DATA LICENSING AGREEMENTS AT U.S. GOVERNMENT AGENCIES AND RESEARCH ORGANIZATIONS

Paul B. Massell and Laura Zayatz[1], U.S. Census Bureau
Paul B. Massell, Statistical Research Division, U.S. Census Bureau, Washington D.C. 20233
paul.b.massell@ccmail.census.gov

ABSTRACT

Federal statistical agencies and research organizations often collect data under a promise that the data will be kept confidential. At the same time, the agencies and organizations wish to release as much of this information as possible for research, while upholding the confidentiality pledge. When the risk of disclosing confidential information associated with a research data file is considered too high to allow public release of the file, there are several restricted access methods that allow a limited number of researchers to use the data (Jabine, 1993). One of these, data licensing, involves the signing of a formal contract between the data providing organization and a research team. Licensing allows researchers access to data that have undergone minimal disclosure limitation procedures and are therefore most useful for detailed analyses. One drawback to licensing is the need to establish a structure for its implementation. This involves the drawing up of a legally binding contract at the data providing organization, the implementation of methods for secure handling of the data, and an enforcement procedure. This paper describes some of the licensing procedures in effect at eight such organizations. Most of the files licensed by these organizations involve demographic data or data from noncommercial institutions (schools, hospitals). However, licensing is also a sensible way to release business establishment files.

Key words: confidentiality, disclosure limitation, restricted data, restricted access

1. INTRODUCTION

Federal statistical agencies and research organizations often collect data under a promise that the data will be kept confidential. At the same time, the agencies and organizations wish to release as much of this information as possible for research, while upholding the confidentiality pledge. Data rich microdata files allow for very detailed analyses by researchers. Unfortunately, the data richness often means a given record representing an individual or establishment reflects a unique combination of characteristics of that individual or establishment, and the likelihood of some reidentifications (disclosures of confidential information) is increased in comparison to less data rich files. One way of satisfying both concerns, the desire of researchers to have access to such files and the desire to prevent disclosures, is for the agency or research organization to release the file under highly controlled conditions. This may involve the development of a licensing arrangement under which the file is lent to a small group of researchers, typically at a university. The licensing agreement must be signed by the principal researcher for the project and, possibly, certain administrators at the university. Together these licensees have the responsibility of ensuring that (1) the data file is used only by the small group of designated file users and (2) the data are kept confidential as required by the licensing agreement as well as by applicable laws and promises made to survey respondents.

The main purpose of this paper is simply to describe the key features of the licensing agreements that have been developed and used for licensing data files in recent years by major U.S. government agencies and a few research organizations. Most of the files contain social science data but there has been licensing of biomedical data (e.g., National Health and Nutrition Examination Survey -- NHANES) and public health data (e.g., the National Longitudinal Survey of Adolescent Health).

In addition, we have presented some comments by data managers and others who are experienced with licensing. Their comments describe things they have learned from that experience that could benefit managers who are new at licensing or are considering it.

Our target population was federal statistical agencies that have licensed data. For variety, we also wanted to include some non-governmental research organizations. Most of the information was supplied by colleagues from the Interagency Confidentiality and Data Access Group (ICDAG), (recently renamed the Confidentiality and Data

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion.

Access Committee (CDAC)). Other information was extracted from web sites. (An earlier version of this paper, (Massell, 1999), contains some details omitted from this version). The following agencies and organizations were included in the study:

Social Science Research Organizations:
- Inter-university Consortium for Political and Social Research (ICPSR)
- Survey Research Center at the University of Michigan (SRC-UMICH)

U.S. Government Agencies:
- National Center for Education Statistics (NCES)
- National Science Foundation (NSF)
- Department of Housing and Urban Development (HUD)
- Health Care Financing Administration (HCFA)
- Social Security Administration (SSA)
- Bureau of Labor Statistics (BLS)

2. COMMON THEMES IN THE LICENSING AGREEMENTS

We noticed the following common themes in the licensing forms that we collected. Of course, any given form may differ from some of these themes.

2.1 Demonstration of the need for the data

The principal researcher must demonstrate that the data is required for research; i.e., public use data is not adequate. The goals of the research that require non-public data must be stated in the application. The licensor must approve of the research before the application process can proceed.

2.2 Designation of the group of people that will have access to the data

The principal researcher (PR) must supply a list of names of people who will be authorized to use the data. Those people must be informed of their responsibility not to share the data with people outside the group. The PR must indicate the group's experience, if any, with handling other licensed datasets.

2.3 Legal aspects of the agreement

The agreement specifies which people in the licensee's institution must sign the form. It also includes a statement concerning which law(s) protects the data (e.g., Privacy Act of 1974).

2.4 Data security, enforcement, and sanctions

A data security program must be developed and implemented. The licensee's institution must allow inspections of the area where the data are used and stored. Penalties for violations of aspects of the agreement are listed on the form (e.g., denial of use of other data from the licensor, fines, prison terms).

2.5 Restrictions on use of the data

There is a requirement that no attempt will be made to determine the identity of respondents. In general, the licensee is not allowed to link the licensed data to other microdata files.

2.6 Restrictions on release of the research results

Articles, reports, and statistical summaries must be reviewed by the agency before they are published or otherwise communicated. The results must adhere to the agency's disclosure limitation practices (e.g., all non-zero cells in a publicly released table must represent some minimum number of respondents).

2.7 Returning the data

There is a specified limit to the duration of the license. It is often less than two years. The licensee is often required to return or destroy the original and any derived files.

2.8 Cost to the licensee (not usually described in the form)

Some licensors require user fees. One type is an up-front fee in the form of a security bond as surety for maintaining confidentiality. Also the licensee must cover the cost of creating and maintaining a secure data handling environment.

3. INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (ICPSR)

3.1 File A: The Community Tracking Study (CTS) Physician Survey

This study is conducted by the Center for Studying Health System Change (HSC). It is sponsored by the Robert Wood Johnson Foundation (RWJ). ICPSR serves as an agent for RWJ, and RWJ owns the data. HSC promises anonymity to survey respondents.

The public use microdata files contain no geographic information. The restricted use files do contain some geographic information and can be used for sub-national analysis. The use of these files requires a contract. Researchers must show a need for the additional variables contained in the restricted use file and must agree to the conditions in the agreement. Also, the researcher must be employed by an organization that has a National Institutes of Health (NIH) Multiple Project Assurances (MPA) Certification Number (typically research universities.)

Licensees must not make any attempt to identify any person, family, household, business, or organization. Any inadvertent identification will be reported to ICPSR. Also, no attempt will be made to link the restricted data to any other dataset. The duration of the agreement is at most 24 months. Any violations of the agreement will require the immediate return or destruction of the data. If this is not done, legal remedies will be taken. Sanctions include the denial of all future access to CTS Restricted Data. A report of the violation is sent to the researcher's institution's office for scientific integrity and a request from ICPSR that sanctions be imposed on the researcher.

3.2 File B: The National Archive of Criminal Justice Data

Some public use data files from the Archive are available after variables that present confidentiality problems are deleted, aggregated, blanked, masked or otherwise modified. For some research, these public use files may not suffice. Researchers may request, in writing, a private use data collection. Their research interests must be documented in a written research proposal and receive human subjects' protection approval from the requester's institution. The requests are reviewed by the Archive and by the agency providing the data in question to the Archive. If the request is approved, the Requester and his/her Institutional Review Board (IRB) Chairperson or Institutional Coordinator must sign a Transfer Agreement consistent with the requirements of Section 28 of the Code of Federal Regulations.

4. SURVEY RESEARCH CENTER- UNIVERSITY OF MICHIGAN (SRC-UMICH)

Researchers may be eligible to receive restricted datasets from the Health and Retirement Study (HRS) if they meet certain requirements.

