ACCESS METHODS FOR UNITED STATES MICRODATA

Daniel H. Weinberg
U.S. Census Bureau

John M. Abowd
Cornell University and U.S. Census Bureau

Sandra K. Rowland
U.S. Census Bureau (retired)

Philip M. Steel
U.S. Census Bureau

Laura Zayatz
U.S. Census Bureau

contact information: <daniel.h.weinberg@census.gov>

6 August 2007

ABSTRACT
Beyond the traditional methods of tabulations and public-use microdata samples, statistical agencies have developed four key alternatives for providing non-government researchers with access to confidential microdata to improve statistical modeling. The first, *licensing*, allows qualified researchers access to confidential microdata at their own facilities, provided certain security requirements are met. The second, *statistical data enclaves*, offer qualified researchers restricted access to confidential economic and demographic data at specific agency-controlled locations. Third, statistical agencies can offer *remote access*, through a computer interface, to the confidential data under automated or manual controls. Fourth, *synthetic data* developed from the original data but retaining the correlations in the original data have the potential for allowing a wide range of analyses.

NOTE: This paper was prepared for the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit) Workshop on Data Access to Micro-Data, Nuremberg. Germany, August 20-21, 2007. It includes the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. The authors wish to thank Arnold Reznek and Tommy Wright for their helpful comments and suggestions.

Access Methods for United States Microdata

I. Introduction

Survey data on individuals and institutions are collected by many organizations. Government agencies, such as the U.S. Census Bureau, collect such data under strict statutory guidelines that require confidentiality -- the protection of a respondent's identity from public disclosure. The necessity for protection leads the agency in the direction of less openness. Yet the public good, and the reason an agency is funded, push the agency in the other direction: release as much information as possible.

This duality must be resolved within the bounds of the agency's enabling legislation. The U.S. Census Bureau uses five of the six basic approaches to providing information to the public. First is the provision of tabulations from the data that are collected, sometimes accompanied by explanatory text. Second is the provision of "public use microdata samples" (PUMS) from which independent researchers can produce their own analyses. The Census Bureau pioneered the use of PUMS in the 1960s and now produces a wide variety of such data files for household surveys. Among the techniques used to protect a respondent's identity on the PUMS are variable suppression, top- and bottom-coding, re-categorization, noise infusion, swapping, and geographic aggregation. Yet the computer revolution of the past half-century, especially the Internet, has made it increasingly possible to decode the information on such files, and agencies have responded by reducing the amount of information on such files. Furthermore, microdata files from business surveys have only rarely been made public, so independent research using representative business data has been difficult or impossible.

The four other methods for disseminating the results of surveys are the subject of this paper. Section II discusses licensing – whereby the statistical agency provides restricted data directly to individuals or organizations under a confidentiality protection agreement. Section III discusses research data centers – statistical enclaves where "outsiders" can enter the inner sanctum of a statistical agencies for research purposes. Section IV discusses remote access, wherein researchers can submit analysis requests (typically computer programs) to agencies and receive the results of those analyses. Section V discusses the most recent development – the creation of synthetic data that mirrors the properties of the collected data yet fully protects the confidential data provided by respondents.

This paper focuses on U.S. practice. Eurostat's Centre of Excellence for Statistical Disclosure Control (CENEX) *Statistical Disclosure Control Manual* (Hundepool et al. 2007) contains a general discussion of European approaches to statistical disclosure control, research data centers, remote execution, remote access, and licensing, with a specific reference to German official statistics.

II. Licensing

If public use microdata samples cannot provide sufficient information to researchers, agencies will sometimes "license" organizations to analyze "restricted-use" confidential microdata. The U.S. National Center for Education Statistics (NCES) uses this method for a large number of its confidential datasets.[1] As noted on their website, "The goal is to maximize

1. The NCES surveys for which restricted-data licenses are available are: Baccalaureate and Beyond, Beginning Postsecondary Students, The Common Core of Data, Early Childhood Longitudinal Studies, Education Longitudinal Study of 2002, High School and Beyond, High School Transcript Studies, National Assessment of Educational Progress, National Education Longitudinal Survey of 1988, National Household Education Survey, National

the use of statistical information, while protecting individually identifiable information from disclosure."[2] Typically, the license document

- defines the information subject to the agreement;
- specifies the individuals who may have access to subject data;
- describes limitations of disclosure and clearance procedures;
- lists administrative requirements;
- requires that copies of publications based on the data be sent to the sponsoring agency;
- requires the organization to contact the sponsoring agency in case of (suspected) breaches of security;
- requires the organization to agree to unannounced and unscheduled inspections; and
- reviews the security requirements for the maintenance of, and access to, subject data, and describes penalties for violations.

The U.S. Bureau of Labor Statistics (BLS) has established a similar program for access to its National Longitudinal Surveys (NLS) of Youth. As its web site notes,

> BLS has established a licensing system through which legitimate researchers at universities and other research organizations in the United States can use NLS data with geographic information at their own facilities, provided that the research project and physical and electronic security measures described in the NLS geocode application are approved by BLS. ... To protect the confidentiality of respondents, the BLS only grants access to geocode files for researchers in the United States who agree in writing to adhere to the BLS confidentiality policy and whose projects further the mission of BLS and the NLS program to conduct sound, legitimate research in the social sciences. ... Applicants must provide a clear statement of their research methodology and objectives and explain how the geocode data are necessary to meet those objectives. Researchers who are

---

Postsecondary Student Aid Study, National Study of Postsecondary Faculty, Postsecondary Education Quick Information System, Public Libraries Survey, School Library Media Centers, School Survey on Crime and Safety, Schools and Staffing Survey, Trends in International Mathematics and Science Study.

