Restricted Access Procedures

Federal Committee on Statistical Methodology
Confidentiality and Data Access Committee

SUMMARY: This document discusses methods for providing access to statistical data while limiting the risk of disclosure of confidential information. Research data centers (RDCs), permit onsite use of confidential files a closely delimited area with specialized equipment and extreme security. Remote access over secure electronic lines to dedicated computers is a more recent development, as is electronic access to data bases previously subjected to disclosure limitation techniques. Fellowships and post-doctoral programs, in which researchers can be treated as agency employees, permit a less restrictive form of access. Finally, some agencies permit the researcher to use confidential data offsite, but under highly restricted conditions as spelled out in a legally binding agreement.

INTRODUCTION

The same laws that require information collected in Federal statistical surveys to be released to the public whenever it is possible to do so, also prohibit the disclosure of personal identities when the data providers have been promised confidentiality. In order to comply with both of these legal (and ethical) imperatives, federal agencies and organizations must find ways to release as much of this information as possible for research, while upholding the confidentiality pledge. Adding to the tension is the fact that while data-rich microdata files allow for very detailed analyses by researchers, they also present an unacceptable risk of disclosure because they contain information on geographic details as well as an extensive set of survey variables. A record representing an individual or establishment reflects a unique combination of many characteristics, and the likelihood of re-identification (disclosures of confidential information) is greater in comparison to less detailed tabular material.

In most cases, the application of statistical disclosure limitation measures have kept pace with the demand for more detailed information in highly powerful and flexible (i.e. electronic) formats. However, such techniques have limitations, for their stringent application to preserve respondent confidentiality can result in a data set that has little research utility. When this happens the research community as well as the general public can suffer.

One way of satisfying the twin concerns of data access and confidentiality, is for the agency or research organization to release the file under highly controlled conditions. Until recently, much valuable information could not be made available to researchers and others. Within the last decade, however, refinements and innovations have been made that have significantly increased the range and depth of access to materials previously not released or available only to very few users. In an effort to make federal agencies and others aware of some of those developments this

report discusses four of them as they are found at selected agencies and institutions[1]: Research Data Centers, Remote Access and On Line Query Systems, Licensing Agreements, and Research Fellowships and Post Doctoral Programs.

RESEARCH DATA CENTERS (RDC)

RDC's allow researchers to use restricted access data at the offices of the data holder, or at a site under it's control, under highly restricted conditions. The essential characteristics of these centers are:

- submission of research proposal;
- formal agreement covering work to be done, data used, and types of output;
- the use of data files stripped of personal identifiers;
- limitations on types of analysis;
- special procedures governing use of any outside (linkable) data brought in by researcher;
- dedicated computers;
- disclosure review of output;
- inspection of material removed from site; and
- physical presence of and oversight by agency staff.

Research review, formal agreements, data and analytical limitations: Before access can be granted, it is necessary to have a clear idea of the nature of the research proposal so that a decision can be made to grant access. Researchers sometimes misunderstand the type of data to which they may have access as well as the output that can be seen within and removed from the RDC (though few question the need to strip names and addresses from any files made available). Furthermore not all analytical approaches and associated output are permitted (disclosure review of tabular data is seen by some as problematic - see discussion of Census RDC). Formal agreements specifying persons who will be permitted inside the RDC, duration and timing of the research, data to be used, analytical techniques, and detailed output help to insure the appropriateness of the research and resources to be employed.

Linkable data, and internal procedures: Frequently, researchers want to enhance data available in the RDC with data from another source. For example, a national health survey of individual families may be linked to a file with health care provider characteristics of the counties within which the sampled families reside. However, the RDC cannot provide the researcher the means to link these two files without violating assurances of confidentiality. The RDC can, however, merge the two files and then permit access within the RDC (with county identifiers removed). Within the RDC, the files accessed reside on carefully configured

[1] Discussion of research data centers in this paper is restricted to those developed by U.S. federal health and statistical agencies. For information on such centers at Statistics Canada, the University of Michigan, the Department of Justice, the Department of Agriculture, and the Immigration and Naturalization Service, see papers presented at the 2001 Joint Statistical Meetings session entitled "Enhancing Researchers Access to Confidentiality Data: Five Case Studies" at http://www.amstat.org/meetings/jsm/2001/index.cfm?fuseaction=activity_details&activityid=237&sessionid=200619

computers, permitting no external contact and with supervision of printouts. Researchers are not permitted to bring any materials to the RDC that would permit disclosure of individual information nor can they remove printouts that have not been reviewed by RDC staff.

