 2003b).

PUMS data make it possible for many users to create their own special tabulations as well as to conduct research using the files. These microdata samples are extremely useful for creating statistics when small geographic areas or detailed crosstabulations for small populations is not necessary. Users of PUMS data "have almost the same freedom to manipulate the data that they would have if they had collected the data in their own sample survey, yet these files offer the precision of census data collection techniques and sample sizes larger than would be feasible in most independent sample surveys" (U.S. Census Bureau, 2003b, p. 2-2). Users requiring more detail require a different mode of access, either a special tabulation or non-employee access to internal data.

¹⁰ Individual records from the population censuses can be made publicly available on microfilm after 72 years. For example, data from the 1930 census were made public on April 1, 2002, through the National Archives.

2.3 Business Public Use Microdata

While the Census Bureau has been able to make public use microdata available for individuals and households, it has not been able to make available comparable public use microdata files for the data it collects on businesses. The skewed nature of business information and the public visibility of many firms make businesses easy to identify, and any protections imposed on the data tend to make them less than useful for most research purposes, even in aggregate tabulations. For example, in the County Businesses Patterns data released by the Census Bureau in 2002, approximately 1.5 million county-industry cells had exact employment figures suppressed (Isserman and Westervelt, 2006). To combat this, the Census Bureau has been investigating the use of synthetic data using noise infusion to protect individual respondents' information. For example, the product OnTheMap, a web-based tool that allows users to produce maps and reports concerning where people live and work, is based on synthetic data with statistical noise infusion.¹¹ For some types of research, the synthetic data may suffice; however, there will be research projects for which only the internal microdata can be used.

2.4 Non-employee Access to Internal Data

The Census Bureau is allowed through Section 23(c) of Title 13 to use non-Census Bureau employees in assisting the Census Bureau conditional on those non-Census Bureau employees being sworn to observe the limitations outlined in Section 9 of Title 13. These sworn non-Census Bureau employees are referred to as having "Special Sworn Status (SSS)." In order to obtain SSS status, a researcher must undergo a full background check. The Census Bureau has long used SSS individuals to help with all aspects of data collection and tabulation, as well as with some research projects. However, it was not until the early 1980s that the Census Bureau began to make internal data available to SSS individuals on a more regular basis for research.

A key event in the history of non-employee data access was the opening of the Center for Economic Studies (CES) in 1982. CES was established to house new longitudinal business databases, develop them further, and make them available to qualified researchers with SSS. John Haltiwanger, the Census Bureau's first Chief Economist, noted that "Shirley Kallek, who was the Associate Director for Economic Programs in the early 1980s when CES was founded ... argued convincingly that the Economic Directorate was sitting on a goldmine of micro business data that needed to be developed and analyzed to help understand the underpinnings of U.S. economic growth and fluctuations" (U.S. Census Bureau, 2008a, p.vi). Qualified academic researchers were thus able to visit the Census Bureau to begin fulfilling those visions. Using these new longitudinal databases, they produced analyses that contributed to a revolution of empirical work in the economics of industrial organization.

2.4 Research Data Centers

The establishment of Census Bureau Research Data Centers (RDCs) greatly expanded researcher access to confidential microdata while ensuring the strict terms of access required by the legislation and regulations governing the Census Bureau and other

¹¹ OnTheMap was introduced in 2006 and is updated annually using data from a voluntary federal-state partnership known as the Local Employment Dynamics. See <http://lehd.did.census.gov/led/>.

providers of data.¹² The RDCs are secure Census Bureau controlled locations with a Census Bureau employee onsite acting as an administrator. The Census Bureau RDC Network currently includes nine facilities with plans for additional facilities.¹³ RDC researchers must obtain Special Sworn Status and they must have an approved project of benefit to the Census Bureau. Proposals for access to data within the RDCs are subject to rigorous review that requires meeting the following five criteria: benefit to the Census Bureau, feasibility, scientific merit, need for non-public data, and disclosure risk avoidance. Proposals to use data collected by or for other agencies generally must also be reviewed by the data provider.¹⁴ RDC researchers are required to conduct all of their research at the RDC and only aggregate information (in the form of summary statistics and model estimates) are released. There are currently more than 400 active researchers associated with the Census Bureau's RDC Network.