2. See the website <http://nces.ed.gov/pubsearch/licenses.asp> and its data use manual.

> granted access to NLS geocode files may use them at their own facilities, provided that the facilities meet BLS security requirements.[3]

Other BLS confidential datasets can be accessed only at BLS headquarters, once an application is approved. The older NLS cohorts can only be accessed through Census Bureau Research Data Centers (see section III), as the Census Bureau does not use licensing because of the way its authorizing statute is written. The U.S. National Science Foundation Division of Science Resource Statistics also uses restricted data licenses to allow researchers to access some of its confidential data.[4]

Licensing is also used in non-government settings. The University of Michigan Institute for Survey Research (ISR) licenses the use of a confidential (geocoded) version of its Panel Study of Income Dynamics (PSID).[5] As its web site notes,

> Due to our desire and obligation to protect respondent anonymity to the fullest extent allowable by law, the Geocode files are not available in general public release at the PSID Website or through the [Inter-university Consortium for Political and Social Research]. Rather, special contractual arrangements must be made to ensure that analysts maintain respondent anonymity. ... The process is somewhat lengthy and typically takes a couple of months. The timeframe is dependent on contract language issues and the responsiveness of the requesting institution. The analyst must submit a CV, a research plan, a sensitive data protection plan, a human subjects review clearance/waiver, and a completed signed contract. In addition, there is a non-refundable administrative fee due at the time the contract is submitted.

Other surveys use this method, such as the Fragile Families and Child Wellbeing Study conducted by Princeton University, which releases geographic identifiers to the public via a

---

3. See <http://www.bls.gov/nls/nlsfaqs.htm#anch25>.

4. These data include the Survey of Earned Doctorates, the Survey of Doctoral Recipients, the National Survey of Recent College Graduates, and the Scientists and Engineers Statistical Data System Integrated Data File. See <http://www.nsf.gov/statistics/license/start.cfm>.

5. The PSID is collected by ISR, but is publicly funded, by the U.S. National Science Foundation and other agencies. See <http://psidonline.isr.umich.edu/>.

restricted use data agreement.[6] Another example is the National Data Archive on Child Abuse and Neglect at Cornell University and its Longitudinal Studies of Child Abuse and Neglect.[7]

III. Research Data Centers

The statutory provisions under which U.S. statistical agencies collect data prevent the release of the full detail of survey data (e.g., names, addresses) in order to protect the confidentiality of respondents. As administrative data about individuals becomes more and more available through the Internet, statistical agencies must reduce the detail about individuals available through public use microdata. The availability of such data through the research enclaves can help ensure that valuable research can continue. Further, since business microdata has only rarely been in the public domain, the enclaves allow microeconomic research on businesses that could not otherwise take place.[8]

The Census Bureau now has nine data enclaves around the U.S., termed Census Research Data Centers (RDCs).[9] The RDCs are partnerships with academic and non-profit organizations. They are Census Bureau facilities managed by the Census Bureau's Center for Economic Studies (CES), staffed by a Census Bureau employee, and meet all physical and computer security

---

6. See <http://www.fragilefamilies.princeton.edu/restricted.asp>.

7. See <http://www.ndacan.cornell.edu/NDACAN/Datasets/Agreements_Licenses/LONGSCAN_Data_License.pdf>.

8. One example of public business microdata is the National Employer Survey for 1994 and 1997, available at http://www.irhe.upenn.edu/Library.php.

9. The CES web site contains additional information about the RDC program: <http://www.ces.census.gov/ces.php/rdc#objectives>. The nine locations currently open are Ann Arbor, Michigan; Berkeley and Los Angeles, California; Boston, Massachusetts; Chicago, Illinois; Ithaca and New York City, New York; Research Triangle, North Carolina, and Census Bureau headquarters in Suitland, Maryland (outside Washington, DC).

requirements for restricted access. RDCs offer qualified researchers restricted access to confidential economic and demographic data collected by the Census Bureau and other federal agencies in their surveys and censuses.

RDCs are aimed at researchers in academia; in independent research organizations such as the Urban Institute or the National Bureau of Economic Research; and in federal, state, and local government agencies. Tabulations of confidential data are generally not allowed to be removed from the RDCs, and therefore estimation of statistical models is the focus of their activities.[10] All researchers are required to become Special Sworn Status employees of the Census Bureau, and as such are subject to the penalty provisions of its authorizing legislation, should there be a confidentiality violation (e.g., a fine of up to US$250,000 and/or up to 5 years in prison).

The objective of the Census Bureau RDCs is to increase the utility and quality of Census Bureau data products. Access to microdata encourages knowledgeable researchers to become familiar with an agency's data products and data collection methods. More importantly, providing qualified researchers access to confidential microdata enables research projects that would not be possible without access to respondent-level information. This increases the value of data that has already been collected. Access to the microdata also allows for data linking not possible with aggregates – both cross-survey linkages and longitudinal linkages.[11] These linkages

10. Special tabulations can be arranged on a reimbursable basis.

11. All of the actual processing of data for approved proposals is conducted on servers located in the Census Bureau's secure central computer facility. Researchers located in the RDCs use thin clients (terminals) to access these servers via Virtual Private Networks.

leverage the value of existing data.[12] Creative use of microdata can address important policy questions without the need for additional data collection.

In addition, 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 microdata records available in a controlled, secure environment to sophisticated users who, by employing the microdata records in the course of rigorous analysis, will uncover the strengths and weaknesses of those records. Each set of observations is the end result of many decision rules covering definitions, classifications, coding procedures, processing rules, editing rules, disclosure rules, and so forth. The validity and consequences of all these decision rules only become evident when the Census Bureau's micro databases are tested in the course of analysis. Exposing the conceptual and processing assumptions that are embedded in the Census Bureau's microdata databases to the light of research constitutes a core element in the Census Bureau's commitment to quality.