In the U.S., three agencies have established RDCs: the Census Bureau, the National Center for Health Statistics, and the Agency for Health Care Research and Quality. The agencies operate their RDCs in similar but somewhat different ways. The following discusses the way each agency operates its RDCs.

U.S. Census Bureau - the Center for Economic Studies (CES) and its RDCs

The Census Bureau pioneered the RDC concept, and this section has two parts. The first part briefly summarizes the history of the Census Bureau's Center for Economic Studies and its RDC program. The second part describes how the Census Bureau establishes and operates RDCs.

History of CES and its RDCs: CES was established in 1983 to provide restricted access 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 the demand could not be satisfied systematically until the early 1980s.

To accommodate increasing demand, and to make costs of accessing RDCs more equitable for those located far from Washington, D.C., the Census Bureau established two pilot RDCs away from Census Bureau headquarters: the Boston Research Data Center, 1994 at the Census Bureau's Boston Regional Office, and in 1997, at Carnegie Mellon University's Heinz School of Public Policy and Management.

With the success of the two pilots, by 1997 the Census Bureau had received expressions of interest in starting RDCs from a number of locations. By 2000, two additional RDCs had been established in California (at UCLA and UC Berkeley) and in North Carolina (Duke University).

RDC Operations: 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.

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. The following discussion is a relatively brief summary of RDC procedures (for more detail, the reader is referred to Reznek and Nucci (2000)).

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.

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. Another, equally important, requirement for SSS is that the researcher's project must benefit the Census Bureau's data programs.

The Census Bureau stations a CES employee, the RDC administrator, at each RDC. 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 researchers 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 administrator consults with the CES management and staff members as appropriate. To function effectively, the administrator must have a research background, so the administrators are researchers in their own right.

The RDC provides a secure computer network. Researchers may not bring laptop computers, zip drives, or other portable mass storage devices, including devices with wireless modems into the RDC facility. 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.

In response to increasing concerns about security, and to promote efficiency, the Census Bureau RDC system is now converting from secure local networks of PCs and Unix workstations to a "thin client" environment. Under this environment, no confidential data will be stored at the RDCs. Instead, the data will be stored on a secure server (or servers) at Census Bureau headquarters. This server will be located in a "demilitarized zone," with firewalls separating it from both the external Internet and the Census Bureau's internal network. From the RDC offices, researchers will use X-terminals ("thin clients") to access the data authorized for their projects, which resides on the central server. The RDCs are connected to the server via dedicated T-1 lines. Researchers are accountable for their computer use, through the use of passwords and system logs. Researchers may print only when an RDC administrator is present.

Researchers must submit proposals to carry out projects at RDCs and these proposals must follow a set of guidelines. The Census Bureau and the RDCs approve 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. For more details on the proposal selection process, see the CES web site at http://www.ces.census.gov.

The RDC administrators emphasize (and reemphasize) to researchers that it is possible to release a much greater range of information for analytic results (e.g. regression coefficients) than for tabular data. Indeed, it is typical to release only a minimal amount of tabular output because the assessment of secondary disclosure risk is very difficult. Census Bureau 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. Moreover, fees may be reduced 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. No restrictions are placed on the project-related analyses the researcher may carry out on-site (one exception: casual "browsing" of the data sets is not allowed; but in any case the research data sets do not contain obvious identifiers such as name and address.) Researchers must submit all research output to the RDC administrator for clearance to remove from the RDC, and must work with the administrator to ensure that the clearance goes as smoothly as possible.

