An Examination of Monitored, Remote Microdata Access Systems

Sandra Rowland

Prepared under contract to the
Committee on National Statistics
National Academy of Sciences (NAS)

Presented at the NAS Workshop on
Access to Research Data: Assessing Risks and Opportunities

October 16-17, 2003

An Examination of Monitored, Remote Microdata Access Systems

1. Introduction

Types of Microdata Access
Sampling of Systems for Monitored Remote Access to Restricted Microdata
Methodologies Commonly Employed in Monitored Remote Access Systems
Usage of Monitored Remote Access Systems

2. Monitored Remote Access in Foreign Agencies

Luxembourg Income Study
Statistics Canada
Statistics Denmark
Statistics Netherlands
Australian Bureau of Statistics
Statistics Sweden

3. Monitored Remote Access in US Federal Agencies

National Center for Health Statistics
National Center for Education Statistics
Census Bureau

4. Sample of Research Projects in US

Digital Government
Cornell Retricted Access Data Center
Integrated Public Use Microdata Series - International
Internet 2 and Next Generation Internet
USGenWeb

5. Conclusions

Appendices

Appendix 1A: Monitored Remote Access System Methodology – U.S. Federal Agencies
Appendix 1B: Usage of Monitored Remote Access Systems – U.S. Federal Agencies
Appendix 2A: Monitored Remote Access System Methodology – Foreign Agencies
Appendix 2B: Usage of Monitored Remote Access Systems – Foreign Agencies

An Examination of Monitored, Remote Microdata Access Systems

Sandra Rowland
Prepared under contract to the
Committee on National Statistics, National Academy of Sciences

October 2003

1. Introduction

Many national statistical offices (NSOs) disseminate microdata in three ways: public use microdata files on CDROM or on-line, research centers or licensed sites, and remote access to restricted microdata. This paper covers a sampling of systems in NSOs that permit monitored remote access to restricted microdata. The sample includes six foreign systems and three systems in the United States (US). Section 1 of the paper covers foreign systems in the following order: Luxembourg Income Study, Statistics Canada, Statistics Denmark, Statistics Netherlands, Australian Bureau of Statistics, and Statistics Sweden. Section 2 covers US systems in federal agencies in the following order: National Center for Health Statistics, National Center for Education Statistics and Census Bureau. The type of methodology employed in each of the systems is reviewed for each country because the methodology influences the kinds of access and results given to users. The usage of each system and the kinds of research that have benefited from their use are reviewed for each country, in so far as such information is available. A few research programs in the US bearing on remote access to restricted microdata are also briefly reviewed in Section 3 because they are examples of the kinds of research that contribute to practical applications in the public sector. Appendices 1A and 1B have a summary of methodology and usage of US systems and Appendices 2A and 2B have a summary of methodology and usage of foreign systems.

Types of Microdata Dissemination

Public use microdata files contain data from surveys and subsamples from censuses and are usually edited through perturbation of data, addition of random noise, top and bottom coding, rounding, variable suppression, and adding, removing and swapping records (Horm 1999). Public use files are usually available on CDROM and in some cases are available on the Internet. The Federated Electronic Research, Review, Extraction and Tabulation Tool (now known as Data Ferrett) is one example of the latter. It makes public use files from several US federal agencies available on the web.

Research data centers (RDC) were created to allow users to access restricted microdata files that are not available on CDROM. The files may have confidentiality edits but the detail on the files is much greater than is permitted in a public use microdata file. Users are required to submit a research proposal to the NSO that maintains the research center and if approved must carry out their work at the center. In some cases NSOs will license and inspect or monitor user sites to obtain and use restricted microdata sets for approved purposes (Seastrom and Kaufman 2003).

Remote access systems make it possible for users to analyze restricted microdata without visiting an RDC. The systems used for remote access to restricted microdata are monitored automatically and/or manually for disclosure avoidance. They employ automated and manual filters that block certain kinds of queries and results. The files available are usually edited for disclosure avoidance using the same techniques as those used for public use files. They provide more detail to researchers than public use files, but less detail than is usually available in an RDC. The files reside in the NSO and extracts of microdata and direct access to the records are not permitted.

Sampling of Systems for Monitored Remote Access to Restricted Microdata

The Luxembourg Income Study is reviewed first because it is the oldest of the programs that give users remote access to restricted microdata. It began in 1983 and utilizes the LISSY remote access system. The system has served as a model for many other systems (consciously or unconsciously) currently in use and under development abroad and in the US. Canada and Denmark have given users remote access to restricted microdata since 2001. The Netherlands, Sweden and Australia began pilots in the use of remote access systems in 2002 and 2003. In the US, the National Center for Education Statistics (NCES) and the National Center for Health Statistics (NCHS) gave users remote access to restricted microdata beginning in 1997 and 1998, respectively. The Census Bureau began disseminating Census 2000 microdata in 2003, after pilot tests that took place in 2002.