The opportunities for researchers to carry out unique research come at a price. Research conducted at RDCs takes place under a set of rules and limitations that are considerably more constraining than those prevailing in typical research environments. Research proposals are reviewed on the basis of five major standards:

1. **Benefit to Census Bureau programs.** Proposals must demonstrate that the research is likely to provide one or more benefits to the Census Bureau.[13] These benefits can include:

12. Researchers may also bring data into the RDCs and arrange for linkage to Census Bureau datasets.

13. Title 13, United States Code, the Census Bureau's authorizing legislation, permits the Census Bureau to employ Special Sworn Status employees for the purpose of carrying out its mission. Specifically, Section 23(c) states "The Secretary [of Commerce] may utilize temporary staff, including employees of Federal, State, or local agencies or instrumentalities, and employees of private organizations to assist the Bureau in performing the work authorized by this title, but only if such temporary staff is sworn to observe the limitations imposed by section 9 of this title."

a. Evaluating concepts and practices underlying Census Bureau statistical data collection and dissemination practices, including consideration of continued relevance and appropriateness of past Census Bureau procedures to changing economic and social circumstances;

b. Analyzing demographic and social or economic processes that affect Census Bureau programs, and that evaluate improvements to the quality of products issued by the Census Bureau;

c. Evaluating or analyzing public programs, public policy, and/or demographic, economic, or social conditions to identify potential complementary datasets, improve data quality, enhance data collection techniques or develop innovative estimation procedures;

d. Conducting or facilitating census and survey data collection, processing or dissemination, including through activities such as administrative support, information technology support, program oversight, or auditing under appropriate legal authority;

e. Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census, or estimate;

f. Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census, or estimate;

g. Enhancing the data collected in a Title 13, Chapter 5 survey or census. For example: (1) improving imputations for non-response; (2) Developing links across time or entities for data gathered in censuses and surveys authorized by

Title 13, Chapter 5;

h. Identifying the limitations of, or improving, the underlying Business Register, Master Address File, and industrial and geographical classification schemes used to collect the data;

i. Identifying shortcomings of current data, collection programs and/or documenting new data collection needs;

j. Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5;

k. Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;

l. Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5;

m. Developing statistical weights for a survey authorized under Title 13, Chapter 5.

2. **Scientific merit.** This criterion relates to the project's likelihood of contributing to existing knowledge. Evidence that a Federal agency such as the National Science Foundation or the National Institutes of Health has approved the proposal for support usually constitutes sufficient indication of scientific merit.

3. **Clear need for non-public data.** The proposal should demonstrate the need for and importance of non-public data. The proposal should explain why publicly available data sources are not sufficient to meet the proposal's objectives.

4. **Feasibility.** The proposal must show that the research can be conducted successfully with the methodology and requested data.

5. **Risk of disclosure.** Output from all research projects must undergo and pass disclosure review.

   a. Tabular and graphical output presents a higher risk to disclosure of confidential information than do coefficients from statistical models.

   b. The Census Bureau is required by law to protect the confidentiality of data collected under its authorizing legislation.

   c. Some data files are collected under the sponsorship of other agencies. In providing restricted access to these data, CES must adhere to all applicable laws and regulations.

   d. Researchers may be required to sign non-disclosure documents of survey sponsors or other agencies that provide data for their research projects.

Both Census Bureau and external experts on subject matter, datasets, and disclosure risk review all proposals to use Census Bureau data. Relevant data sponsors and data custodians also review proposals that request certain datasets. Any proposals approved by Census Bureau staff seeking to use datasets that contain federal tax information must also be reviewed for approval by the Internal Revenue Service (IRS); if data are collected for another federal agency, approval must be obtained from that agency as well. The Census Bureau's legal authority to provide access to IRS tax data in its custody requires that the Title 13, Chapter 5 benefit be the *predominant purpose*; criteria (e)-(m) listed above as a benefit can be used to justify access. Proposals must also pass a review by the Census Bureau's policy office to ensure consistency with agency mission. If a proposal is not approved, it can be resubmitted if revised to address noted deficiencies.

The proposal review process can be cumbersome and time consuming, and the consequent delays in getting access to the data at the RDCs are frustrating to researchers. Average and median review times are falling so this obstacle has become lower. Also, all projects must, by law, have a benefit to the Census Bureau. Therefore, some worthy research projects with questionable benefits must be rejected.

The output of RDC projects can be methodological or statistical and includes both scientific papers and benefit statements addressing the Census Bureau's needs. Output undergoes disclosure review under rules established by the Census Bureau's Disclosure Review Board, which may review particularly difficult situations.

While the Census Bureau contributes approximately 55 percent of the full costs of the RDC network, the remaining costs must be recovered from sources outside the Census Bureau. The university and non-profit organizations which operate the non-headquarters RDCs typically contribute the space in which the RDCs operate, and provide "release time" to the professor or individual who serves as the RDC's Executive Director. But they must also pay the salary of the RDC Administrator, raising those funds in a variety of ways – as a direct contribution of the partner institution, through membership fees from a funding consortium, by charging fees for access, or a combination of these methods.[14]

One recent development that will increase the utility of the RDC network to researchers is the decision to allow the confidential data of other federal agencies to be available through the

14. A typical fee for academic access is US$15,000 per year, plus any other costs (cleaning the file, creating the metadata, linkage between files). Some universities provide infrastructure funds in exchange for free access by their faculty and graduate students.