Researchers are expected to submit papers for the CES Discussion Paper series and to submit their final published research papers and reports to CES, to maintain a record of the research results and to ensure that the benefits of the research accrue to the Census Bureau's data programs and to future researchers.

National Center for Health Statistics RDC

In 1998 an RDC was established at the National Center for Health Statistics (NCHS). The principle underlying its formation is that under certain rigidly observed restrictions access to detailed data can be provided to qualified researchers with little or no appreciable risk of re-identification of NCHS respondents. Within a secure monitored facility external researchers are allowed access to internal restricted data files for approved projects. Restricted data files are those which contain information not released for unrestricted use, such as detail for smaller geographic areas (e.g., state, county, or lower), or socio-demographic variables such as age, race, or occupation, but do not contain personal identifiers (e.g., name, street address, social security

number). Restricted data files may be used in the RDC by researchers wishing, for example, to control for detailed geographic area in their models. In addition, data supplied by the researcher may be merged onto the NCHS-collected data files for enhanced analyses.

Use of the NCHS RDC is governed by strict policy and procedures. Researchers must first submit a research proposal for approval; no materials may be brought into the RDC, and no materials (data products) may leave the RDC without disclosure review; researchers may be asked to sign a confidentiality affidavit; and the RDC can only be used when staff are available for supervision.

While researchers may include small geographical areas in their analytical design, except in very unusual circumstances, they are not provided actual names, nor are they permitted to see analytical output with the names of small geographic units. A statistical model would employ, say, county level characteristics as an independent variable, but output would be in the form of an overall index or coefficient for the entire statistical domain - not for small areas themselves.

Should a researcher request that an NCHS data file be merged with external data for the geographic areas in which NCHS survey respondents are located, RDC staff conduct the merge and then remove the geographic identifiers leaving the researcher access to a file that consists of the NCHS data merged with the additional data. Should the researcher need clustering variables to stratify on geography, RDC staff construct a set of dummy geographic indicators.

The NCHS RDC has been able to accommodate a large range of users that need hands-on access to the data. A laboratory has been developed that permits researchers to conduct their work on-site. The NCHS RDC has its own computer system with no connections to any other computer system and has been designed with a number of firewalls and fail-safe mechanisms that allows the researchers access to authorized data only. The system does not contain any data that are not being actively used; archived data files and inactive data files are kept on magnetic tape in a secure room that is accessible only by RDC staff. The user workstations have had the floppy disk drives and parallel ports disabled and all printouts are routed to a single printer in another room that can only be accessed by RDC staff.

In the NCHS RDC, many different types of research projects can be addressed, user-supplied data can be merged with NCHS data (although merging is done by NCHS staff), short term as well as long term projects are acceptable, and virtually all NCHS data (without personal identifiers) can be made available. More details are available at http://www.cdc.gov/nchs/r&d/rdc.htm

The Agency for Health Care Research and Quality - Center for Cost and Financing Studies Data Center

The Agency for Health Care Research and Quality opened a data center (the Center for Cost and Financing Studies Data Center, or CCFS-DC) in December 2000. The purpose of the CCFS-DC is to provide access beyond public use files to those researchers using the Medical Expenditure Panel Survey. Within a secure, monitored facility, researchers with approved projects can access non-public use data needed to complete a specific project. In addition to public data, the

data provided at the CCFS-DC generally fall into one of two categories: requests to use data that had not been prepared for public use, but which have no risk of identifying respondents, and/or requests for demographic or geographic data at a level not released. In addition, researcher supplied data can be merged onto MEPS data.

Use of the CCFS-DC is governed by detailed procedures. A researcher must submit a detailed proposal, which is reviewed for both feasibility and consistency with the legislation and regulation that authorized the survey. Once the project is approved, AHRQ constructs for the user an analytic file containing the needed variables. Direct geographical variables (i.e. state or county codes) are typically not provided, but variables that describe characteristics of geographic areas such that no one specific area can be identified are provided. Estimates cannot be produced for sub-national areas. CCFS staff conduct all data mergers requiring access to geographical identifiers.