Given the success of the RDC model in increasing research in the business realm, the Census Bureau expanded the data made available within the RDCs in the late 1990s to include individual and household data. Since that time the amount of data available through the RDC Network has expanded both in terms of surveys and censuses available as well as years of data within these surveys and censuses. Two federal agencies, the National Center for Health Statistics (NCHS) and the Agency for Healthcare Research and Quality (AHRQ), have recently partnered with the Census Bureau and have begun to make their data available through the Census Bureau Research Data Centers. By building on existing infrastructure, NCHS and AHRQ have benefited from expanded access to their data across the country for a minimal expenditure.¹⁵ The RDC Network has gained by increasing utilization and expanding the base of support beyond the traditional social sciences and into schools of health. The research community, primarily in health-related fields, has benefited from expanded access to incredibly rich microdata with the ability to conduct this research closer to their home institutions.

The Census Bureau has worked to reduce the costs to researchers of using an RDC. These costs include lab fees, the time involved in preparing the proposal to conduct the research project, the time involved in learning the structure of the data, and usually travel costs. Hence, there is a trade-off to obtaining more detailed information than is publicly available. One innovation that helps to reduce these costs was developed by researchers at Cornell University. These researchers have developed a computational infrastructure that is known as the VirtualRDC. The VirtualRDC is designed to "provide access to

¹² Other statistical agencies in the U.S. have research data centers but the Census Bureau's RDC network is the largest in scope. Some statistical agencies outside the U.S. have research data centers. For example, Statistics Canada has an extensive RDC network with 26 locations. France's INSEE has plans to operationalize an RDC by 2010.

¹³ The existing nine RDCs are located in the following cities: Washington DC, Boston MA, New York NY, Ithaca NY, Raleigh-Durham NC, Berkeley CA, Los Angeles CA, Ann Arbor MI, and Chicago IL. The Stanford University branch of the Berkeley RDC will open in Fall 2009 another RDC is scheduled to open in early 2010.

¹⁴ The CES website contains more information about the proposal review process: <http://www.ces.census.gov/index.php/ces/researchprogram>.

¹⁵ NCHS does have a remote submission program for researchers to submit analyses and receive the output from these analyses. However, there are limitations to this system. For more information, see the NCHS remote submission webpage: <http://www.cdc.gov/nchs/r&d/rdcremote.htm>.

synthetic data over the internet, to assist potential Census RDC users in preparing their proposals, and to train new users in the operating system environment, data, and software available on the real Census RDC (<http://www.vrdc.cornell.edu/news/welcome/>).” No confidential information is available on the VirtualRDC, but users can minimize the amount of time that they actually need to be in the physical RDC by conducting research using synthetic data and writing programs using zero observations data sets outside of the physical RDC. For some users, the synthetic data may be sufficient; hence, those users would never need to use the physical RDC. Over 500 users have accessed information on the VirtualRDC.

2.5 Microdata Analysis System

The Census Bureau is in the process of developing a remote Microdata Analysis System (MAS) where users will be able to conduct statistical analyses on microdata without seeing the actual underlying observations. Currently the goal is to include all of the Census Bureau’s data sets and to allow all types of analyses (Lucero, Singh, and Zayatz, 2009). To use the MAS, users would submit queries to the server in order to obtain statistical results. These results would be automatically checked by the system to ensure that the results pose no risk of disclosure. Currently the Census Bureau has a prototype of MAS using data from the Current Population Survey March 2000 Demographic Supplement and the American Community Survey, with some basic statistical analyses available for use (e.g., correlation coefficients, ordinary least squares regression) (Lucero, Singh, and Zayatz, 2009). The confidentiality protections required are very complex since they need to prevent users from using combined results from multiple analyses to identify information about individual respondents.¹⁶

When the MAS is available, it will be another mode of access; however, as with the other modes of access, there will be limitations to its use. Data users need to consider all of the modes of access as well as the trade-offs between increasingly detailed information and the cost of access to that information.