RDCs.[15] So far, the U.S. National Center for Health Statistics (NCHS) and the Agency for Healthcare Research and Quality have reached an agreement with the Census Bureau to make its confidential data available in that way.[16]

The National Institute of Child Health and Human Development has established a research data center network that allows access to qualified researchers to data collected by its grantees who have collected demographic data. Their Data Sharing for Demographic Research (DSDR) project operates at three locations – the Carolina Population Center at the University of North Carolina at Chapel Hill, the Minnesota Population Center at the University of Minnesota, and the Population Studies Center at the University of Michigan.[17] DSDR also provides limited data through licensing.

The National Opinion Research Corporation (NORC, a not-for-profit U.S. enterprise) and the U.S. National Institute of Standards and Technology (NIST) Advanced Technology Program (ATP) have just established a data enclave, in order to provide restricted access to U.S. business microdata collected by ATP and others using licensing through remote access or onsite use.[18] While researchers supported by private and public foundations (e.g., the Kauffman Foundation) and other research-supporting institutions (e.g., the National Science Foundation and the

---

15. A Committee on National Statistics panel has recently issued a report, *Expanding Access to Research Data: Reconciling Risks and Opportunities*, which contains praise for the RDC program, and recommended that "the Census Bureau and other statistical agencies should explore ways to house confidential data from as many agencies as possible in a single supervised location in a number of host institutions in order to add to their value for research use." (Committee on National Statistics, 2005; page 77)

16. NCHS has its own RDC at its headquarters, but no other locations.

17. Their website is <http://www.icpsr.umich.edu/DSDR/>.

18. Their website, still under development, is <http://dataenclave.norc.org/>. Onsite access is at either of NORC's two offices (Chicago and Washington DC).

National Institutes of Health) and some U.S. federal agencies could place their confidential data at this enclave, because of statutory restrictions (such as the law authorizing the Census Bureau to collect confidential information), not all such data could be placed at that enclave. The NORC enclave can be a viable access method for data not constrained by statute.[19]

Other countries have adopted the RDC approach. By far the most advanced (in some ways surpassing the U.S. approach on which it was based) is the Canadian RDC network.[20] This is a true network in which the leadership is by a coordinating council of partner institutions, and the central statistical office, Statistics Canada, plays a facilitating rather than a lead role (and hosts a "federal" data center), with primary funding coming from the partner institutions and granting agencies. The United Kingdom has also established a Virtual Microdata Laboratory, where academics and government officials can access confidential firm-level (business), controlled-access census, and potentially other microdata files under special license (Ritchie, 2006). The relatively new Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) – a pilot program from 2004 to 2006 – has stated its goal as "to facilitate access to BA and IAB micro data for non-commercial empirical research using standardized and transparent access rules "[21]

IV. Remote Access to Microdata

19. Its founding sponsor, NIST, is not a statistical agency and therefore its data collection activities are not covered under the Comprehensive Information Protection and Statistical Efficiency Act of 2002.

20. Full information about the Canadian network can be found at <http://www.statcan.ca/english/rdc/index.htm>. Canadian researchers do not have access to microdata on businesses.

21. Information about the German FDZ can be found at <http://fdz.iab.de/en/pageTextModulRight.asp?PageID=1>; see also Kohlmann (2005).

Remote access systems make it possible for users to analyze restricted microdata without visiting a statistical enclave. The remote access systems provided by statistical agencies employ automated and manual filters that block certain kinds of queries and results and must be monitored automatically and/or manually for disclosure avoidance; extracts of microdata and direct access to the records are not permitted. The data files available for analysis are usually edited in advance to reduce the possibility of disclosure using the same techniques as those used for public use files. However, they tend to provide more detail to researchers to carry out their analyses than do public use files, but less detail than is usually available in a RDC.

The Luxembourg Income Study (LIS) is the oldest of the data suppliers that give users remote access to restricted microdata.[22] LIS began in 1983 to harmonize income data on household surveys from a number of countries. Its data managers developed the LISSY remote access system to allow users from around the world to analyze the household surveys included in its database. That system has (consciously or unconsciously) served as a model for many other systems currently in use and under development. Canada and Denmark have given users remote access to restricted microdata since 2001. The Netherlands, Sweden, and Australia began pilots in the use of remote access systems in 2002 and 2003.[23] In the U.S., the National Center for Education Statistics (NCES) and the National Center for Health Statistics (NCHS) gave users remote access to restricted microdata beginning in 1997 and 1998, respectively. The Census

22.The LIS is a non-profit cooperative research project with a membership that includes 30 countries on four continents (Europe, America, Asia, and Oceania); more information can be found at <http://www.lisproject.org/>. Hundepool et al. (2007) refer to the LIS system as “remote execution”.

23. For example, the Australian Remote Access Data Laboratory can be accessed at <http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL)?OpenDocument>.

Bureau began disseminating Census 2000 microdata tabulations via remote access in 2003, after pilot tests in 2002.[24]

There are two common methodologies in use among these systems. One type usually consists of an email interface that allows users to send programs as part of the body of the email or in an attachment. These systems usually accept standard statistical programs such as SAS, SPSS, STATA, and GAUSS, chosen because they are commonly used by researchers and lend themselves to automated review of input programs and statistical results. These "remote job execution systems" systems typically prohibit or modify certain commands, thus limiting the kinds of outputs that users may receive; this type is used in all of the non-U.S. applications and by NCHS. Processing is usually done in batch mode ("off-line") rather than interactively ("on-line"). Results are returned within minutes or days depending on the size of the program and the degree of manual intervention needed to assure the confidentiality of the data is maintained.

The other, less common, type of system consists of a web interface with custom-built or custom-tailored (commercial) software that requires users to learn how to use the program and/or user interface. The web-based systems produce tabular or model results with percentages and/or means and may also provide variances and correlation matrices. The web applications are used in the NCES and Census Bureau systems. Processing is done while the user is on-line and results are returned within seconds or minutes depending on the size of the tabulation. There is no manual intervention.