Only persons with formal affiliations with institutions that have been certified by the NIH as having met procedural criteria for Institutional Review Boards/Human Subjects Review Committees may have access to the data. The person primarily responsible for the research project must be a Principal Investigator or Co-Principal Investigator on a current federal government research grant or contract. This is because the primary sanction is notification (probably through the National Institute on Aging) to federal research funding agencies, with a possible recommendation of termination of current, and denial of future research funding to the investigators.

Applicants must agree not to link data from one HRS restricted dataset with data from other restricted HRS datasets. If the HRS restricted data contains information from Social Security Administration records, no linkages are to be made with any dataset containing geographic information at a level of aggregation more detailed than Census Division.

5. NATIONAL CENTER FOR EDUCATION STATISTICS (NCES)

As a historical note, let us mention that data licensing among federal agencies began with the NCES in 1990 (Duncan 1993). NCES's licensing system is based on the transfer of legal responsibilities to the licensee and related authorized users. Sanctions include a prison sentence of up to five years and/or a fine of up to $250,000.

To receive a loaned restricted use data file, an external user must file an application that includes a statement describing the proposed statistical use of the data, documents signed by a senior official at the user's institution and the principal project officer, notarized affidavits signed by all authorized users, and a security plan describing the handling and use of the data. Licenses are issued for periods of up to five years.

Licensees also agree to participate in unannounced on-site visits by security inspectors. NSF uses the same security inspectors as NCES. These inspectors are automating their report filing system to allow direct input to the master database of licenses. It will allow timely updating of authorized users, a problem at universities where many of the authorized users are research assistants.

The linking of licensed datasets to other datasets is allowed, but all use of the restricted data is limited to uses consistent with the statistical purpose described in the researcher's analysis plan.

Researchers are required to submit any reports or other releases of information derived from the restricted data to NCES for review for disclosure risk. NCES requires a minimum cell size of three (unweighted) cases with the provision that subsequent tabulations must not provide additional information which would disclose individual identities. And in the case of sample survey data, a minimum cell size of 30 is driven by variance considerations.

NCES and NSF are working on a joint project to revise their agency specific procedures manuals to produce a generic manual that will be used by both agencies. This generic manual could be used by other agencies as well. Printed versions of the manual will contain some details specific to each agency. The online version will show the details for the agency from which the Website was entered.

NCES and NSF have been sharing the same team of security inspectors for onsite monitoring of licensees. Since many of the licensees are at academic institutions, it is cost effective to combine inspection visits across agencies. Over the last year the contractor has staffed this activity with trained security officers who have experience handling information security issues for some of the more traditional security agencies in government. The agencies have been very pleased with the level of professionalism exhibited by this group.

6. THE NATIONAL SCIENCE FOUNDATION (NSF)

All available information is described in the previous section (5).

7. THE DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT (HUD)

Microdata on low-income public housing and on Indian housing may be obtained through a licensing procedure. Research organizations interested in the data must submit written requests to the Assistant Secretary for Policy Development and Research. These requests must include the following: name and title of the principal researcher, the research purpose, the list of persons who will have access to data, the description of data security procedures and, in particular, how HUD data will be kept separate from other data, description of the organization's experience in safeguarding confidential data, a statement that no attempt will be made to obtain the identity of any individuals in the dataset, a statement agreeing to provide HUD's Assistant Secretary for Policy Development and Research with copies of the portions of publications and other data products produced with these data.

A Memorandum of Understanding and Confidentiality, regarding the licensed use of microdata on individual households receiving housing subsidies, is signed. The licensed use is in effect for three years. The data are subject to the Privacy Act of 1974. Regarding published tabular results, no cell describing 10 or fewer cases (whether directly or through subtraction of other cells) can be released to a non-authorized user. Improper disclosures are to be reported to HUD's Office of the Assistant Secretary for Policy Development and Research. HUD reviews annually the organization's ability to maintain the confidentiality of the data.

8. THE HEALTH CARE FINANCING ADMINISTRATION (HCFA)

Microdata on Medicare beneficiaries' entitlement, enrollment, and claims paid are available through a "Data Use Agreement" which is similar to the licensing arrangements at other agencies. Under this agreement, HCFA retains ownership of the files. A "Custodian" of the file on behalf of the user must be designated. The user must notify HCFA of any change in the custodian.

No linking of records to files with identifiers can be performed unless specifically authorized by the HCFA System Manager or unless such linking is part of the approved protocol. The user agrees not to use the original or derivative data files without prior written approval. The user agrees to submit a copy of all findings to HCFA within 30 days of making such findings (research results). In terms of publishing results, the user cannot publish any lists of beneficiaries from which the identity of any beneficiary can be deduced. In particular, the user must be careful about revealing detailed information on: geography indicator, age, sex, diagnosis, procedure, admission/discharge date(s), and date of death. The user agrees that HCFA will be the judge of identifiability of an individual from the microdata (with a reasonable degree of certainty).

9. THE SOCIAL SECURITY ADMINISTRATION (SSA)

A Condition of Use agreement was recently approved but has not been used yet. It will be used in conjunction with a Memoranda of Understanding in some cases and alone in other cases. The parties to the agreement must agree that SSA retains ownership of the data files.

The recipient must identify all purposes for which the data are to be used. An individual must be designated as Custodian of the files on behalf of the recipient. The Custodian is responsible for enforcing all conditions of use and for the establishment and maintenance of security arrangements. The recipient will provide SSA with a list of all persons having access to the data. The recipient agrees not to try to identify individuals from tabular data supplied by SSA. The recipient agrees not to publish or otherwise release any listing of information extracted or derived from an individual record. The recipient agrees to provide to SSA with an annual inventory and accounting of all SSA original files and derivatives or copies of these files.

10. BUREAU OF LABOR STATISTICS (BLS)

10.1 File A: Census of Fatal Occupational Injuries (CFOI)

The licensed microdata from this census includes demographic characteristics of the decedent and information on how the fatal injury occurred. It contains no personal identifiers, and the lowest level of geography identified is Census region. "Authorized persons" are those individuals who are authorized access to the confidential information and have signed a BLS non-disclosure affidavit. The recipient's project coordinator will forward all signed affidavits to the BLS Project Coordinator prior to receiving access to the confidential information.

In terms of publishing research results, the recipient is to make every effort to release statistical information derived from the CFOI research file in such a way as to avoid inadvertent disclosure of individual persons or establishments. In particular, attention will be paid to tabulations with small cell sizes, and cells with fewer than three cases may not be published. All reports prepared using the CFOI file will be submitted to the BLS project coordinator for confidentiality review prior to publication or release to individuals other than authorized persons. The recipient will transmit to BLS copies of any final reports, research articles, or other media upon its publication or release.

10.2 File B: The National Longitudinal Survey of Youth (NLSY)

The NLSY is a national probability sample of more than 12,000 youth aged 14-21 as of January 1, 1979. Sponsored by the Bureau of Labor Statistics (BLS), these youth have been interviewed annually since 1979. The NLSY geographic-environmental (Geocode) data file contains information, which in certain circumstances, could permit identification of individuals who have participated in the survey. The Geocode data file includes the location of residence of each respondent for each survey year at the Standard Metropolitan Statistical Area or county level, and codes for schools attended.

11. OBSERVATIONS OF THOSE EXPERIENCED WITH LICENSING

Monitoring all of the licensing agreements can be quite a burden on a small staff, so be sure that the research can reasonably be expected to produce sufficient public benefits. Many "researchers" are interested in using survey data but most do not actually become productive information creators because they do not have training in the use of survey data. In addition, creating meaningful new aggregates requires a lot of work and analysis.