The CCFS-DC can only be used when a CCFS staff member is on-site, and the facility is staffed at all times when in use. Materials brought into the data center are subject to inspection. No micro data can leave the facility at any time. Tabular output is reviewed before it is removed. Users are asked to sign a confidentiality affidavit as well as a Data Center Use Agreement which describes data center policy and restrictions in detail.

The facility itself is secure. In addition to being staffed when in use, there is 24 hour video surveillance. Entrance is by key card, and users are not provided with a card so they must be admitted by the staff member. The computing facilities include a separate LAN which is completely apart from the general purpose networks at AHRQ. Each Data Center user has a separate account, and has no access to other accounts. Users are given a Xyloc card which functions as a proximity device. The device is "signed out" each day. The workstations contain no floppy disk drives, CD readers, or parallel ports. All print jobs are directed to a single printer in a separate, locked room. The print outs are reviewed by a staff member before being given to the data center user. The print outs are reviewed again before they are removed from the CCFS-DC.

In addition to onsite use, the CCFS-DC provides limited access to programmers who will execute statistical programs for external users. The same data restrictions apply. Users are charged a per-hour fee for these services.

Census, NCHS and AHRQ Data Centers Compared

All three agencies control access to their RDCs, requiring close review of proposed research, formal signed agreements, physical and electronic security, and review of output. There are some differences, however. The Census model is more appropriate for extended stay within the RDC. Moreover, the researcher is required to have a status equivalent to that of a Census employee - subject to the same legal restrictions and sanctions. At NCHS and AHRQ, on the other hand, any properly qualified researcher (who signs a confidentiality pledge and enters into a formal agreement) can be granted access, given that data provided them is not "identifiable" under the circumstances for obtaining them within the RDC. Regardless of their legal status, the security employed by all three systems is quite similar. Duration of stay at the NCHS and

AHRQ stay can be as short as one week and tabular as well as regression analysis and output are permitted.

The NCHS RDC provides data for virtually any NCHS survey, whereas the AHRQ was designed to provide access to its Medical Expenditure Panel Survey. The Census RDC provides access to Title 13 data such as the Current Population Survey, and economic data bases such as the Longitudinal Research Database, the Longitudinal Business Database (LBD), and the Survey of Manufacturing Technology.

Costs are difficult to compare, given the variation in duration of stay at the various RDCs. They tend to be higher at the Census, but depending on the potential to contribute to Census programs, can be lower. Census also provides a subsidy for students.

REMOTE ACCESS AND ON-LINE QUERY SYSTEMS

A few government agencies have remote access systems in operation and others are planned or in development. The National Center for Education Statistics has online query systems (one for each Postsecondary data report) which combines a data base system with a spreadsheet program to allow users to request tabulations and correlation matrices from restricted data files. To avoid the risk of disclosure, the data produced are categorical, all counts are weighted, and estimates are only produced for cells with at least 30 respondents.

In the National Center for Health Statistics, remote access is handled in the Research Data Center. Remote access is governed by strict policy and procedures. Upon approval of a proposal, RDC staff create a data file exactly like the real data - except with fictitious data - for the researcher's use in debugging programs prior to sending them in to be run, thereby reducing the number of iterations on the remote access system. The advantage of the NCHS system of remote access is that it allows the researcher to have the full analytical power of the detail contained in the data set while preventing the disclosure of identifiers into the public domain.

To use the NCHS remote access system, approved researchers submit analytic computer programs written in the SAS language by e-mail, without direct access to the data. SAS was chosen as the analytic language because it is in wide use and is sufficiently well structured that an automated scanning system could be used. For the NCHS remote system, a number of functions available in SAS have been disabled because they can produce unstructured output that can not be readily scanned (reviewed) in an automated system or present an unreasonable risk of disclosure. Disabled functions include PROC TABULATE, PROC IML, PROC PRINT, and LIST.