The Census Bureau's Microdata Analysis System (MAS) is currently in prototype. It allows users to select a study population, collects information from the users on what model the

24. Rowland and Zayatz (2001) discuss the design of this system.

users would like to run and what options he or she would like to invoke (e.g., weighted or unweighted), passes this information through a firewall, constructs the population and the model, tests the request for confidentiality (mostly for the size of population), runs the model, tests the results, and returns them to the user. The system uses a synthetic data procedure to produce residuals and supporting population-specific tables. It also has a mechanism to guard against complementary disclosure.

In addition to statistical protections, such systems require software and security support. The information being passed through the firewall must be verifiable without additions (e.g., a system exit) and the descriptive elements must be genuine. Software run on government systems, particularly those with external interfaces, are subject to a variety of regulations – including a detailed and extensive security plan. Platform dependencies must be held to a minimum, several software applications brought together, a large catalog of metadata constructed and fixed, and a detailed user interface maintained. See Steel (2006) for more information.

Several aspects of the methodology employed vary by system (see the Appendix):

- The use of confidentiality edits to the base files (these usually involve adding noise to the data to reduce the possibility of disclosure).
- Electronic authorization of users, typically requiring the use of user identification and passwords to gain access to the system.
- Email or web user interfaces – the way users communicate what they want from the system.
- The use of standard statistical programs or custom applications for processing.
- The presence of query filters to examine requests and block the user from requesting

certain results prohibited by the statistical agency.

- The presence of results filters to examine results and block any result prohibited by the statistical agency.
- Automated and manual intervention by statisticians for disclosure avoidance at the query submission or output stages.
- Usage logs for disclosure avoidance review that are accumulated and used by statistical agencies to determine if their rules are adequate for disclosure avoidance and to detect possible risks to confidentiality.

Another important methodological aspect of a remote system is automating complementary disclosure review to prevent the possible disclosure of restricted data that could result from combining multiple outputs. Although research in this area has been undertaken (see Duncan et al. 2000), no comprehensive mechanism is known to prevent complementary disclosure. This may be due to the difficulty and expense of automating such procedures. One possibility is being tested for the MAS.

Researchers and statistical agencies are also interested in disseminating data relating to smaller geographic areas while preventing disclosure of restricted data. Karr and Sanil (2001) have done research on automating aggregation of small geographic areas and/or allowing user-defined geographic areas but none of the current systems contain mechanisms to disseminate data for user-defined areas.

Most of the users of monitored remote access systems are researchers and public sector employees. Most statistical offices require user registration and some must officially accept a research proposal before the system can be accessed. There are several aspects of system usage

also covered in the Appendix:

- Whether permission or authorization is required by the statistical agency to access the systems (this may involve signing a research contract and confidentiality agreement or some other kind of registration).

- The types of users permitted to use the systems – ranging from the most exclusive policy of allowing only public sector users, to the most inclusive policy of allowing anyone to use the systems.

- The files available through the systems – ranging from one or two files to many, as well as whether files can be combined and whether user files are permitted.

- The presence of documents or metadata online, varying from detailed user guides to tailored emails.

- Whether assistance is available, including automatic feedback and help desks and user workshops.

- Turnaround time for results ranging from seconds to days.

- Hours of availability – ranging from 24 hours a day, 7 days a week to office hours only (such as 8 hours a day, 5 days a week).

- Cost – ranging from zero to periodic membership fees to fee for service time.

- Examples of benefits derived from use including reports and policy implications.

V. Synthetic Microdata

Creating synthetic microdata is a disclosure avoidance technology that protects confidentiality by replacing actual microdata with data that have been simulated either fully or partially. Rubin (1993) and Little (1993) proposed the technique of synthesizing microdata to reproduce the statistical properties of the underlying confidential data while replacing all sensitive times with simulated values. Although apparently very new, synthetic data actually used for the first time by the U.S. Census Bureau in the preparation of the 1990 Decennial Census Summary Files, where it was called “blank and impute” (see Federal Committee on

Statistical Methodology 2005, p. 32). In addition, the Survey of Consumer Finances has used methods that would now be called “synthetic” for confidentiality protection since editing the 1989 survey (see Kennickell 1991 and 1997, and the discussion in the appendix to Abowd and Woodcock 2001).

Rubin’s technique, now known as “fully synthetic data”, considered all unsampled units in a sampling frame as “missing”. The responses gathered from the survey to be protected are multiply-imputed using the variables in the sampling frame, assumed to be known for all entities in the population, as the conditioning variables. The result is a “completed” data set in which every entity in the population has values for every variable in the survey. Synthetic microdata files are created by sampling from each of the synthetic populations. Rubin showed that a variation on the analysis technique given in his original multiple imputation formulas for missing data (Rubin 1987) could be used to compute estimated values and confidence intervals from the synthetic data. Rubin’s original method relied on a Bayesian bootstrap to simulate the missing data.

Fienberg (1994) elaborated on Rubin’s technology by proposing the use of the posterior predictive distribution from a model-based data analysis. Whereas the Bayesian bootstrap preserved all multivariate relationships among the synthetic variables by reusing the actual survey responses (imputed to other entities in the frame, via the conditioning variables), model-based methods produced synthetic microdata that were not any entities’ actual responses to the underlying confidential survey.

Little’s technique, now known as “partially synthetic” data, considered only the sensitive values or sensitive cases in the data set to be protected. He proposed synthesizing either the

sensitive values or the identifiers of sensitive cases, using the other variables in the original sample as conditioning information and properly accounting for the rule that designated the value as sensitive. Like Rubin, Little provided formulas that could be used for computing estimated values and confidence intervals.