Be aware that the signers of these agreements don't always completely familiarize themselves with all of the conditions contained therein. Be aware that changes in custodianship of the data may not be reported even though this is spelled out in the agreement. You need to constantly remind the researcher about the conditions in the agreement they signed. They tend to innocently forget the rules, but they usually make good use of the data.

The term of the agreement should not be too far in the future. Situations such as lost data or changes in personnel can occur and cause follow up to become very difficult. The most powerful motivator for adhering to the restrictions on data usage is not the criminal penalty that may be imposed for misuse, but rather the possibility that misuse could lead to the denial of future requests for data. Thus, the latter should be included in the penalties. It is now possible to apply a "fingerprint" to a computer file so that if the file is "lent" by a researcher in violation of the licensing agreement, the researcher can be identified. This may prove to be a worthwhile practice.

12. REFERENCES

General references:
Jabine, Thomas B.(1993), "Procedures for Restricted Data Access," *J. Official Statistics*, vol. 9, no. 2, pp. 537-589.

Massell, Paul B.(1999), "Review of Data Licensing Agreements at U.S. Government Agencies and Research Organizations," paper presented at the Workshop on Confidentiality of and Access to Research Data Files, sponsored by the Committee on National Statistics (CNSTAT), Washington, D.C.

"Chapter 6 : Technical and Administrative Procedures," (1993) in George.T. Duncan, Thomas B. Jabine, Virginia A. de Wolf (eds.), *Private Lives and Public Policies*, National Academy Press, pp. 141-179.

Websites with licensing information for specific organizations or surveys
ICPSR: http://www.icpsr.umich.edu/ICPSR/About/Publications/Bulletin/Fall98/article.html
http://www.icpsr.umich.edu/NACJD/Private/private.html
NCES: http://nces.ed.gov/statprog/ : leads to web page entitled: Statistical Standards Program
NSF: http://www.nsf.gov/sbe/srs/srsdata.htm#MICRODATA
ADD Health: http://www.cpc.unc.edu/addhealth/datasets.html.

ESTABLISHMENT SURVEY DATA PRODUCTS OF THE CONFIDENTIALITY AND DATA ACCESS COMMITTEE

Jacob Bournazian, Energy Information Administration,[i] and
Virginia de Wolf, Office of Management and Budget
Contact: Jacob Bournazian, U.S. Dept of Energy, 1000 Independence Ave., SW, EI-42,
Washington, DC 20585 Jacob.Bournazian@eia.doe.gov

ABSTRACT

The Confidentiality and Data Access Committee (CDAC) is an inter-agency committee whose purpose is to promote cooperation and sharing of information concerning data access issues and statistical disclosure methods among Federal agencies. Currently, over 16 agencies have representatives on the Committee. This paper focuses on CDAC's products and activities that pertain to accessing and publishing data collected under a pledge of confidentiality from institutions, organizations, and business firms (called "establishment data"). It concludes with a description of future plans and activities.

Key words: Disclosure limitation methods, Restricted access procedures, Publicly available data

Federal statistical agencies often collect data from institutions, organizations, and business firms (termed "establishment data" in this paper) under a pledge of confidentiality. Each year a wide range of Federal agencies publish an increasing number of statistical data products from establishment surveys. The expanded use of the Internet has also improved the ease with which the public can access establishment data. Additionally, the increased capabilities of computers to access, process, and manipulate data raises important issues concerning how best to protect against the unauthorized disclosures of establishment data collected under an assurance of confidentiality.

Statistical agencies have two main options for protecting such confidential establishment data. The first method is to restrict the content of the published data sets or files. Agencies apply statistical methods to limit disclosure of respondent level data before disseminating data as either public-use microdata files . The other method is to *restrict access* to the microdata files. This involves defining the terms and conditions through a written agreement in which someone may access the agency data. The agreement specifies the purpose, the physical locations, and other appropriate conditions for protecting against the unauthorized release of confidential data. See Duncan et. al., (1993) and Jabine (1993). As advances are made in computer technology, Federal statistical agencies must continually consider what can be released as microdata files or tables using restricted data procedures and when to use restricted access procedures.

The Confidentiality and Data Access Committee (CDAC)[ii] is an inter-agency group comprised of staff members from Federal statistical agencies who work in the "confidentiality area". It operates as a forum where members share ideas on disclosure limitation methodology, and discuss problems, solutions and common approaches to issues concerning confidentiality and data access. CDAC was formed in 1995; over 16 Federal agencies are represented on the Committee. CDAC provides a mutually supportive environment in which individuals can ask questions and/or seek advice across agency boundaries on issues concerning data access and confidentiality issues. In order to encourage the open communication of ideas, only Federal employees may become members and the group's meetings are closed to members and invited guests. The closed meetings promote increased cooperation and sharing of statistical disclosure methods among Federal agencies by serving as a safe environment to discuss sensitive topics such as disclosure limitation methodology and data access issues.

This paper focuses on CDAC's products and activities that pertain to accessing and publishing data collected under a pledge of confidentiality from establishments. First, it describes the "Checklist on the Disclosure Potential of Proposed Data Releases" which was developed to help agencies determine the suitability of releasing public-use microdata files or tables. It then describes a current project to develop "auditing" software that would assess the degree of protection afforded tables that contain confidential information. The third activity is the new CDAC brochure, entitled "Confidentiality and Data Access Issues for Statistical Data," which was designed as a "primer" for people unfamiliar

with data access and confidentiality issues. It concludes with a description of the Committee's future plans to develop a checklist-like document that could be used by agencies to help them determine when, and how, to allow access to confidentiality data under restricted access procedures. It ends with a summary of future

THE CHECKLIST

Statistical agencies do not have full access to the kind of population data which is necessary for an agency to assess the risk of identifiablity when reviewing a data file for release to the public. As a result, many statistical agencies have resorted to rules of thumb or informed intuition to decide whether to permit public access to certain data files. One recommendation of the Federal Committee on Statistical Methodology's 1994 report "Report on Disclosure Limitation Methodology" (SPWP #22) was for each agency to centralize their review of disclosure-limited data products. SPWP #22 suggests that if the number of programs is small, such a review could be handled by one individual; alternatively, if an agency has multiple or large programs, a review panel, team, or board might be needed--in this paper the term Disclosure Review Board is used to refer to formal agency disclosure review process even though such a review might be handled by one person in the agency. The "Checklist on Disclosure Potential of Proposed Data Releases" (called Checklist) was developed as one tool to assist in an agency in such a review and in making a determination as to whether or not to release certain data files.

The Checklist consists of a series of questions that are designed to assist an agency's Disclosure Review Board in determining the suitability for release of microdata files and tabular data collected from individuals and organizations under an assurance of confidentiality. It begins with a cover sheet that asks for basic information about the proposed data release and then follows with three main sections. Section 1 pertains to microdata files that contain information from individuals or establishments, while Section 2 and 3 refer to tabular data from individuals and establishments, respectively.

Section 1: Microdata: Most microdata files contain demographic information. *Some questions in this section may not be applicable for establishment-based files.*

A major part of this section of the Checklist focuses on geographic information because it is the key factor in permitting inadvertent identification. In a demographic survey, few respondents could likely be identified if located within a single State, but more respondents -- especially those with rare and visible reported characteristics -- could be identified if located within a county or other geographic area with 100,000 or fewer persons.

The risk of inadvertent disclosure is higher with a publicly released data set that has both detailed geographic variables and a detailed, extensive set of survey variables. The risk is also often a function of the quality and quantity of "auxiliary" information (data from sources external to the data to be released). This auxiliary information is often difficult to assess for its disclosure risk. "Trimming" a data set by dropping survey variables, collapsing response categories for other variables, and/or introducing "noise" in the data, i.e., statistical perturbation, are techniques that may reduce the risk of inadvertent disclosure (Kim & Winkler, 1995).