While the NCHS remote system is designed to operate entirely automatically, systemic manual checks are performed to insure proper functioning. The system scans the e-mail for arriving computer programs; validates the user; scans the program for non-allowable commands; verifies that the program is not trying to access unauthorized data files; and then, if no problems are found, executes the program against the real data. After execution, the output is automatically scanned for disclosure problems. If none are found, the researcher receives their output within a few hours, depending on staff availability. The current remote access system operates by e-mail

but an Internet-based system is under development and testing. The Internet-based system will offer a more user friendly interface and improved turn-around time.

The Census Bureau is developing an online query system as part of its American FactFinder data dissemination system. This will allow researchers to electronically submit requests for tabular data from Census 2000, the 1990 Decennial Census, the American Community Survey, and the 1997 Economic Census. Data files accessed are mainly summary files with matrices of aggregated data. To avoid the risk of disclosure, the tabulations will come from a previously swapped, recoded, and topcoded microdata file. There will be restrictions on levels of geography, number of table dimensions, total populations counts, mean and median cell size, and percentages of cell counts of one. The system has been tested and automated Internet access to tabulations from microdata files with confidentiality protection has proven feasible. The actual development of the full production system is under way. A fuller description of the "query filter", the "results filter", special software employed, and database security is available in Rowland and Zayatz (2001).

RESEARCH FELLOWSHIPS AND POST DOCTORAL PROGRAMS

In these programs, researchers are funded to work at Agency offices. While there, they are able to explore confidential data otherwise available only to the agencies' staffs. Because they are made fully subject to the agencies' laws (along with its sanctions), they become, in every respect, agency employees and conduct research in residence at the agency, use agency data and facilities, and adhere to the same confidentiality agreements as regular employees. Candidates are required to have a recognized research record and considerable expertise in the area of proposed research. A goal of these programs is to bridge the gap between academic scholars and government social science research. Thus, the program enables staff in the federal statistical agencies to interact with renowned experts, developing long-term relationships for future research collaboration.

Criteria for post-doctoral and senior fellowship programs vary between agencies. The post-doctoral research programs at Bureau of Labor Statistics (BLS) and at the Bureau of the Census (BOC) were developed to train recent Ph.D. graduates in survey methodology. A goal is to promote interest in continuing to work in the federal statistical agencies. Post-doctoral researchers at these agencies must have held a Ph.D. (or equivalent) in a relevant field for less than three years or complete the Ph.D. before the commencement of work as a post-doctoral researcher. Post-doctoral research applicants must submit detailed research proposal for evaluation by agency staff.

The American Statistical Association (ASA), (in collaboration with the National Science Foundation (NSF)), administers Fellowship programs in the BLS, BOC, and at the National Center for Education Statistics (NCES). ASA also sponsors research fellowship programs in the National Center for Health Statistics, (NCHS), and in the Bureau of Economic Analysis, (BEA). For ASA Research Fellowships, candidates must submit a detailed research proposal for competitive evaluation by a Program Review Board. Composition of that board may include representatives of ASA, academia, and other statistical organizations. Additional consideration by staff of the sponsoring agency depends on the type of fellowship. Proposals are evaluated on

the applicability of the research to agency programs, the value of the proposed research to science, and the quality of the applicant's research record. Qualified women and members of minority groups are encouraged to apply.

Areas of potential research vary by agency. At BLS, examples include, but are not limited to:

> Statistical Methodology and Computing - sampling frames, time-in-sample effects, non-sampling errors, time-use, computer-assisted interviewing, statistical quality management, cost-error modeling, item imputation, expert systems for data access and use, information dissemination, statistical graphics and data visualization, estimation, time series methods, statistical methods for data analysis, and statistical disclosure methodology.
>
> Economic Measurement and Research - measurement of labor force characteristics, output definitions, incidence of injuries and illnesses, measurement and analysis of non-wage benefits, measurement of economic growth, productivity research, price measurement, and analysis of labor markets.