In spite of the fact that the techniques were not really a dramatic departure from current practice in the early 1990s, they did not receive much attention from the disclosure limitation community. The technique was reviewed in the original version of Federal Committee on Statistical Methodology Working Paper 22 (1994) without comment. Although data users were prepared to accept the technique, as can be seen from its use by the Survey of Consumer Finances, they were given few opportunities (Abowd and Lane 2004).

There are essentially two standards that synthetic microdata must meet. The first is a disclosure avoidance standard. The synthetic microdata must be shown to protect confidentiality at least as well as other methods, and to ensure that unlawful disclosures are avoided. Second, the synthetic microdata must be shown to provide inferences that are consistent with the inferences an analyst would have made from the original data, although possibly less precise because of the incremental uncertainty associated with the synthesizing process. This second standard is known as analytical validity.

The use of synthetic data techniques began to expand with the advent of large-scale microdata based on longitudinally integrated employer-employee data, like those produced by the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program; see Abowd et al. (2004). These data are particularly difficult to protect using standard microdata disclosure avoidance techniques because the critical new information in the data consists of the

linkages between the various entities that have been integrated. In their 2001 and 2004 papers, Abowd and Woodcock reviewed the disclosure limitation methods that might be applied to such data and concluded that synthetic data provided a viable technique that could simultaneously provide disclosure avoidance and analytical validity if properly applied. They provided examples based on synthesizing confidential integrated employer-employee data from France. The method that they found most effective would now be called “partially synthetic data with missing values”.

Two related research strains honed the toolkit. The first strain involves the specific technique that Abowd and Woodcock used, which is called “Sequential Regression Multivariate Imputation” (SRMI; Raghunathan et al. 1998), and was taken from the missing data literature. SRMI approximates complicated multivariate distributions that mix discrete and continuous variables by a sequence of conditional distributions in a manner similar to Markov Chain Monte Carlo, but is much simpler to implement. Raghunathan et al.(2003) proposed essentially the same technique, and also provided examples of its successful application. The second strain involves appropriately specifying all the inference formulas when synthetic and missing data techniques are applied to the same underlying source data. Reiter (2004) provided the complete set of formulas for combining partially synthetic missing data using multiple imputation.

In 2001 in the U.S., a consortium of the Census Bureau, Social Security Administration (SSA), Internal Revenue Service (IRS), Congressional Budget Office, and Congressional Joint Committee on Taxation undertook a research project to create a public use file from the Survey of Income and Program Participation linked to SSA-supplied longitudinal benefit histories and IRS-supplied longitudinal employee-employer earnings reports (“W-2” data). After investigating

several feasible disclosure avoidance methods, the group agreed to experiment with partially synthetic data. The initial SIPP synthetic data file (the "Beta" version) is now available for testing by researchers (Abowd et al. 2006). The testing protocol grants any person access to the synthetic version of the data provided that the variables required to perform the analysis are on the file. Analyses are conducted on an Internet-accessible computer system (<http://vrdc.ciser.cornell.edu/news>). At the time this manuscript was prepared, the disclosure review boards from the Census Bureau, SSA, and IRS had all approved the protocol; however, the data files themselves had not been placed on the access system.

In 2005, the Census Bureau's LEHD program released its first official synthetic microdata product a public use application – *OnTheMap*, based on its longitudinally integrated employer-employee data (<http://lehdmap2.dsd.census.gov/>). That application relates residence and workplace addresses, allowing the user to map the residence locations of all individuals working at a given location or all the work locations for individuals living at a given location.[25]

As part of a 2004 National Science Foundation grant to the Census Research Data Center partners, an active collaboration between the Census Bureau and university-based researchers from Carnegie-Mellon, Cornell, Duke, Michigan, and Simon Fraser is producing synthetic microdata from a variety of sources.[26] These include the Longitudinal Business Database (LBD), American Community Survey (ACS), and the LEHD Infrastructure File System. Synthetic microdata from the LBD are already available on the Virtual Research Data Center. Synthetic microdata from the ACS are scheduled for release as part of the 2006 Public Use Microdata

25. The synthetic microdata themselves can be downloaded from the Cornell-based Virtual Research Data Center (http://vrdc.ciser.cornell.edu/onthemap/doc/).

26. Grant number 0427889; details at <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0427889>.

Sample for the ACS. The techniques are being used to protect the confidentiality of ACS responses for group quarters residents. Synthetic microdata from the LEHD infrastructure are expected to be released in the fall of 2007.

VI. Concluding Comment

Microdata is the foundation on which our understanding of human and business behavior is based. Yet threats to confidentiality are increasing as computers get more powerful and more and more information is available to the public on the Internet. In the face of such threats, preserving the extent of current microdata becomes less and less likely. Consequently, statistical agencies need to use new and different methods to make their information available to the public, in ways that preserve the ability of social scientists to manipulate the data for research purposes while preserving the confidentiality of respondents. This paper has discussed four such ways that statistical agencies have responded: licensing, research enclaves, remote access, and synthetic data.

References

Abowd, J.M., M. Stinson and G. Benedetto. 2006. *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project* (available at [URL pending Census approval, should be available at the time this paper is presented in August 2007]).

Abowd, J.M. and S. Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data," in P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Amsterdam: North Holland), 215-277.

Abowd, J.M. and S. Woodcock. 2004. "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag), pp. 290-297.

Abowd, J.M. and J. Lane. 2004. "New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers," in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases* (Berlin: Springer-Verlag), pp. 282-289.