For surveys of establishments the issues are generally different as such entities are often selected from very skewed populations. For example, in the U.S., there are only a handful or so of hospitals with 1,000 or more beds and inadvertent disclosure in a survey of hospitals might be possible using detail on the number of beds and geographic information as large as a Census region.

Section 2: Tabular Data from Persons or Households: This section pertains to tables that are based on data collected from persons or households (referred to as demographic data) under a pledge of confidentiality. There are two types of tables. Frequency count data tables show the number in the population with certain characteristics or, equivalently, the percent of the population with certain characteristics. Magnitude data tables present the aggregate of a "quantity of interest" over all units in the cell. Equivalently, the data may be presented as an average by dividing the aggregate by the number of units in the cell. Demographic data are typically reported as frequency count data.

Section 2 of this Checklist should always be completed if the tabulations are based on a complete count or an enumeration of the target population. Its use should also be considered when:

- the tabulations will identify small geographic areas, e.g., areas with populations less than 100,000;
- a large sampling fraction was used, as in the case of the decennial census long-form sample;
- the tables will have a large number of dimensions or cells; or
- the tables will cover especially sensitive topics.

Section 3: Tabular Data from Establishments or Other Types of Organizations: This section pertains to tabular data that are collected from organizations under a pledge of confidentiality. As with demographic data, tables can be of two types. Tables of frequency count data contain the number of units in a cell, such as a table of the number of establishments within the manufacturing sector by industrial classification group. Tables of magnitude data present the aggregate of a "quantity of interest" over all units in the cell, such as a table that presents the total value of shipments for those establishments in the same cells. Different statistical disclosure limitation methods can be used depending on the type of data being presented, although for practical purposes entirely rigorous definitions are not necessary.

For establishment data, an agency has to balance the level of detail for non-geographic variables in the file against the level of geographic detail to publish or allow public access. The first step in reviewing a request to release establishment data is to set the minimum population size for each geographic area and identify the variables on the file which may be used as geographic identifiers. These variables may be used as a set to easily identify individual respondents within a published cell. If a member of the population is unique with respect to a given characteristic set and is also in a survey sample, that respondent will be identifiable. In that case, the respondent would be unique to both the sample and the popultation. However, a respondent whose characteristic set is unique in the sample may not necessarily be unique in the population and therefore, not necessarily identifiable. Although it is rare that public use establishment data files are released, the checklist is an appropriate place to comprehensively review the issues associated with the public release of establishment data files.

Completing the Checklist

The Checklist was developed with the following considerations:

- It should be completed by a person who has appropriate statistical knowledge and who is familiar with the microdata file or tabular material in question (i.e., branch chief, survey manager, statistician, or programmer). While this implies a considerable familiarity with survey and statistical terminology, those without such background will nonetheless be able to understand much of what it is intended to accomplish. (Those who need a "primer" on statistical disclosure limitation methods, should see Chapter 2 of SPWP # 22, or information on the use of noise for tabular data should see Evans, Zayatz, and Slanta [1996]. Two additional useful resources are the 1996 Eurostat publication and the book by Willenborg and de Waal [1995]).

- Responses to questions in the Checklist are not intended to supply all the information that might be required by a Disclosure Review Board before a microdata file or table is released to the public. Some additional questions may need to be answered and/or given special consideration. Nonetheless, if files and tabular material are reviewed with the aid of the Checklist early enough, the need for time-consuming and costly re-programming of the data to be released can be avoided. This allows additional time for coordination with collaborators and other potential users.

Of course, there are two sides to the application of statistical disclosure limitation procedures: limiting disclosure risk and releasing data that are analytically useful. Products that meet the criteria for public release will often not meet all user requirements. When there are several methods of meeting requirements for public release, the alternative that provides data with the greatest analytic value clearly is preferred.

Uses for the Checklist

Users should complete the cover sheet and answer all questions for the applicable section(s). (Obviously, if it is distributed as a paper document, those who need more space for an answer would attach a continuation sheet and identify the number of the question.) The completed document should be submitted to the Disclosure Review Board.

In addition to helping an agency's Disclosure Review Board to determine the disclosure potential of proposed data releases, the Checklist has other uses:

- it can serve an important educational function for program staff who complete the Checklist;
- it can provide documentation when an agency is considering release of related data files and tabulations; and

- it could be very useful in defending legal challenges to an agency's decision to withhold certain tabular data or restrict data contained on a public-use file.

Note that the Checklist reflects the current standards of the Census Bureau and the National Center for Health Statistics for the release of data for public use.

SUPPRESSION AUDIT PROGRAM

Many Federal agencies conduct surveys on businesses where the individual survey responses must be kept confidential. When these data are published in tables or files released to the public, the published cells may contain values which disclose company level data that the Federal agency is obligated to protect as confidential. Cell suppression is a common technique for protecting frequency count and aggregate magnitude data from being used to disclose company level data. The suppression of these sensitive cells, called primary suppressions, does not always protect the confidentiality of the published data. To guarantee that a linear combination of cell estimates along a row or column may not be used to reveal primary suppressed cells, it may be necessary to suppress additional cells that are called complementary suppressions. Most agencies have developed automated software which processes the data through a disclosure sensitivity algorithm based on specific rules. Software also exists which determines necessary complementary suppression cells to prevent derivation of primary suppressed cells using various mathematical formulas.

Federal agencies are publishing complex survey data using various survey processing systems. Establishment data is plentiful with identifiable information. Sometimes clusters of data characteristics can be used to identify company level data. While Federal agencies use suppression software to protect against disclosing confidential information to the public, no information is available concerning the quality of the suppression patterns for the data that is released to the public. The protection range of a suppressed cell is one measure of the quality of a suppression pattern and may be defined as the minimum difference between the upper and lower value to the actual cell value. Many Federal agencies do no audit the suppression of their published data because of a lack of available low cost, easy, reliable, and efficient software. Without an automated system which is easily modifiable, agencies currently need to spend considerable resources to perform an audit of the suppressions for all tables in a publication.

CDAC launched an inter-agency project during 1999 to develop a suppression audit software for use by any Federal agency. Several system requirements were specified including the use of PROC LP from SAS software and the need to efficiently audit data tables of up to 5 dimensions. The first phase of the project was to develop a methodology for importing tabular cell data into the SAS auditing system software. This includes describing the format layout that all input data tables must conform to be read as an ASCII file input file to the SAS program. The first phase also includes specification of the program code; output data sets and descriptions. Seven agencies are curremtly participating in this project - Energy Information Administration, Bureau of Labor Statistics, Bureau of Economic Analysis, Bureau of Census, National Science Foundation, National Center for Education Statistics, and Statistics of Income, Internal Revenue Service. Four agencies, Energy Information Administration, Bureau of Labor Statistics, Bureau of Economic Analysis, Bureau of Census, have contributed test data sets for this project. The methodology being developed requires a CSV (comma separated value - ASCII) file as input to read in the SAS import routine.

The project is completing the first phase for designing the software module. A model is being developed from linear relationships within and among tables to generate the requisite computational data structures for the optimizer, PROC LP. The requisite computational data structures will be coded to allow for the future option of using alternate optimizers

besides PROC LP. Use of a common input format (i.e., sparse format according to SAS) allows input to alternate optimizers; however, that is just a side advantage because SAS PROC LP requires this format in order to operate. Specifically the specification describes in detail the mathematical solution to a 5 dimensional suppression audit problem and identifies the sequence of steps required to solve a 5 dimensional problem. Once the model design specification is approved by the participating agencies, the module or modules will be coded and integrated with the linear programming module which solves the problem. The integration of the these modules with an output display that summarizes suppression quality by primary and complementary suppressions will create a fully functioning auditing software. CDAC intends to make this software freely available on its web site after it has been developed and tested.