Senior research fellows and post-doctoral researchers become temporary employees and thus, are subject to the confidentiality pledges that all employees must sign. In NCHS, ASA Research Fellows are hired under an Intergovernmental Personnel Agreement (IPA), with their academic institution. Research Fellows working under an IPA designation are considered to be agents of the federal government and are therefore permitted to work with confidential data.

Salaries are paid by the sponsoring agency. Fellows and most post-doctoral candidates receive salaries commensurate with their qualification and experience. Fringe benefits, travel and relocation costs are negotiable.

Time limitations vary, with fellowship support ranges from 6 months support at BLS to one-year with the potential of one-year extension at NCHS. Post-doctoral research appointments are limited to two years, but can be extended with special considerations.

Citizenship requirements vary by type of fellowship or program.

LICENSING AGREEMENTS

For many researchers the most desirable arrangement is one in which they are allowed to access confidential data at their own institution thereby avoiding conducting research under unfamiliar if not uncomfortable conditions. This section describes the key features of licensing agreements that permit such access developed in recent years by U.S. government agencies and research organizations. In some cases (e.g. the National Center for Education Statistics), the agencies' organic legislation provides for such agreements. In others, special arrangements have been developed. The research upon which this discussion is based utilized information from six

federal agencies and three social science research organizations[2]. More detailed information on these programs is available in Massell (1999) and Massell and Zayatz (2000).

Researchers obtain access to restricted access data by signing an agreement or license. Such agreements require:

- a demonstrated need for sensitive data;
- authorization for all users at the requesting institution;
- signature by a senior level official and key staff;
- a data security plan;
- agreement by researchers not to identify individual research subjects or to link data received with other microdata files; and
- review of all statistical output before publication.

The license is for is a specified period of time and data files must be returned or destroyed. Some licensors require fees and/or approval by an institutional review board.

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.

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.

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

[2] The U.S. government agencies are the National Center for Education Statistics (NCES), the National Science Foundation (NSF), the Department of Housing and Urban Development (HUD), the Health Care Financing Administration (HCFA), the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). The social science research organizations are the Inter-university Consortium for Political and Social Research (ICPSR), the Survey Research Center at the University of Michigan (SRC-UMICH), and the University of North Carolina, The Carolina Population Center.

data (e.g., Privacy Act of 1974).

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).

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.

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).

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

Other requirements: Some licensors require user fees. The licensee must cover the cost of creating and maintaining a secure data handling environment. Others licensors require prior approval of the research plan by an institutional review board.

DISCUSSION AND CONCLUSIONS

The restricted access procedures described in this document have been developed to meet a variety of needs under differing legal and organizational circumstances, making comparison difficult. Still, they represent valuable experience and some suggestions may be offered as to the range of needs they satisfy and the relative ease of access they afford.

For the researcher, some approaches provide distinct advantages over others when judged against criteria for ideal access such as:

- convenient and speedy approval of research proposal;
- access to the maximum amount of detail;
- ability to perform a variety of statistical procedures;
- access within a research setting most conducive to productive research;
- availability to a wide variety of researcher; and
- reasonable cost.[3]

At the same time, data providers confront their own requirements, that include:

- assuring that pledges of confidentiality made to respondents are strictly observed;
- arranging for the dissemination of data on the widest possible basis;

[3] It has not been possible to effectively compare the relative costs of these approaches. That is outside the scope of this paper. The reader is invited to contact the appropriate agencies and their web sites for information of this sort.

- providing data access to researchers qualified to make maximum analytical contributions; and
- availability of the appropriate legal framework, properly trained staff, and equipment.

Considering these different approaches first from the standpoint of the researcher, the licensing approach offers the opportunity to conduct research using a variety of analytical techniques in familiar surroundings with all of the customary research resources readily available. The detail available in this type of arrangement is not uniform, with varying degrees of disclosure limitation measures applied to the data. Not surprisingly, because the data accessed are to some degree identifiable, the application process is thorough and selective, open to a restricted class of researchers and involves an enforceable and highly proscriptive legal document, provisions for site inspection and, in some cases, the payment of fees.