3. Multi Access Mode Strategies: The Longitudinal Business Database

Data users are heterogeneous and agencies must provide a range of data products via a variety of modes. Before computers, users accessed only aggregated tabulations of basic variables that were available through official print publications. NSOs, such as the Census Bureau, began making Public Use Micro Sample (PUMS) files available to academic and other sophisticated users as universities and other large institutions installed mainframe computers. Today, however, nearly every citizen living in a developed economy has access to the computing power needed to process large quantities of data.

Data users today rarely access the products of NSOs via printed publications. Users acquire data electronically, typically via the Internet, and analyze and manipulate them on cheap, ubiquitous and networked computers. Since most users now, or will soon, possess

¹⁶ The MAS is actually an example of a collaborative effort between the Census Bureau and external experts from academia. The Census Bureau has contracted with academics who are experts in the fields of disclosure analysis and statistics in order to develop the confidentiality protections.

the technology, if not the training and skills, to obtain and analyze even the detailed microdata that underlie official statistics, the main consideration for NSOs is how to maximize the amount of information released while maintaining confidentiality.

We use the Census Bureau's Longitudinal Business Database (LBD) as an illustrative example of a data product where multiple access modes, each catering to a different subset of data users, are not only part of its dissemination strategy, but also a critical part of its development and ongoing improvement. The development of the LBD at CES responded to the need to extend longitudinal analyses of businesses beyond manufacturing once Census Bureau and academic researchers demonstrated the value of creating longitudinal files for manufacturing establishments (see Dunne, Roberts and Samuelson, 1989 and Davis, Haltiwanger and Schuh, 1996). Initially, the LBD was explicitly designed to meet the needs of researchers requiring high-quality linked longitudinal establishment data (see Jarmin and Miranda, 2002). The value added in longitudinally linking the annual snapshot files from the Census Bureau's Business Register became quickly apparent through several research papers that used the LBD to provide a rich, detailed portrait of the U.S. economy.¹⁷ Moreover, through the dissemination of these papers, it also became apparent that there was a strong demand for publicly accessible data products based on the LBD.

Given the many potential users of the LBD information, all with diverse needs and analytical abilities, it is clear that different modes of access are necessary to best serve this community. Currently, there are four different access modes through which users can obtain LBD products: 1) public-use tabulations available on the Internet, 2) special tabulations for international comparisons, 3) public-use synthetic micro data via the Cornell VirtualRDC, and 4) gold-standard confidential microdata via the RDC network.

The Census Bureau's Business Dynamics Statistics (BDS) program makes a number of public use tabulations from the LBD available to the general public via the BDS website (<http://www.ces.census.gov/index.php/bds>). The BDS data are unique in that they include tabulations by firm size and age and allow users to better measure the process of firm entry, growth, decline and exit. The Census Bureau also provides a special tabulation from the LBD according to agreed-upon specifications for the Organization for Economic Cooperation and Development (OECD). The OECD then makes these tabulations publicly available along with similar and comparable tabulations provided by other OECD countries for its Structural and Demographic Business Statistics and Entrepreneurship Indicators Programs.¹⁸ These publicly available tabulations provide users with information about the life-cycle of firms in the U.S. and also allow for international comparisons using the OECD data. For some users, this should suffice. However, for research that requires this information at a more detailed geographic or industry level, this information may not meet their needs. For that, users may require

¹⁷ Recent examples include: Davis, Haltiwanger, Jarmin and Miranda (2006) and Haltiwanger, Jarmin and Krizan (forthcoming).

¹⁸ See http://www.oecd.org/document/17/0,3343,en_2649_34233_36938705_1_1_1_1,00.html and http://www.oecd.org/document/0/0,3343,en_2649_34233_39149504_1_1_1_1,00.html

access to microdata, whether it be synthetic or the gold-standard, in order to accomplish their research objectives.