Abowd, J.M., J. Haltiwanger and J. Lane. 2004. "Integrated Longitudinal Employee-Employer Data for the United States," *American Economic Review Papers and Proceedings*, Vol. 94, No. 2 (May), pp. 224-229.

Committee on National Statistics, Panel on Data Access for Research Purposes. 2005. *Expanding Access to Research Data: Reconciling Risks and Opportunities*. National Academy Press.

Duncan, G., S. Roehrig, and K. Kannan. 2000**.** Final Report on the American FactFinder Disclosure Audit Project for the U.S. Census Bureau.

Federal Committee on Statistical Methodology.1994. *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22 . <http://www.fcsm.gov/working-papers/wp22.html>.

Federal Committee on Statistical Methodology. 2005. *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22 (Revised). <http://www.fcsm.gov/working-papers/spwp22.html>

Fienberg, S.E. 1994. "A Radical Proposal for the Provision of Microdata Samples and the Preservation of Confidentiality." Carnegie Mellon University Department of Statistics *Technical Report* No. 611.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P-P. De Wolf. 2007. *Handbook on Statistical Disclosure Control*, version

1.01. A [Eurostat] Centre of Excellence for Statistical Disclosure Control. March. <http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf>

Karr, A. F. and A. P. Sanil. 2001. "Web Systems that Disseminate Information but Protect Confidential Data." National Institute of Statistical Sciences Working Paper.

Kennickell, A.B. 1991. "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," SCR Working Paper, prepared for the Annual Meeting of the American Statistical Association, Atlanta, GA, August <http://www.amstat.org/Sections/Srms/Proceedings/papers/1991_001.pdf>.

Kennickell, A.B. 1997. "Multiple Imputation and Disclosure Limitation: The Case of the 1995 Survey of Consumer Finances" Chapter 8 in Federal Committee on Statistical Methodology (eds.) *Record Linkage Techniques - 1997*. Proceedings of an International Workshop and Exposition, Arlington, VA, March 20-21. <http://www.fcsm.gov/working-papers/akennickell.pdf>

Kohlmann, A. 2005. "The Research Data Centre of the Federal Employment Service in the Institute for Employment Research." *Schmollers Jahrbuck* 125, pp. 437-447. <http://www.ratswd.de/download/schmollers/Kohlmann.pdf>

Little, R.J. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9(2), pp. 407-426.

Raghunathan, T.E., J.M Lepkowski, J. Van Hoewyk, and P. Solenberger. 1998. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." Survey Research Center, University of Michigan.

Raghunathan, T. E., J. P. Reiter, and D. B Rubin. 2003. "Multiple imputation for statistical disclosure limitation" *Journal of Official Statistics* 19, pp. 1-16.

Reiter, J. P. 2004. "Simultaneous use of multiple imputation for missing data and disclosure limitation" *Survey Methodology* 30, pp. 235-242.

Ritchie, F. 2006. "Statistical Disclosure Detection and Control in a Research Environment." Paper for 2006 International Conference on Comparative Analysis of Enterprise (Micro) Data [CAED], Chicago, IL. <http://www.ces.census.gov/docs/caed/papers/pap_162_Felix_Ritchie.pdf>

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys* (New York: Wiley).

Rubin, D.B. 1993. "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics* 9(2), pp. 461-468.

Rowland, S. and L. Zayatz. 2001. “Automating Access with Confidentiality Protection: The American FactFinder.” *Proceedings of the Government Statistics Section, American Statistical Association*.

Steel, P. M. 2006. “Design and Development of the Census Bureau's Microdata Analysis System: Work in Progress on a Constrained Regression Server.” Presentation at Federal Committee on Statistical Methodology Statistical Policy Seminar. November 28-29.

Appendix 1A: Monitored Remote Access System Methodology – U.S. Federal Agencies

	National Center for Health Statistics (NCHS)	**National Center for Education Statistics (NCES)**	**Census Bureau**
System Name	Analytical Data Research by Email (ANDRE)	Data Analysis System (DAS)	Advanced Query (AQ)
Date Begun	April 1998	1997; revised 2003	April 2003
Electronic Authorization	Email access certification	None needed	User name and password Predefined domains (IP address)
Software for User Interface	SAS, SUDAAN	Custom-built application; accommodates various surveys	Commercial software tailored for use (SQL processing)
Output format	SAS commands	Tables with standard errors and correlation matrices	Tables
Email/Web	Email	Web	Web
Confidentiality edits to base file	Yes	Yes	Yes
Query filter	SAS log commands not permitted are modified and suppressed	Must be enough cases to allow calculations	Web interface restricts data available
Output filter	Cell and row suppression	Row suppression	Whole table suppressed
Intervention	Automatic; questionable output routed to staff for resolution	Automatic	Automatic
Complementary Disclosure Techniques	Yes	No	No
Usage Logs for Confidentiality Review	Yes	No	Yes
User Defined Areas	Information not available	No	No

Appendix 1B: Usage of Monitored Remote Access Systems – U.S. Federal Agencies

	NCHS – ANDRE	**NCES – DAS**	**Census Bureau – AQ**
Authorization Required	Registered subscribers. Proposals received and approved	None	Registered users
Principal Types of Users	Anyone may apply	Anyone can run tables. Principle users are policy analysts and researchers	State Data Centers; State Legislatures; Census Information Centers
Number of Users	45 users in past 5 years; 10,000 SAS programs run		Over 500
Files available	All NCHS files and a few others	8 surveys dealing primarily with post-secondary education	Census 2000 hundred percent and sample
User files accommodated	Yes	No	No
Most popular files	National Family Growth Survey, National Health and Nutrition Examination Survey	National Post Secondary Student Aid Study	Census 2000 Sample
Metadata On-Line	Yes	Yes	Yes
Assistance Available	Documents and personal assistance via email	Web help and personal assistance via email	User guide and automatic feedback on web
Turnaround	A few hours	Seconds to minutes	Seconds to minutes
Cost	$500 per month	Free	Free
Availability	24/7	24/7	24/7
Benefits derived	Cheaper and more convenient than RDC access	Can access tables required without a restricted-use license	User-defined tables from Census 2000
Types of analysis	Research in numerous health topics such as marriage, family formation, family health and fertility	Post-secondary Education Descriptive Analysis Reports require use of DAS	Public analysis and planning in all areas covered by Census 2000
User assessments	Information not available	1999 Customer Satisfaction Survey	Assessment of Census 2000 Sample data test