Methodologies Commonly Employed in Monitored Remote Access Systems

There are two common methodologies among the systems reviewed in the paper. One type usually consists of an email interface that allows uses to send programs as part of the body of the email or in an attachment. These systems usually accept standard statistical programs such as SAS, SPSS, STATA and GAUSS, chosen because they are commonly used by researchers and lend themselves to automated review of input programs and statistical results. These systems return SAS, SPSS, STATA and GAUSS results and may prohibit or modify certain commands thus limiting the kinds of outputs that users may want. The email systems are used in all of the foreign applications and by NCHS. Such systems have been referred to as "remote job execution systems" (Schouten and Cigrang 2003, p. 6). Processing is usually done in batch mode rather than on line, so the systems may also be referred to as off-line. Results are returned within minutes or days depending on the size of the program and the degree of manual intervention.

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

There are several aspects of the methodology employed that vary by system (Refer to Appendices 1A and 2A). Many of these aspects are reviewed in the paper for each system depending on the information available. Methodologies employed may include:

confidentiality edits to the base files accessed that usually involve adding noise to the data to reduce the possibility of disclosure,

electronic authorization of users that requires the use of user identification and passwords to gain access to the system,

email or web user interfaces that provide a facility to the user to communicate what they want from the system,

standard statistical programs or custom applications for processing that use statistical software packages or custom programs to process the user request,

query filters to examine requests and block the user from requesting certain results prohibited by the NSO,

results filters to examine results and block any result prohibited by the NSO,

automated and manual intervention for disclosure avoidance at the query submission or output stages that are totally automated or involve the review of input and/or output by statisticians, and

usage logs for disclosure avoidance review that are accumulated and used by NSOs to determine if their rules are adequate for disclosure avoidance and to detect possible risks to confidentiality.

Another important methodological aspect of a remote system is automating complementary disclosure review to prevent the possible disclosure of restricted data that could result from combining multiple outputs. Although research in this area has been undertaken (Duncan, Roehrig, and Kannan 2000) none of the systems reviewed contain mechanisms to prevent complementary disclosure. This may be due to the difficulty and expense of automating such procedures.

Researchers and NSOs are also interested in disseminating data from smaller geographic areas while preventing disclosure of restricted data. Research on automating aggregation of low level geographic areas and/or allowing user-defined geographic areas has been carried out using real data (Karr and Sanil 2001). None of the systems reviewed contain mechanisms to disseminate data for user-defined areas.

Usage of Monitored Remote Access Systems

Most of the users of monitored remote access systems are researchers and public sector staff. The systems are sometimes limited to official users but most often are not. Most

statistical offices require user registration and some must officially accept a research proposal before the system can be accessed. Some of the foreign systems reviewed have little information on usage because the systems are new or in the pilot stage. Others such as those used in Statistics Canada and the Luxembourg Income Study have detailed information. The Census Bureau has statistics on the usage and a little information on customer satisfaction. The NCES has a customer satisfaction survey with some usage statistics and the NCHS has usage statistics.

There are several aspects of system usage covered in the paper depending on the information available (Refer to Appendices 1B and 2B):

permission or authorization required by the NSO to access the systems that involves signing a research contract and confidentiality agreement or some kind of registration,

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

files available through the systems ranging from one or two files to many, as well as combining files for use and permitting user files,

documents or metadata available online varying from detailed user guides to tailored emails,

assistance available including automatic feedback and help desks and user workshops,

turnaround time for results ranging from seconds to days,

hours of availability ranging from 24 hours a day, 7 days a week (24/7) to office hours only (8/5),

cost ranging from zero to membership fees to fee for service time, and

benefits derived from use including reports and policy implications.

2. Monitored Remote Access in Foreign Agencies

In recent years the European Advisory Committee on Statistical Information in the Economic and Social Spheres (CEIES) has encouraged research in the dissemination of microdata because, "significant research can only be undertaken with microdata" (CEIES, 2002). The committee also notes that, "we are now moving to a situation where there are technological solutions which can produce a 'virtual safe setting' over the Internet… an area that needs to be explored as a cheaper and much preferred alternative

to [physical] safe settings". It recommended that Eurostat establish the feasibility of a virtual safe setting (CEIES 2002 pp. 2 and 3).

CEIES holds periodic conferences on dissemination of microdata and confidentiality that encourage cooperation. Knowledge and research are shared among European and other countries such as Canada, Australia and the US that contribute to efforts to build and utilize remote access systems. Given encouragement by Eurostat and the example of the Luxembourg Income Study, several European countries, Canada and Australia have begun to use or test remote access to restricted microdata.