BROCHURE

CDAC newest product is its brochure entitled "Confidentiality and Data Access Issues for Statistical Data" which was written as a primer for people unfamiliar with these issues. Many businesses and establishments have to report on a wide variety of surveys that are conducted by several agencies. For example, a company may be filing survey forms with the 3 different Federal agencies and each agency may have different obligations concerning preserving the confidentiality of the data. The public is generally unaware that some statistical agencies are bound to preserve the confidentiality of their survey responses while other agencies may withhold release of the data only if there is an exemption under the Freedom of Information Act (FOIA). The brochure provides a comprehensive review of the disclosure limitation and restricted access policies and procedures that various government agencies follow to preserve the confidentiality of the reported data. It is one information product which addresses businesses and establishments concerns over the confidentiality of their survey responses. As establishment surveys struggle to main response rates each year, distribution of the brochure may be useful for gaining the trust and cooperation of respondent.

Checklist for Restricted Access Procedures

There are instances when the tabular/microdata products which summarize the establishment data are not adequate for certain types of research uses. To accommodate such users, some agencies have developed restricted access procedures that place conditions on who can use the data, for what purposes, at what locations, etc. (see Jabine, 1993, for an description of a broad array of restricted access procedures used by U.S. statistical agencies). There are two general types of restricted access procedures: licensing the data user before granting access; and limiting access to the data through secure sites called "Research Data Centers." The Census Bureau's Research Data Centers are one example of using restricted access procedures to enable researchers to have access to establishment microdata under carefully controlled conditions at one of three data centers. The Bureau added two other remote sites in California during 1999 and plans to open another in North Carolina in 2000. The National Center for Health Statistics opened a similar research data center in 1998. The National Center for Education Statistics licenses their researchers to use, at their university or research center, data sets that contain more detailed information than the "standard" public-use microdata file. Under licensing, the researcher must sign an agreement with the agency which permits the installation of the restricted data on their computer in return for meeting the agency's conditions relating to maintaining confidentiality of the data. One future project for CDAC will be to develop a checklist that Federal agencies may use to decide whether (and how) to provide restricted access to microdata files that the agency cannot release to the public due to the need to protect confidential information obtained from respondents.

Summary and Future Plans

CDAC has several projects for developing common procedures and methodologies for issues relating to accessing and publishing establishment data. The Checklist is not a "fixed" document and agencies are encouraged to modify the Checklist to suit their particular needs. The development of an audit suppression software will also be an easily modifiable program for an agency to use to audit the quality of the suppressions for the data being released. The Brochure on data confidentiality will be a useful information tool for conducting establishment surveys and publishing the survey results. All of the information products from CDAC are available through its web site at http://www.fcsm.gov/cdac_index.html. CDAC will continue to function as a resource for all Federal agencies and develop information products which are relevant to current and emerging issues and trends for issues relating to data confidentiality and privacy, and data access.

References and Resources

de Wolf, V.A. (1997). The "Interagency Confidentiality and Data Access Group," *American Statistical Association, 1997 Proceedings of the Government Statistics Section and Social Statistics Section*, 323-328.

Duncan, G.T., Jabine, T.B., & de Wolf, V.A. (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, D.C.: National Academy Press.

Eurostat. (1996). *Manual on Disclosure Control Methods*. (Catalogue #: CA-94-96-283-EN-C). Luxembourg: Eurostat.

Evans, T., Zayatz, L., & Slanta, J. (August 1996). "Using Noise for Disclosure Limitation of Tabular Data," *Proceedings of the 1996 Annual Research Conference and Technology Interchange*. Washington, DC: U.S. Department of Commerce, Bureau of the Census, 65-86.

Federal Committee on Statistical Methodology. (May 1978). *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. (Statistical Policy Working Paper 2). Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Federal Committee on Statistical Methodology. (May 1994). *Report on Statistical Disclosure Limitation Methodology*. (Statistical Policy Working Paper 22). Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

Jabine, T. B. (1993). "Procedures for Restricted Access," *Journal of Official Statistics*, 9(2), 537-589.

Kim, J.J. & Winkler, W.E. (1995). "Masking Microdata Files," *American Statistical Association, 1995 Proceedings of the Section on Survey Research Methods*, 114-119.

i For information about CDAC, contact Jacob Bournazian at EI-42, Forrestal Building, Washington, D.C. 20585, or alternatively, Jacob.Bournazian@eia.doe.gov, (202) 586-1256.

ii CDAC serves as a special interest committee on data access and confidentiality issues of the Federal Committee on Statistical Methodology (FCSM). The FCSM is an inter-agency committee, sponsored by the Office of Management and Budget, whose goals are to: disseminate information on statistical practices among all Federal statistical agencies; introduce new methodologies in Federal statistical programs to improve data quality; and provide a mechanism for statisticians in different Federal agencies to meet and exchange ideas. The FCSM reviews and analyzes a wide range of methodological issues concerning the publication of data by government agencies. In general, deliberations of FCSM subcommittees result in a Statistical Policy Working Paper (SPWP). Currently, twenty-six SPWPs have been released covering a wide range of topics. There are several working papers which relate to establishment data such as SPWP # 11 (*A Review of Industry Coding Systems*), SPWP #13 (*Federal Longitudinal Surveys*), SPWP #15 (*Quality in Establishment Surveys*), SPWP #18 (*Data Editing in Federal Statistical Agencies*), SPWP #21 (*Indirect Estimators in Federal Programs*), and, SPWP #22 (*Report on Statistical Disclosure Limitation Methodology*) For information on FCSM and its how to obtain copies of its SPWPs, go to its web site (*http://www.fcsm.gov*).

INCREASING ACCESS TO LONGITUDINAL BUSINESS SURVEY MICRODATA: THE CENSUS BUREAU'S RESEARCH DATA CENTER PROGRAM[1]

Arnold P. Reznek
U.S. Census Bureau, CECON Room 206 WPII, Washington, D.C. 20233-6300; rezne001@ces.census.gov

Since the early 1980s, the Census Bureau's Center for Economic Studies (CES) has developed research files of economic (business and establishment) microdata and has provided researchers with restricted access to these files – while maintaining appropriate safeguards to protect confidentiality and providing benefits to the Bureau's data collection programs. In the 1990s, the success of the CES program and increased demand for access prompted the Census Bureau to expand the program by establishing a limited number of Research Data Centers (RDCs) outside the Washington D.C. area. RDCs are secure sites, operated jointly by CES with institutional partners, where researchers may obtain restricted access for approved research projects. This paper describes the RDC program and how it provides access while protecting the confidentiality of the underlying microdata.

Key Words: Confidentiality, Disclosure Analysis

1. INTRODUCTION - THE CENTER FOR ECONOMIC STUDIES AND THE RESEARCH DATA CENTER PROGRAM

The U.S. Census Bureau's Center for Economic Studies (CES) carries out four main activities.

CES collects, archives, and develops research microdata files, from the historical data files underlying the Census Bureau's surveys. These files are often longitudinal. Until very recently, they were almost exclusively from the Census Bureau's surveys of business establishments and firms (surveys), but the files now include data from the 1990 Decennial Census and from a growing number of the surveys of households and individuals (demographic surveys). This paper is almost exclusively concerned with the economic data.

CES operates a research program using these longitudinal research microdata sets. This research program is carried out both by CES permanent staff and by research associates from outside the Census Bureau. Access is provided in secure facilities -- with appropriate safeguards to protect data confidentiality.

The safeguards are necessary because under U.S. law, microdata from Census Bureau's data programs are confidential. They may be used only for statistical purposes and only at secure sites by Census Bureau employees or by individuals whom the Bureau has granted Special Sworn Status (SSS). The law provides specific penalties for violations.[2] Although the law protects the privacy rights of respondents to Census Bureau survey forms, an important practical reason to ensure confidentiality is that publication (however inadvertent) of microdata would most likely reduce co-operation with data collection programs. In addition, some files are collected under the sponsorship of other agencies, or they may contain various amounts of administrative data from other agencies. In providing restricted access to these data, we must adhere to all applicable laws and regulations, as must all Census Bureau programs.