Synthetic data can be especially useful for business information since it is extremely difficult for NSOs to release public-use business microdata products that are not synthesized due to the ease of re-identifying large units. Hence, recent developments by statisticians provide agencies with the ability to create synthetic versions of microdata products that can be released. The basic idea is to fit models for the sensitive information in the collected data, simulate replacement values from these models, and release the simulated data for public use. This can preserve confidentiality, since identification of businesses and their sensitive data is very difficult when the released data are not actual, collected values. Furthermore, with appropriate data generation methods, this approach enables data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. The Synthetic LBD (see Kinney et. al. 2009) was developed as part of a National Science Foundation funded project and was a large collaborative effort involving statisticians, economists and computer scientists from around the U.S. The first version of this innovative new product was approved for release in the summer of 2009 and should be deployed to the Cornell University VirtualRDC (<http://www.vrdc.cornell.edu>) by the end of 2009. To test and improve the analytic validity of the Synthetic LBD, users will be asked to provide code run on the synthetic data to be run on the gold-standard confidential LBD, so the results can be compared. In this way, researchers can test the limitations of the synthetic data to better determine when access to the gold-standard microdata is required.

The full set of gold-standard confidential microdata from the LBD are available to approved researchers working on approved projects through the Census Bureau's RDC Network. To date, there have been more than 400 researchers associated with close to 90 active and completed projects who have used the gold-standard LBD data.

4. Future Access Modes

Given the ever-changing landscape in technology, it is hard to anticipate how the world of data access will look even a few years from now. At the Census Bureau, we are continuously working on improving existing modes of access as well as exploring new modes of access. By staying connected to the international community, we have tried to stay current on all of the latest developments in this arena.

4.1 Improvements to Existing Modes

Improvements to public use tabulations include providing interactive graphic displays of tabulations in order for researchers to see the effects of changing certain parameters (e.g., time). For the Business Dynamics Statistics, we are currently developing interactive national maps that users can change over time. This allows users to visualize the dynamics of these changing statistics without having to download the full data set. In the past, slower connection speeds and slower CPUs would have made viewing these graphs intolerable for the average user, but advances in technology now make it possible to provide information that is more visual. These types of graphics can also provide information to users in a way that adds some protection to the underlying data since users

do not necessarily need access to the full information to examine trends over time. Moreover, as with OnTheMap, the underlying data can be based on synthetic data without significant information loss.

In general, the Census Bureau is exploring further development of synthetic data, both in terms of new data sets as well as improving existing synthetic data. While there are some synthetic data products like OnTheMap, the Synthetic LBD, and the SIPP Synthetic Beta File¹⁹, there is still more work to be done. Work is on-going to compare results from analyses using the synthetic data to results using the microdata. This will help the development of products on two fronts: 1) to determine the limitations of the synthetic data, and 2) to determine how much detail can be provided to make the synthetic data more useful for research purposes. For the Synthetic LBD, researchers are investigating of feasibility of providing more geography (currently the data have been released at the national level) and additional firm characteristics.

The Census Bureau is also developing an automated system to produce the special tabulations created for the OECD's Structural Business Statistics and Entrepreneurship Indicators Programmes. The automated system will produce tabulations on job creation and destruction, survival, and growth using the LBD that are directly comparable with analogous statistics from other countries.

There have also been several improvements to the RDC Network that have lowered the "costs" of access to researchers. We have developed an on-line project management system, including a disclosure avoidance review tracking system (DARTS), in order to better track the status of proposals throughout the review process as well as other requests once the proposal has been approved. We are continuously adding new data products as well as years of existing data sets. For example, we have recently begun to make the Longitudinal Employer-Household Dynamics micro data available to researchers. By linking employers to their employees, these data allow researchers to examine the combined dynamics of employers and their employees when analyzing the economy. There have also been recent improvements to the computing environment, with the installation of clustered servers and increased disk space for those clustered servers. By clustering servers, we reduce redundancy by not cordoning off resources for individual RDCs. Hence, if one RDC is at peak load and another RDC is experiencing a lull (which happens frequently given that the RDCs span multiple time zones), the resources for the RDC experiencing the lull are not simply sitting idle. Moreover, the RDC Network facilitates collaboration by researchers across the country by allowing all researchers on the same project access to the same project space and data files. Improvements in these four areas have greatly improved data access at the Census Bureau, but we are also looking into new modes of data access.