Luxembourg Income Study (LIS)

The Luxembourg Income Study is an international program that makes microdata from 66 household income surveys available for research from the 25 countries that participate in the program. The files may be restricted use or public use depending on the country. The Current Population Survey is used for the US. The program began in 1983 and utilizes the Luxembourg Income Study System (LISSY) to disseminate data. The LISSY is an independent system that can be and is being used by other programs. The programs include the Luxembourg Employment Study (LES), the DIW (Deutsche Institute fur Wirtschaft) in Berlin for the German household panel, and the EUROSTAT/London School of Economics for remote access to EUROSTAT data (Cigrang 2003). This section of the paper covers the use of LISSY in the LIS.

LISSY was designed to handle statistical software programs thus allowing users to write their own programs. Today LISSY can accept SAS, SPSS and STATA programs. The LISSY began accepting and returning programs from users on diskette in 1983. In 1987 it used email on the European Academic Research Network (EARN/BITNET) allowing users with access to EARN/BITNET to email their programs to the system. Email on the Internet was used in LISSY version 4.

The heart of the LISSY is the mail router (Post Office). The Post Office

> "retrieves the email requests from the mail server,
>
> prepares these requests for processing by checking for all security issues like clearly identifying a user, checking for the use of illegal statistical commands, and checking for the usage of sequences of commands or variables or any other combinations not allowed,
>
> returns any job that breaches security to the sender along with an error message explaining the violation,
>
> distributes the requests to the batch processor computers,
>
> examines the output file size and contents for output acceptable to security rules,

returns acceptable statistical results to the proper (registered) user email addresses,

sends suspicious output to the review queue for manual review instead of returning, results to the user, and

maintains critical databases needed for the overall operation" (Cigrang and Schouten 2003, p. 8).

LISSY can use public use files and restricted files. It accesses the standard geographic areas that are defined in the program files ranging from national to subnational areas and has no facility for generating user-defined areas. Although it generates usage logs, it does not automatically check multiple outputs for complementary disclosure. Cigrang and Schouten examined some research on automatic checking for complementary disclosure and found that "complete evaluation of multiple queries may be too complex, time consuming or restrictive to implement" (Cigrang and Schouten 2003, p.11).

The advantage of LISSY is that it allows users to submit their own programs using familiar well-known statistical packages. Users can be supplied with synthetic or dummy files to test the syntax of their programs. They may obtain every type of analysis that the packages render within the bounds of the security rules defined for the program. For example, in LIS all types of analysis are acceptable except extracts. However, the LIS query filter does not permit certain commands, word sequences and variables.

In LIS, users must submit a research proposal and sign a contract as well as a confidentiality pledge. Users are primarily academic researchers. There is no cost to the user, but only users from participating countries that pay an annual fee may use the data. The program has income files from 25 countries. User files are not accommodated. There is available documentation for key variables that have been made as comparable as possible by the program. LIS sponsors workshops and has a full-time person on a help desk to assist users. LISSY output can be returned within minutes depending on the size of the programs submitted, the number of programs being submitted at the same time and the number of servers on line. The system is available 24 hours a day, 7 days a week.

The LIS has a 20-year history of data analysis. Recent statistics show that from January 1, 2001 through June 30, 2003, 213 users submitted 36,280 programs on average per year. The highest usage was by the US, the UK and Germany. US researchers alone submitted 10,047 programs or 28 percent of the average number of programs per year, during that period. The number of LIS working papers published per year has grown from around six in 1985 to over 45 in 2002 for a total of approximately 350 working papers (LIS 2003).

Four fifths of the papers written are on inequality of income and poverty. "There was a slight preponderance of poverty issues in the early 90s and of inequality issues before and thereafter" (Forster and Vieminckx 2003, slide 2). Groups at risk are a main concern and have made up around 30 percent of the studies since the inception of the LIS. There is an increasing focus on families and children at risk compared with studies of the elderly

since the early 1990s. Analysis of comparative trends among countries made up just under 30 percent of the topics studied in the late 1980s but dropped to under 20 percent in the early 2000s. The LIS has contributed to four major fields of study in its 20 year history:

> "refinement of the income concept,
>
> proliferation of equivalence scales,
>
> conceptualization and measurement of income inequality and poverty, and
>
> proper identification of international rankings and trends" (Forster and Vieminckx 2003, slide 7).

Statistics Canada

Statistics Canada began to consider the use of remote access for research on restricted microdata in the 1990s because researchers had complained about the lack of detail in the Public Use Microdata Files (PUMF) and the inability to produce exact variances for their analyses. "Remote access entails researchers e-mailing their analytical programs to Statistics Canada where they are run on survey master files residing within the Agency's secure internal network. Researchers do not have direct access to the confidential survey microdata. The program outputs are vetted for confidentiality before being e-mailed back to the researchers" (Tambay, Goldman and Potter 2003, p.7).