CES researchers – termed research associates – are Census Bureau customers and data users with analytical interests. By providing secure restricted access to microdata, CES creates a constructive interchange among the Census Bureau, researchers, program sponsors, and other major data users and policy makers in the government, business, and academic research communities. This provides an opportunity for unique research that advances science and informs policy while providing basic measurement research and feedback that can advance Census Bureau data programs. Triplett (1991) makes the case for analytic research programs within statistical agencies.

CES supports a system of Research Data Centers (RDCs). The RDCs provide researchers with access to data at secure sites in and outside the Washington, D.C. area. The "remote" sites increase capacity and reduce researchers' costs of relocating to carry out research projects. We operate the following remote RDCs jointly with institutional partners:

- The Boston Research Data Center (BRDC), operated with the National Bureau of Economic Research (NBER), opened in January, 1994.
- The Carnegie Mellon Census Research Data Center (CMCRDC), operated with the Heinz School of Public Policy and Management at Carnegie Mellon University, opened in January, 1997.
- The California Census Research Data Center (CCRDC), operated with the University of California, opened in early summer, 1999. The CCRDC has two sites, one at UCLA and one at UC Berkeley.
- A new Triangle Census Research Data Center, (TCRDC), to be operated jointly with Duke University, has been approved for the Research Triangle area of North Carolina, will open in 2000.

CES directly supports Census Bureau data collection programs, through projects undertaken by CES staff and research associates – sometimes jointly with staff of the Census Bureau's data production divisions.[3]

This rest of the paper proceeds as follows. Sections 2 and 3 describe some of the databases at CES and briefly characterize some of the research results obtained. Sections 4 and 5 summarize the historical development of the program and describe the program's operating model. Section 5 gives conclusions and future directions.

2. CES DATABASES

CES has developed a large number of research databases from the Census Bureau's economic data. Most are longitudinal, containing linked observations over time on individual business units. They can be linked at the business

unit level, increasing the value of any individual database. Researchers themselves played key roles in their development.

In the early years, CES databases centered around the *Longitudinal Research Database* (LRD). The oldest and perhaps best-known database, the LRD is a panel of annual establishment-level cost and output data from the Census of Manufactures (since 1963) and from the Annual Survey of Manufactures (since 1972). Several other manufacturing databases have been linked to the LRD, including the *Survey of Manufacturing Technology (SMT) Database*, the *Pollution Abatement Cost and Expenditures (PACE) Database* (Streitwieser 1996), the *Manufacturing Energy Consumption Survey (MECS) Database*, and the *Research and Development (R&D) Database* (Adams and Peck 1994).

Major initiatives are underway to expand the establishment and firm databases to include business sectors beyond manufacturing. Several databases reach beyond the manufacturing sector, including the *Quarterly Financial Reports (QFR) Database* and the *Characteristics of Business Owners (CBO) Database* (Headd 1999). We are currently developing a *Longitudinal Business Database (LBD)* that includes economy wide data from two major sources: the Quinquennial Economic Census establishment information from the Censuses of Construction, Manufactures, Wholesale Trade, Retail Trade, and Services; and the Standard Statistical Establishment List (SSEL; U.S. Bureau of the Census [1979]), a database of all U.S. business firms and their establishments that is used as a sampling frame for the Economic Censuses and over 100 current Census Bureau surveys. As always, research projects are crucial in building the LBD.

In the last several years, a major focus has been the development of *employer-employee* data sets that combine information from both the demand (firm) and supply (worker) sides of the labor market. These data sets are valuable for studying a variety of issues in labor economics -- firm and worker contributions to pay determination; the effects of technological change on worker pay and earnings inequality; the returns to worker and firm from worker training, and more. Bayard, Hellerstein, and Troske (2000) and McKinney (2000) describe these efforts.

3. SOME RESEARCH RESULTS

In this paper, we can only characterize CES research very generally. The research has a common theme: plant and firm behavior is very heterogeneous. Jensen and McGuckin (1997) synthesize several years of recent research on firm performance, firm and industry evolution, and economic growth. Pointing out the differences between research using longitudinal microdata and more traditional economic research using aggregated data, they summarize a growing body of evidence that microdata are essential for understanding economic performance and competition -- aggregate data mask most of the "action" in the data. Haltiwanger (1997) summarizes recently emerging evidence that the heterogeneity observed in longitudinal microdata is crucial for understanding macroeconomic as well as microeconomic fluctuations. He describes a new macroeconomic measurement and modeling approach that exploits these findings, and proposes a long range longitudinal micro database collection, processing, and measurement strategy for supporting it.

CES research results have influenced and been influenced by the development of both micro- and macroeconomic theories that explicitly allow heterogeneous behavior by individual businesses. For example, see Jovanovic (2000) on the asynchronous adjustment of complementary inputs, and a series of papers by Caballero, Engel, and Haltiwanger that are described in Haltiwanger (1997).

CES research has generated several new Census Bureau data products: an index of product diversification (Gollop and Monahan 1991); an index of high technology trade (McGuckin et al 1992) and a set of data series on gross job creation and destruction in the manufacturing sector (Davis, Haltiwanger, and Schuh [1996]).[4]

CES research has resulted in feedback to the Census Bureau survey programs described above. To enhance and formalize this feedback, CES staff members attend survey meetings and serve on various program-related committees and task forces. How to make this aspect of the program work best is an ongoing issue, illustrating the problem of integrating basic research into a production environment.

The capabilities to provide survey sponsors with direct access to survey microdata and to link surveys have increased the interaction between data users and data collectors. For example, Federal Reserve Board researchers have investigated quality and usefulness of the FED-sponsored Survey of Plant Capacity (e.g., Doyle 2000). Moreover, access has helped to attract two new surveys. (1) The National Employer Survey (NES), sponsored by the Department of Education, is the first nationally representative survey of employers (establishments) that documents their practices and expectations concerning education and training, as well as the quality of their workforces.[5] (2) The second survey consists of employer components of the Medical Expenditure Panel Survey Insurance Component (MEPS-IC), sponsored by the Agency for Healthcare Research and Quality (AHRQ). This survey collects data on health insurance plans obtained through employers, unions, or other private sources. The data for both surveys are available for use at RDCs.[6]

Much CES research investigates measurement issues. A long line of CES research has contributed to and continues to investigate issues related to the North American Industry Classification System (NAICS). Most recently, Klimek and Merrell (2000) presented work at ICES-II that uses the microdata to provide NAICS-based classifications for 1992 data. In addition, CES staff members are involved in the Census Bureau's efforts to measure the developing digital economy. Mesenbourg (2000; ICES-II) presents a summary of the Census Bureau's efforts in this area.

4. DEVELOPMENT OF CES AND THE RDC PROGRAM

CES was established in 1983, to provide restricted access[7] to the Census Bureau's economic (business establishment and firm) microdata for the manufacturing sector. Access to these data had been granted sporadically at times since the 1950s, but demand could not be satisfied systematically until the early 1980s (Kallek 1982; McGuckin and Pascoe 1988).

The CES program was quite successful, attracting a significant number of researchers by the early 1990s. However, it was (and is) often very expensive for researchers to relocate to CES headquarters for the time required for projects, and space is limited there. To mitigate this problem, McGuckin (1992) CES proposed what essentially is the current RDC operating model, and the Census Bureau agreed to a limited pilot program. The first pilot was the BRDC In Boston (January 1994), established with support from a grant from the NSF(NSF) to the NBER. The CMCRDC (January 1997) was the second pilot, and the first RDC not at a Census Bureau facility. The Heinz School of Public Policy and Management at Carnegie Mellon University volunteered to host that pilot site.