¹⁹ The SIPP Synthetic Beta file was produced as part of a collaborative effort between the Census Bureau, the Social Security Administration, and the Internal Revenue Service. For more details, see the Census Bureau's website on this topic: http://www.census.gov/sipp/synth_data.html.

4.2 Exploring New Modes

There are constantly new innovations to be considered for providing data access. Some data providers are currently using remote desktop and/or remote submission to enable users to access microdata. For example, the National Opinion Research Center (NORC) has developed a remote desktop system that allows users to log into data servers to conduct analyses directly on the servers. In so doing, the system prevents the user from downloading, copying, or printing information from the server on their local personal computer.

While the Census Bureau is currently developing one type of remote access system (the MAS), there are other types of systems that the Census Bureau will also explore. In all of these systems, researchers are able to use microdata without ever having direct access to the microdata. The difference in the systems lies in the amount of interaction required by the host agency. Thus, all aspects of the process under the MAS will be done automatically and thus require very little input from the host agency. However, there are other remote access systems where users submit programs to be run on the microdata but employees at the host agency run the analysis and conduct the disclosure review. For example, this is the type of remote access system used at NCHS.

Another potential new mode of access relies on an innovation in personnel. Most projects at the RDCs include graduate students working for a more senior researcher. One of the costs of using the confidential data is the start-up costs involved in learning about the micro data which often are not as well documented as public-use datasets. These graduate students develop expertise in the confidential data associated with their project that could be used in other projects. One idea is to have graduate students at RDCs who are not dedicated to one particular project, but who instead are available to work on projects using data within their sphere of expertise. In this manner, researchers at remote locations could direct the empirical work of the graduate student research assistant working onsite at the RDC.

5. Conclusions

Given the conflicting objectives of increasing data access and increasing data protections, no one data access mode will ever serve as a panacea. For this reason, using different modes of access seems to provide the most efficient use of resources to meet the heterogeneous needs of data users. Moreover, innovation is the key to increasing access while maintaining the confidentiality of respondents – synthetic data and interactive graphic displays are recent innovations that help achieve both objectives. Since innovation tends to be driven by exposure to new ideas, sharing experiences and ideas with an already large, international community interested in the data access debate can help to improve both data access and confidentiality protections.

The NSOs of many countries have accumulated considerable experience in providing access to confidential microdata to authorized researchers for approved projects in secure

settings (safe people, safe projects, and safe locations). These researchers have successfully integrated a wide variety of survey data and administrative data sources to conduct rich analyses and assist NSOs in developing new data products.

In the case of firm and establishment data, analyses conducted on confidential statistical microdata have fundamentally changed the way we view the functioning of modern market-based economies. This success, along with an increasingly global economy, has generated considerable interest in integrated microdata from multiple countries. NSOs typically capture the domestic activities of the business operating within their borders, but they often miss large parts of the activities of multinationals. Integrating and analyzing data from multiple countries would allow researchers to better understand the operations of multinationals and their role in both the global and national economies.

From a purely technical perspective, the access modes described above could be used to provide secure access to microdata from multiple countries. This has been done for selected demographic surveys for European Union member countries that are available for scientific analysis at a secure center in Luxembourg. We know of no case where business data from multiple countries are made available where they can be linked and integrated to provide a supranational view of firm operations. Currently, the legal and policy environments that most NSOs operate under preclude integrating and analyzing confidential microdata from multiple countries. Given these constraints, efforts have been made to construct harmonized public-use data products from the confidential data in each NSO. Excellent examples of this are the OECD's Structural and Demographic Business Statistics (SDBS) and Entrepreneurship Indicators (EIP) programs. These efforts are limited, however, and calls from the research and policy communities for analyses requiring integrated microdata from multiple countries will likely increase as the national economies become more integrated.