Appendix 2A: Monitored Remote Access System Methodology – Non-U.S. Agencies

	Australian Bureau of Statistics	**Statistics Canada**	**Statistics Denmark**
System Name	Remote Access Data Laboratory (RADL)	NA	NA
Date Begun	April 2003	2001	March 2001
Electronic Authorization	Yes	Yes	Rented password key
Software for User Interface	SAS, SPSS	SAS, SPSS, STATA (varies by survey)	SAS, SPSS, STATA, GAUSS
Output format	SAS, SPSS	Varies by survey	SAS, SPSS, STATA, GAUSS
Email/Web	Email	Email	Web user interface Email output retrieval
Confidentiality edits to base file	Yes	Yes	Yes
Query filter		No filters. Statistics Canada executes programs in batch runs and reviews outputs for disclosure avoidance.	No filters. Users can make new datasets but not download datasets or individual records. Statistics Denmark reviews outputs for disclosure avoidance.
Output filter	Largely automated confidentiality checks		
Intervention	Automatic triggers for disclosure avoidance; manual inspection of output	Manual inspection of output	Manual inspection of output
Complementary Disclosure Techniques	No	Techniques to prevent linkages	No
Usage Logs for Confidentiality Review	Audit trails and records	Yes	Emails logged and checked by Statistics Denmark
User Defined Areas	No	No	No

Appendix 2A: Monitored Remote Access System Methodology – Non-U.S. Agencies

	Luxembourg Income Study	**Statistics Netherlands**	**Statistics Sweden**
System name	Luxembourg Income Study System (LISSY)	OnSite@Home	Feasibility study
Date Begun	1983; email in 1987 on EARN/BITNET	2006	2002-2003
Electronic Authorization	User ID and Password	Biometric id, smart card and password	User name and password Predefined domains (IP address)
Software	SAS, SPSS, STATA	any	Information not available
Output format	Every type of analysis No extracts of microdata	Any statistical query except extracts	Information not available
Email/Web	Email	CITRIX	Email
Confidentiality edits to base file	Yes	Yes (restrictions weakened upon success of initial tests)	Yes
Query filter	Yes; certain commands, word sequences and variables not permitted	No	Similar to Statistics Denmark
Output filter	Yes	Frequencies and contingency tables will be automated in the future	Information not available
Intervention	Accepts programs, processes and returns but suspicious output reviewed	Manual review of output; will be automated in the future	Information not available
Complementary Disclosure Techniques	No	No	Information not available
Usage Logs for Confidentiality Review	Yes	Yes – Future	Information not available
User Defined Areas	No	No	Information not available

Appendix 2B: Usage of Monitored Remote Access Systems – non-U.S. Agencies

	Australia Bureau of Statistics	**Statistics Canada**	**Statistics Denmark**
Authorization Required	Yes	Contact required to create files for use	Granted on need-to-know basis for specific projects
Principal Types of Users	Information not available	Government and university researchers	Authorized institutional environments (ministries, research institutes, universities, non-governmental organizations); no private individuals or foreigners.
Number of Users	Information not available	Information not available	3/01-3/03 - 43 authorizations
Files available	Microdata files held at ABS	Small number of surveys	Same files available for on-site research. Usually sample data
Most popular files	Information not available	National Population Health Survey (NPHS) Survey of Labor and Income Dynamics (SLID)	Integrated Database for Labor Market Research
Metadata On-line	Information not available	Survey documentation Data dictionary Synthetic dummy file	Information not available
Assistance Available	Information not available	Workshops Macros for variance estimation	Information not available
Turnaround	Information not available	1-2 working days	Information not available
Cost	Information not available	Yes	Information not available
Availability	8/5	8/5	8/5
Benefits derived	Access to more detailed data	Less physical and administrative restrictions than RDC	Information not available
Types of Analysis	Information not available	Research in health and employment issues	Research in health, employment and business

Appendix 2B: Usage of Monitored Remote Access Systems – Foreign Agencies

	Luxembourg Income Study	**Statistics Netherlands**	**Statistics Sweden**
Authorization Required	Contract with confidentiality pledge	Yes	Yes
Principal Types of Users	Academic users	All government ministries	Public authorities; researchers; users of regional statistics
Number of Users	Jan 2001 to Jun 2003: 213 users, 36,280 programs (highest usage by U.S., U.K., Germany)	Pilot – University of Tilberg	Just began evaluation with actual users
Files Available	Income files from 25 countries	Social allowances	Information not available
Most popular files	Income Study	Information not available	Information not available
Metadata on-line	Yes (documentation of key variables; synthetic microdata files)	Yes	Information not available
Assistance Available	Yes (workshops, help desk)	Yes	Information not available
Turnaround	Minutes	½ working day	Information not available
Cost	Annual country fee	Information not available	Information not available
Availability	24/7	Information not available	Information not available
Benefits derived	Cross country analysis of similar data	Information not available	Information not available
Types of Analysis	Contributions to international inequality and poverty research	Information not available	Information not available