With the success of the two pilots, by 1997 we received expressions of interest in starting RDCs from a number of locations. However, we could not accommodate them all due to the limited resources available to maintain and support RDCs. To expand, we continued our longstanding partnership with NSF, developing a plan with several goals. First, we wanted to ensure fair, objective and appropriate RDC placements -- in areas that promised the best combination of high quality research and benefits to the Census Bureau's mission. Together, the Census Bureau and NSF developed a call for proposals under which various sites would compete for the opportunity to host RDCs. NSF provides some potential seed funding and proposals are evaluated by the Census Bureau and by NSF's peer review process. Successful RDC partners demonstrate the ability to work with the Census bureau to provide fair and objective access to researchers; the existence of a regional research community of sufficient size and quality to yield high quality research output; the presence of a coherent plan for long term funding; an understanding of the need to work with the Census Bureau to protect the confidentiality of the underlying microdata and the need to ensure that researchers use the data only for statistical purposes (i.e., not for administrative purposes such as regulation or enforcement).

Second, we needed to improve our support mechanisms to support an expanded system of RDCs, by standardizing our project selection process, improving and updating our database documentation, expanding the range of data available at RDCs, and upgrading our computing capacity. This process is ongoing.

Given these criteria for expansion, the Census Bureau decided it could support up to four new RDCs over the next few years. At the end of 1997and in January 1998, we announced the availability of the new RDC placement opportunities These announcements consisted of a letter and prospectus mailed to a large number of potential RDC partners (academic and research organizations), postings on the Census Bureau web site, and a Federal Register announcement. Applications would be accepted and reviewed every six months under NSF's regular social science proposal review procedures (submission deadlines are January 15 and August 15).

Under this process, we have chosen two new RDCs, with three RDC sites: the CCRDC (summer 1999), which is one RDC with two facilities – at UCLA and UC Berkeley; and the TCRDC, which should open in 2000.

5. THE RDC OPERATING MODEL

Based on six years of experience in operating RDCs, we have developed an operating model for RDCs , which we describe in this section. The essentials of this model were first proposed by McGuckin (1992).

5.1. Selecting New RDCs.

New RDC sites are chosen using the procedures described in the previous section.

5.2. Establishing and Operating New RDCs

Before research begins at an RDC, the Census Bureau and its new RDC partner prepare the site to receive and support data and researchers. We develop a Memorandum of Understanding (MOU) that specifies the way the RDC will be operated and what each partner will provide. The partner is responsible for covering direct costs of operation at the site, including office space, local computer hardware and software, the salary for a CES employee to be stationed on site as the RDC administrator (the administrator's duties are described below), and any other personnel, such as an Executive Director. These costs are covered by funds raised by the local research community. During the first three years of operation, the NSF provides up to 100,000 per year as "seed money" to partially cover these costs. RDCs recover these costs through laboratory fees on research projects. The Census Bureau contributes databases; new database development and documentation; and subject matter, data, and administrative support. The RDC Administrator is recruited jointly by the Census Bureau and its partner but is hired by the Census Bureau.

Each RDC has a panel that provides overall guidance and oversight. Panel members represent the RDC partner and other organizations with an interest in the RDC. The panel ensures open and fair access and assesses operations, and panel members review proposals for scientific merit and feasibility. The Census Bureau Executive Staff and the Census Advisory Committee of Professional Associations (particularly the American Economic Association subcommittee) provide oversight at CES headquarters. We are also developing mechanisms through which the panels relate to each other and through which they have a voice in Census Bureau policies toward RDCs.

5.3. Maintaining Confidentiality and Security

Protecting the confidentiality of the data is paramount at RDCs, at all stages in establishing and operating the RDC and at all stages in the life of RDC research projects. Reznek and Nucci (2000) give details; here we provide a relatively brief summary. Ensuring confidentiality involves providing a physically secure office; imparting to researchers at the RDC the Census Bureau "culture of confidentiality;" and putting in place policies and procedures for protecting confidentiality protection and for releasing research output.

An RDC is a secure physical location, as certified by the Census Bureau security office. Protecting security requires providing physical (office) security, computer security, and data security. Each RDC has a security plan developed and

approved according to established Census Bureau procedures. The RDC office is in a secure (locked) room (or rooms) with a security system that meets Census Bureau specifications.

Access to the RDC facility is given only to Census Bureau employees or other persons to whom the Census Bureau has granted Special Sworn Status (SSS) - including researchers carrying out approved projects at the RDC and certain others who have a need to enter (e.g., specified local computer staff or RDC staff members). To be granted SSS, an individual must obtain a security clearance and sign a sworn agreement to preserve the confidentiality of the data. SSS allows researchers access only to the confidential data needed for their approved research projects. Persons with SSS and Census Bureau employees are subject to the same legal penalties for revealing confidential information as are regular Census Bureau employees.[8] Another, equally important, requirement for SSS is that the researcher's project must benefit the Census Bureau's data programs.

The RDC provides a secure computer network. It is not possible to access confidential data from outside the RDC (e.g., by remote terminal). Researchers are accountable for their computer use, through the use of passwords and system logs. Researchers may not bring into the RDC facility laptop computers, zip drives or other portable mass storage devices, including devices with wireless modems. The RDC computers are set up to prevent copying of data to removable storage media. Also, approved procedures exist for storing and disposing of confidential data, and for transferring these data from one secure location to another.

5.4. The Role of the RDC Administrator

The RDC administrator is essential for RDC operations. To maintain security and confidentiality of the data, the administrator instills the Census Bureau's "culture of confidentiality" into the researchers and provides guidance to the researchers regarding security and confidentiality restrictions. The administrator examines any results the research associates wish to remove from the secure facilities, ensuring that Census Bureau policies are followed to prevent release of confidential data. This examination of research output is called *disclosure analysis*. The administrator also provides local administrative, computer, data, and subject matter support, and acts as a liaison between the researchers and CES (as well as the rest of the Census Bureau).

In carrying out all duties, the administrators consult as needed with CES management and staff members. To function effectively, the administrators must have a research backgrounds, and they are researchers in their own right.

5.4. Carrying out Research Projects at RDCs

Researchers must submit proposals to carry out projects at RDCs; these proposals must follow a set of guidelines.[9] We currently choose new research projects roughly every two months according to a proposal review cycle. Proposals undergo careful review at the RDC, at the Census Bureau, and sometimes by outside agencies, including research funding agencies such as NSF or agencies that sponsor Census Bureau surveys. The selection criteria include need for access to confidential Census Bureau data; potential to benefit Census Bureau data programs; scientific merit, feasibility, and risk of disclosing confidential information.

We emphasize (and reemphasize) to researchers that we can release a much greater range of information for analytic results (e.g. regression coefficients) than for tabular data (it is hard to untangle an $X'X$ matrix). Indeed, we typically release only a minimal amount of tabular output because secondary disclosure is very difficult: the operating divisions release a large amount of tabular data, severely limiting the number of possible extra tabulations.

A formal agreement is written for each approved project, specifying the scope of the project, the data and services to be provided by both researcher and CES, reports and other obligations of both parties, and the project term (including duration and intensity of laboratory use). Projects are charged laboratory fees to cover the costs of support. The fee structure is identical for all RDCs, and the fees go directly to the RDC where the project takes place. The fees cover the direct costs of operating the RDC -- personnel, space, computing facilities. Extra fees are charged for projects that require unusual amounts of support – for example, obtaining new data sets that require special programming efforts by CES or other Census Bureau personnel; or special data or subject matter consultation such as aid in matching to outside data sets. On the other hand, RDCs subsidize a limited number of graduate school researchers. Also, we may reduce fees for projects that make a particularly valuable contribution to CES or other Census Bureau data programs, such as database development and documentation.