Appendix: Legal and Policy Environment

Federal agencies in the U.S., including the Census Bureau, are legally required to protect identifiable, individual-level information based on a number of different statutes. As described in Section 2, the data collected by the Census Bureau is protected by Title 13 of the United States Code. In addition to Title 13, there are many other legal protections for data provided to the Census Bureau, and in some cases, provided to researchers through Census Bureau facilities. These are outlined in the following section.

A.1 Confidentiality Restrictions and the Privacy Act, HIPAA, and CIPSEA

The codes concerning the collection, use, dissemination, and maintenance of person-level individually identifiable information are contained in the United States Code, Title 5, Section 552a (“The Privacy Act of 1974”). Among other things, the Privacy Act limits and restricts the disclosure of personally identifiable individual-level data that is maintained by various Federal government agencies. It allows an exemption for disclosure to the Census Bureau “for purposes of planning or carrying out a census or survey or related activity pursuant to the provisions of title 13” (Title 5, Section 552a.13(b)).” In addition to the Privacy Act of 1974, some Federal agencies are covered by Title 42, The Public Health and Welfare.²⁰ This specifically pertains to data provided to Census for internal use and for use by Special Sworn Status researchers in the Census RDCs by the National Centers of Health Statistics (NCHS) and by the Agency for Healthcare Research and Quality (AHRQ).

NCHS informs respondents to the National Health and Nutrition Examination Survey (NHANES) about the protections that are in place to protect their data:

“Public laws keep all information participants give confidential. We will hold all data we collect in the strictest confidence. We gather and protect all information in keeping with the requirements of Federal Law: the Public Health Service Act (42 USC 242k) authorizes collection and Section 308(d) of that law (42 USC 242m), the Privacy Act of 1974 (5 USC 552A), and the Confidential Information Protection and Statistical Efficiency Act (PL 107-347) prohibit us from giving out information that identifies you or your family without your consent. This means that we cannot give out any fact about you, even if a court of law asks for it. We will keep all survey data safe and secure. When we allow researchers to use survey data, we protect your privacy. We assign code numbers in place of names or other facts that could identify you.”²¹

In addition to these protections, some of the data collected by these agencies include medical records of individuals. The confidentiality of medical records is regulated under the Health Insurance Portability and Accountability Act (HIPAA) of 1996: “Identifiable medical information may be disclosed for research purposes only with the written consent of the person providing the information or in a limited set of circumstances in which an

²⁰ More information about Title 42 is available at http://uscode.house.gov/download/pls/Title_42.txt. See Chapter 6A as it pertains to the Public Health Service, including NCHS and AHRQ.

²¹ <http://www.cdc.gov/nhanes/pQuestions.htm#Is%20my%20information%20confidential>

institutional review board determines that the identifiable medical information is essential to the conduct of the research and the disclosure presents minimal risk to the individual” (NRC, 2005, p. 23). In general, the HIPAA Privacy Rule applies to health plans, health care providers, and health care clearinghouses, yet it is pertinent here since it covers the transmission of medical records from these organizations to agencies like NCHS to be used for public health and research purposes. These data then fall under the protections covering data collected by these agencies.

While there are protections in place for data collected by Federal agencies, very few of these have penalties as severe as Title 13. This was one impetus for the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002; hence, the first part of CIPSEA (Subtitle A) establishes minimum standards for confidentiality protections for information collected by federal agencies. “CIPSEA thereby provides statutory protection to the many statistical agencies that previously has only custom or other nonstatutory authority to back up pledges of confidentiality” (NRC, 2005, p. 23).