In contrast, RDCs permit access to more detailed data, with somewhat fewer restrictions as to the type of researcher (for NCHS and AHRQ) or approval by other entities (e.g. approval by institutional review boards, IRBs). In this case, data protection is primarily the responsibility of the RDC rather than the researcher, access is not as long as with licensing, is available to fewer people per project and limited to a site designated by the agency. A number of factors (data preparation, researcher unfamiliarity with data files, misunderstandings as to what specific data can be accessed) can make for delay in obtaining access through the RDC. However, what is lost in convenience and stimulation of the "home" research setting can be counterbalanced by the richness of the data and the possibility for contact with agency researchers.

Participants in post-doctoral and senior fellowship programs represent perhaps the most favorable mechanism for researchers, for they can access the same data as most other agency researchers. Their working conditions are less restrictive than those of RDCs, and there are numerous possibilities for interaction with agency colleagues (as well as others in the Washington area). They are not limited in statistical techniques they may employ, but research topics must be consistent with agency priorities. The latter requirement is not a major obstacle, however, as these priorities are defined in relatively broad terms. The principal drawback to gaining data access through this mechanism is, of course, the limited number of fellowships available.

From the standpoint of data providers, with respect to both breadth of data dissemination and confidentiality, those agencies with the proper legislation and infrastructure, may be favorably disposed to licensing for it permits making relatively detailed data available to a significant number of researchers in various research centers. Data protection is chiefly the responsibility of the researcher and is addressed by legal and administrative means (on-site inspections, possible loss of future data access, security requirements).

Senior and post-doctoral Fellows are legally and administratively indistinguishable from agency employees, but are necessarily limited in number. Nonetheless, these programs are important, not only for the high quality research that results, but because of the strong and enduring ties that may be forged between agencies and the larger research community.

Somewhere between licensees and Fellows are the research data centers which require the agency to make a relatively costly investment in a properly equipped research site staffed by experienced agency staff who are researchers in their own right. Those willing to make this investment can, however, make relatively rich data sets (often more detailed than those available to licensees) available, albeit under less than ideal - but nonetheless accommodating - research conditions. For on-site RDCs, geographic proximity and the availability of sufficient travel resources heavily influence those who can take advantage of their facilities. Though these circumstances impose severe restrictions on some, it must be recognized that for extremely sensitive data sets (e.g. data for establishments, genetic data, small area studies) the RDC may be the *only* way that agencies can permit outsider to access those data.

An exciting development in restricted access has been the development of remote access systems and on-line query systems. Whereas the NCHS remote access system requires that researchers have the capability to write and de-bug their own computer code, on-line query systems do not. The latter also have the advantage of not requiring advance approval of an application form. In the case of the American FactFinder, for example, the system itself carries out all the necessary data preparation and statistical manipulation. The strength of the on-line system is that disclosure risk has already been addressed in the preparation of the data on which it is based, insuring that no tabulation available to the user represents a disclosure.

The user who is dissatisfied with a particular tabulation used in Census publications, can design his own tabulation using this system. While more sophisticated statistical analyses are not yet possible with the Census system, the NCES system does permit the calculation of statistics needed for the interpretation of complex surveys as well as correlation matrices. In the case of the NCHS remote access system, there are few limits on the statistical techniques that may be applied. From the standpoint of the researcher, these systems have the advantage of permitting access from the researcher's desk top, without the need for the development of a research proposal, etc. and at minimal cost. The providing agency benefits from more direct control over confidentiality and the dissemination of its data to a wide audience.

In the near future there is reason to expect that remote or on-line systems may be developed further. Already work is underway (Karr, et al., forthcoming) on a system that would to provide access to data with varying degrees of detail to users with differing needs and confidentiality clearance. Such systems would provide for one subcomponent with the maximum amount of detail that would be available only to agency staff and its contractors. A second subcomponent would provide less detail to research collaborators and other authorized persons, while a third (the most general) system would be accessible by any member of the public.

As we have seen, few of the institutions mentioned here rely exclusively on a single means of providing restricted access to their data. The methods described fit a variety of needs and circumstances and each must be judged in terms of the situation it was designed to address. The reader is directed to the references cited below and to the agencies mentioned for further information.