At project start, the researcher is given SSS; the RDC administrator gives the researcher "awareness training" on security and confidentiality policies – including procedures for release of research output; and the researcher is given access to the software and data needed for the project. We do not restrict project-related analyses the researcher may carry out on site. (One exception: we do not allow casual "browsing" of the data sets, which in any case do not contain obvious identifiers such as name and address). Researchers submit all research output to the RDC administrator for clearance to remove it from the RDC, and they work with the administrator to ensure that the clearance goes as smoothly as possible. We also expect researchers to submit papers for the CES Discussion Paper series and to provide CES with their final published research papers and reports. This maintains a record of the research results and ensures that the project benefits the Census Bureau's data programs and future researchers.

6. CONCLUSIONS

We have described CES and its RDC program, which grants researchers restricted access to Census Bureau confidential microdata at several secure facilities. The program has been successful; one indication is that in the U.S.,

the National Center for Health Statistics has established its own RDC, and other U.S. Federal statistical agencies have expressed interest. Canada and the Netherlands have established RDCs, and other countries are considering it.

Nevertheless, much more remains to be done, particularly in the U.S. Although increasing access to U.S. Federal microdata should yield great benefits for both data users and statistical agencies, we feel such efforts should be coordinated and directed toward the establishment of "Federal RDCs." At FRDCs, data from multiple statistical agencies would be made available to researchers. Such interagency collaboration would avoid costly duplication of facilities, and would provide one way to satisfy demands for a more coherent, integrated statistical system. However, establishing and operating FRDCs would generate considerable problems of coordinating the policies and procedures of all agencies involved. U.S. data sharing legislation now under consideration – it has passed the House of Representatives and is now in the Senate – would allow a limited number of statistical agencies to share their confidential micro data for specific purposes. We do not know whether this legislation would allow FRDCs, or whether FRDCs could be established without such legislation. Given permission to establish FRDCs, we would face issues similar to those discussed above in expanding our current system of Census Bureau RDCs. We feel that the solution should be similar, including the procedures for establishing and operating RDCs.

NOTES

1. This paper reports the results and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties of research and to encourage discussion. I thank Robert McGuckin (former Chief of CES), John Haltiwanger (formerly Chief Economist of the Census Bureau), and J. Bradford Jensen (currently Director of CES), who developed many of the ideas, policies, and initiatives summarized in this paper. Any errors of fact or interpretation are mine.
2. The relevant law is Title 13, U.S.C., Section 214. Under this section and Title 18 U.S.C Sections 3559 and 3571, violations are punishable with a fine of not more than $250,000 and imprisonment of not more than five years, or both.
3. For more information on CES and the RDC program, including the databases and research projects, see the CES web pages, under "C" in the A to Z list on the Census Bureau's web site, www.census.gov.
4. Updated statistics are available on Haltiwanger's web site, http://www.bsos.umd.edu/econ/haltiwanger/.
5. The NES is carried out through educational research institutes associated with the University of Pennsylvania. For more information on the NES, see the survey web site: http://www.irhe.upenn.edu/research/research-main.html.
6. For more information, see the AHRQ website: http://www.meps.ahrq.gov.
7. The requirement to maintain confidentiality of the data they collect prevents the Census Bureau and other statistical agencies from releasing to the public all the data they collect. In the face of these requirements, agencies have developed two ways of providing access to their data, termed *restricted data* and *restricted access*. For an excellent discussion of restricted access and restricted data, see Duncan, Jabine, and deWolf, 1993, pp. 141-42 .
8. In reality, research associates may face greater penalties than Census Bureau employees for disclosing confidential information, including loss of professional credibility and denial of future access to the microdata.
9. The guidelines are on the CES web pages, on the Census Bureau's web site. Interested researchers may also send email to proposals@ces.census.gov.

REFERENCES

Adams, James D. and Suzanne Peck (1994). "A Guide To R&D Data At The Center For Economic Studies U.S. Bureau of The Census." CES Discussion Paper 94-9 (August).

Davis, Steven, John Haltiwanger, and Scott Schuh (1996). *Job Creation and Destruction*. Cambridge: MIT Press.

Bayard, Kimberly, Judith Hellerstein, and Kenneth Troske (2000). "The New Worker-Employer Characteristics Database." Paper prepared for the International Conference on Establishment Surveys - II, June 17-21, 2000, Buffalo, NY.

Duncan, G.T., T.B. Jabine, and V.A. deWolf, eds. (1993). *Private Lives and Public Policies*. Washington, D.C., National Academy Press.

Gollop, Frank M., and James L. Monahan (1991). "A Generalized Index of Diversification: Trends in U.S. Manufacturing." *Review of Economics and Statistics*, Vol. LXXIII, No. 2 (May): pp. 318-330.

Haltiwanger, John (1997). "Measuring and Analyzing Aggregate Fluctuations: The Importance of Building from Microeconomic Evidence." *Federal Reserve Bank of St. Louis Review*, (May/June), pp. 55-78.

Headd, Brian (1999). " The Characteristics of Business Owners Database, 1992." CES Discussion Paper 99-8 (August).

Jensen, J. Bradford and Robert H. McGuckin (1997). "Firm Performance and Evolution: Empirical Regularities in the U.S. Microdata." *Industrial and Corporate Change*, Volume 6, Number 1, pp. 25-47.

Kallek, S., "Objectives and Framework" (1982). In U.S. Bureau of the Census, *Development and Use of Longitudinal Establishment Data*. Economic Research Report ER-4. Washington, D.C.: U.S. Government Printing Office.

Klimek, Shawn D. and David R. Merrell (2000). "Industrial Reclassification from the Standard Industrial Classification System to the North American Industry Classification System." Paper for presentation at the International Conference on Establishment Surveys- II, Buffalo NY, June 17-21, 2000.

Jovanovic, Boyan (2000). "Optimal Adoption of Complementary Technologies." *American Economic Review.* Vol. 90, No. 1, (March), pp. 15-29.

McGuckin, Robert H. (1992). "Analytic Use of Confidential Data: A Model for Researcher Access with Confidentiality Protection." CES Discussion Paper 92-8 (August).

McGuckin, Robert H., Thomas A. Abbott, Paul Herrick, and I. Leroy Norfolk (1992). "Measuring Advanced Technology Products Trade: A New Approach." *Journal of Official Statistics*, Vol. 8, No. 2, pp. 223-233.

McGuckin, Robert H. and George A. Pascoe, "The Longitudinal Research Database: Status and Research Possibilities" (1988). *Survey of Current Business*, November 1988, pp. 30-37.

McKinney, Kevin (2000), "The Longitudinal Employer - Household Dynamics Project : Using State UI Data." Presentation at the International Conference on Establishment Surveys - II, June 17-21, 2000, Buffalo, NY.

Mesenbourg, Thomas L. "Satisfying Emerging Data needs: Measuring Electronic Business. Paper for presentation at the International Conference on Establishment Surveys II, Buffalo NY, June 17-21, 2000.

Reznek, Arnold P. and Alfred R. Nucci (2000). "Protecting Confidential Data at Restricted Access Sites: Census Bureau Research Data Centers." *Of Significance*. Association of Public Data Users (APDU) (forthcoming).

Streitwieser, Mary L. (1996). "Evaluation and Use of the Pollution Abatement Costs and Expenditures Survey Micro Data." CES Discussion Paper 96-1, January.

Triplett, Jack E (1991). "The Federal Data System's Response to Emerging Data Needs." *Journal of Economic and Social Measurement*. Vol. 17, pp. 155-77.

U.S. Bureau of the Census (1979). *The Standard Statistical Establishment List Program.* Bureau of the Census Technical Paper 44. Washington, D. C: U.S. Government Printing Office.