The following excerpt from Subtitle A—Confidential Information Protection, SEC. 511(b) explains the purpose of the subtitle:

(b) PURPOSES.—The purposes of this subtitle are the following:

- (1) To ensure that information supplied by individuals or organizations to an agency for statistical purposes under a pledge of confidentiality is used exclusively for statistical purposes.*
- (2) To ensure that individuals or organizations who supply information under a pledge of confidentiality to agencies for statistical purposes will neither have that information disclosed in identifiable form to anyone not authorized by this title nor have that information used for any purpose other than a statistical purpose.*
- (3) To safeguard the confidentiality of individually identifiable information acquired under a pledge of confidentiality for statistical purposes by controlling access to, and uses made of, such information.*

Under CIPSEA, the penalties for willful disclosure of specific information prohibited under this title are similar to those under Title 13 – including the possibility of a fine not more than \$250,000 and/or imprisonment up to 5 years.

Since Census Bureau confidential data are already under statutory protection (Title 13 has been enacted into positive [statutory] law), CIPSEA ultimately impacts the Census Bureau indirectly through the second half of the act – Subtitle B—Statistical Efficiency. data sharing arrangements. The second part of CIPSEA (Subtitle B) allows a limited number of federal agencies to share confidential business data for statistical purposes. Thus data access has been enhanced for the agencies that are currently covered by CIPSEA. This increases statistical efficiency by allowing the agencies to collect complementary data that can be shared to reduce the amount of resources expended on data collection (rather than having multiple agencies collecting similar information).

There are currently three agencies covered by CIPSEA: Census Bureau, Bureau of Economic Analysis, and Bureau of Labor Statistics. The first such data-sharing project matched data from BEA's surveys of Foreign Direct Investment the U.S. and U.S. Direct Investment Abroad with the Survey of Industrial Research and Development which is collected by the Census Bureau and sponsored by the National Science Foundation. By combining these datasets collected by different agencies, this pilot project "produced a potentially rich source of information for analyzing the consequence of Foreign Direct Investment (FDI) for innovation and other outcomes" (Atrostic, 2008).

A.2 Statistical Use of Administrative Records Data

Section 6 of Title 13 provides for the authority for programs under the Census Bureau to use administrative data from other agencies conditional on this use being protected not only by Title 13 but also by the protections required by the source agency. The Census Bureau relies upon administrative records data from the Internal Revenue Service (IRS), Social Security Administration (SSA), and the states to fill in crucial areas that are not covered by Census Bureau collections. The Census Bureau uses IRS business data for single-unit firms in its register of all businesses in the U.S. (the Business Register). Using IRS data, for example, means that the Census Bureau must follow rules outlined under Title 26. Datasets, such as the Business Register, that include both Census Bureau data and IRS data, are considered to contain commingled data. The commingled data are then all considered Title 26 protected data.

On the individual side, the Census Bureau uses data from both IRS and SSA to maintain its listing of individuals and households as well as for research purposes. These research purposes include using the administrative records to examine the accuracy of information reported by respondents. For example, the SSA provides earnings information to examine the accuracy of income reports in the Survey of Income and Program Participation.

The Longitudinal Employer-Household Dynamics (LEHD) program at the Census Bureau creates innovative statistical products using linked employer-household data. The LEHD program combines federal and state administrative data on employers and employees with Census Bureau surveys and censuses. The Census Bureau enters into agreements with individual states to share data under the Local Employment Dynamics (LED) Partnership. Under these agreements, the states provide administrative records on workers and employers to the Census Bureau. These state records include the state's Unemployment Insurance (UI) wage records and the ES-202/Quarterly Census of Employment and Wages (QCEW) employer records. The ES-202/QCEW data are cleansed of any CIPSEA-covered data elements. Once they have been combined with the Census Bureau data, the confidentiality of the state data is protected under the Privacy Act and Titles 13 and 26 U.S.C.

The Census Bureau and the state partners are currently in negotiations to develop a National Memorandum of Understanding (MOU) that would unify and harmonize the individual state agreements into one national agreement. In addition to the protections noted above, the National MOU outlines provisions for producing publicly available

documents that use these data. Documents that use state data in an analysis at the state-level or the sub-state level are subject to state review. By contrast, documents that use state data in an analysis that is aggregated to a multi-state or national level do not require state review.