REFERENCES AND CONTACTS

General

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

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

Committee for Data Access and Confidentiality http://www.fcsm.gov/cdac/
Chair 2002-2003 Jake Bournazian, Jacob.Bournazian@eia.doe.gov

Licensing

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.
Massell, Paul B., Laura Zayatz, (2000), "Data Licensing Agreements at U.S. Government Agencies and Research Organizations," *Proceedings of ICES-II (International Conference on Establishment Surveys).*

National Center for Education Statistics, "Restricted Use Data Procedures Manual." http://nces.ed.gov/statprog/rudman/index.asp

U.S. Census Bureau Statistical Research Division:
Paul B. Massell, paul.b.massell@census.gov, (301) 457-4954
Laura Zayatz, laura.zayatz@census.gov (301) 457-4955.

University of Michigan, Inter-university Consortium for Political and Social Research: http://www.icpsr.umich.edu/NACJD/Private/private.html
National Center for Education Statistics: http://nces.ed.gov/statprog/confid6.asp
National Science Foundation: http://www.nsf.gov/sbe/srs/srsdata.htm#MICRODATA
University of North Carolina, Carolina Population Center, National Longitudinal Study of Adolescent Health: http://www.cpc.unc.edu/addhealth/datasets.html.

Research Data Centers

Beyene, Negasi, John Horm, and Deanna Dick, "The National Center for Health Statistics Research Data Center: New Research Opportunities.

Reznek, Arnold, "Increasing Access to Longitudinal Business Survey Microdata: the Census Bureau's Research Data Center Program." Proceedings of the International Conference on

Establishment Surveys - II: Survey Methods for Businesses, Farms, and Institutions. Buffalo, New York, June 17-21, 2000. (November 26, 2001).

Reznek, Arnold., Joyce Cooper, and J. Bradford Jensen. "Increasing Access to Longitudinal Survey Microdata: the Census Bureau's Research Data Center Program." *American Statistical Association 1997 Proceedings of the Section on Government Statistics and Section on Social Statistics.* Alexandria, VA, 1997, pp. 243-248.

Reznek, Arnold and Alfred R. Nucci, "Protecting Confidential Data at Restricted Access Sites: Census Bureau Research Data Centers." *Of Significance.* December 2000. (With Alfred R. Nucci).

U.S. Bureau of the Census:
Arnie Reznek, rezne001@ces.census.gov, 301-457-1235
Laura Zayatz, laura.zayatz@ccmail.census.gov, 301-457-4955
National Center for Health Statistics:
Ken Harris, kwh1@cdc.gov, 301 458 4262.

On-line Query Systems

Rowland, Sandra and Laura Zayatz "Automating Access with Confidentiality Protection. The American Factfinder." Proceedings of the Section on Government Statistics of the 2001 Joint Statistical Meetings (forthcoming).

Karr, Alan F., Jaeyong Lee, Ashish P. Sanil, Joel Hernandez, Sousan Karimi and Karen Litwin "Web-Based Systems that Disseminate Information from Data but Protect Confidentiality", in William McIver and Ahmed Elmagarmid (Eds), *Advances in Digital Government: Technology, Human Factors, and Policy*, Kluwer Academic Publishers, Amsterdam (Forthcoming)

Fellowships and Post Doctoral Programs

Bureau of Labor Statistics: Steve Cohen, Cohen_Steve@bls.gov (202) 691-7400

National Center for Health Statistics:
Research Fellowships: Jacqueline Smith, , (301) 458-4512
Post Doctoral Programs: Dr. Lester Curtin, , (301) 458-4040

Bureau of the Census: www.census.gov/srd/www/fellweb.html
National Center for Education Statistics: Marilyn McMillen Seastrom, marilyn.mcmillen@ed.gov, (202) 502-7303
Bureau of Economic Analysis: www.bea.doc.gov/jobs/rsch.htm

(as of April 4, 2002)