A.3 IT and Physical Security

The Federal Information Security Management Act (FISMA) of 2002 requires the head of each federal agency to provide information security protections commensurate with the risk and magnitude of harm that would occur due to unauthorized data use. Each agency must “develop, document, and implement an agencywide information security program ... to provide information security for the information and information systems that support the operations and assets of the agency.” The National Institute of Standards (NIST) developed the standards and guidelines for security. Each agency must delegate to its Chief Information Officer the authority to ensure compliance with FISMA (<http://csrc.nist.gov/drivers/documents/FISMA-final.pdf>).

In 2004, the Homeland Security Presidential Directive 12 (HSPD-12) established a federal standard for secure forms of identification for employees and contractors who are seeking to gain access to federal secure facilities. Implementation of the directive must be done in “a manner consistent with the Constitution and applicable laws, including the Privacy Act and other statutes protecting the rights of Americans” (http://www.dhs.gov/xabout/laws/gc_1217616624097.shtm)

References

- Atrostic, B.K. (2008). "Measuring U.S. Innovative Activity: Business Data at the U.S. Census Bureau," *The Journal of Technology Transfer* 33:2, 153-171.
- Davis, Steven J., John Haltiwanger and Scott Schuh (1996). *Job Creation and Destruction*, Cambridge: MIT Press.
- Davis, Steven J., John Haltiwanger, Ron S. Jarmin and Javier Miranda (2007). "Volatility and Dispersion in Business Growth Rates: Publicly Traded vs. Privately Held Firms," *NBER Macroeconomics Annual 2006*, Vol. 21.
- Dunne, Tim, Mark.J. Roberts and Larry. Samuelson (1989). "The Growth and Failure of U.S. Manufacturing Plants," *Quarterly Journal of Economics*, Vol. 104, pp.495-515.
- Haltiwanger, John, Ron Jarmin and C.J. Krizan (forthcoming). "Mom and Pop Meet Big-Box: Complements or Substitutes?" *Journal of Urban Economics*.
- Isserman and Westervelt (2006). "1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data," *International Regional Science Review* 29:3.
- Kinney, Saki, Jerry Reiter, Arnold Reznick, Javier Miranda, Ron Jarmin and John Abowd (2009). "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database," mimeo.
- Jarmin, Ron S., and Javier Miranda (2002). "The Longitudinal Business Database," CES Working Paper 02-17.
- Lucero, Jason, Lisa Singh, and Laura Zayatz (2009). "The Current State of the Microdata Analysis System at the Census Bureau," *Journal of Statistical Methods*.
- National Research Council (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Ruggles, Steven and Russell R. Menard (1995). "The Minnesota Historical Census Projects," *Historical Methods*, 28:1, 6-10. <http://usa.ipums.org/usa/voliii/pumshist.shtml>
- U.S. Census Bureau (2001). *A Monograph on Confidentiality and Privacy in the U.S. Census*, July, <http://www.census.gov/history/pdf/ConfidentialityMonograph.pdf>
- U.S. Census Bureau (2002). *Measuring America: The Decennial Censuses From 1790 to 2002*, U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2003a). *Events in the Chronological Development of Privacy and Confidentiality at the U.S. Census Bureau*, report by History Staff, <http://www.census.gov/history/pdf/PrivConfidChrono.pdf>

U.S. Census Bureau (2003b). *2000 Census of Population and Housing, Public Use Microdata Sample, United States: Technical Documentation*.

U.S. Census Bureau (2008a). *2007 Research Report: Center for Economic Studies and Research Data Centers*, U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2008b). “Serious 2010 Census Challenges to Be Swiftly Addressed Clear Internal and External Agreement That 2010 Census Can Succeed With Major Operational and Budgetary Changes,” April 3. http://www.census.gov/Press-Release/www/releases/archives/2010_census/011773.html.

Wright, Carroll D. and William C. Hunt (1900). *History and Growth of the United States Census: 1790-1890*, Government Printing Office, Washington, DC.