Statistics Canada began offering remote access to a small number of surveys, some of which do not produce a PUMF. Disclosure control varies by survey and whether or not a PUMF is available for a survey. If a PUMF is available, unit-level data, minimum and maximum values, location of sample units or clusters and anecdotal information about respondents are not released. Statistics and cell values in tables must be based on a minimum number of cases. Guidelines for the creation of synthetic files given to users for testing have been established.

Researchers are required to contact Statistics Canada to create the files for use. Registered users may submit their programs through email using software or files such as SAS, SPSS, STATA, Foxpro and ASCII. The programs and files users may submit vary according to the survey data being offered. For some surveys, researchers are provided with documentation on the description of the survey, record layout and a dummy (synthetic) file for testing programs. For others they may submit requests without knowing the internal structure of the database (Tambay, Goldman and Potter 2003).

The actual process of receiving programs and vetting results is fairly manual. Statistics Canada executes the programs and reviews outputs for disclosure avoidance before they are returned to the user in one to two working days. The work is done during normal working hours.

The longitudinal National Population Health Survey (NPHS) has the highest use of remote access to date. The NPHS covers health status, use of health services, risk factors and demographic and socio-economic status. "In 2001 and 2002 the number of programs received averaged 99 and 40 per month, respectively" (Tambay, Goldman and Potter 2003, p. 8). Statistic Canada provides dummy files with no analytical use for researchers to test their programs. Workshops are provided for users interested in using the data and macros are provided for variance estimation. There is a fee for remote access to the NPHS microdata.

According to a presentation on NPHS research findings, there were 242 articles from NPHS research in 91 journals. Fifty-nine grants based on the research were identified Hamilton and Humphrey (2002). The presentation did not distinguish among the media used to access the data that includes published reports and public use microdata files in addition to remote access. Important areas of research include cancer and diabetes prevalence and utilization of health services.

The techniques used in the Survey of Labour and Income Dynamics (SLID) are different from those used in the (NPHS). The SLID is the "first Canadian household survey to provide national data on the fluctuations in income that a typical family or individual experiences over time which gives greater insight on the nature and extent of poverty in Canada" (http://www.statcan.ca/english/sdds/3889.htm). The SLID data retrieval system permits users to create results from a single screen without knowing the structure of the database. They may select variables and create both longitudinal and cross-sectional data sets that can be used by their preferred analytical software. Between May 2002 and July 2003, 160 requests for SLID had been submitted remotely.

The principal research areas of the SLID are:

Employment and unemployment dynamics
Life-cycle labor market transitions
Job quality
Family economic mobility
Dynamics of low income
Life events and family changes
Educational advancement and combining work and school.
(http://www.ciqss.umontreal.ca/Documents/acetat_e_2002. ppt.)

Statistics Canada has cooperated with universities to make researchers aware of what is available from both PUMFs and remote access to restricted microdata. The Data Liberation Initiative (DLI) improved access to Canadian data in the universities. The British Columbia Interuniversity Research Data Centre is an example of a university center that provides assistance to researchers. University libraries in Quebec pooled their resources to render PUMFs more usable through the Sherlock system. "SHERLOCK was developed mainly for members of the Quebec academic community to enable them to access and utilize the survey microdata of the DLI and the University consortium for Political and Social Research (ICPSR)" (Drolet 1999, p. 15). Although this cooperation

facilitates data usage, the restricted access microdata files are kept in Statistics Canada and researchers must apply to Statistics Canada for remote use of restricted access microdata and pay a fee.

Statistics Denmark

Statistics Denmark began allowing remote access in March 2001 after completion of a successful pilot that began the year before (Anderson and Thygesen 2003) . The experience using remote access from its inception to date has yielded good results with no breaches of confidentiality. Therefore, Statistics Denmark will eventually replace on-site access to restricted microdata with remote access.

Users of the remote system may submit SPSS, STATA and GAUSS programs and may work with the data freely creating new datasets from the original datasets. All data processing is done at Statistics Denmark and researchers may not download or print datasets or data extracts. Users communicate with the system through the Internet and outputs are returned using email. Output is examined by Statistics Denmark staff and must be aggregated enough to avoid disclosure of information on individuals and enterprises (Anderson 2003).

Remote access is granted by Statistics Denmark only to authorized institutional environments granted on a need-to-know basis for specific projects. Government ministries, research institutions, universities and nongovernmental organizations in Denmark are the types of environments approved. Access is not given to individuals and foreigners may have access only if residing temporarily in an authorized institution in Denmark. From March 2001 to March 2003, Statistics Denmark has given forty-three authorizations.

Most data files made available to researchers are register-based samples that cover labor market research, sociology, epidemiology and business economics. Statistics Denmark has created a number of research databases that link information from several individual registers because users often need linked information. The databases include the Demographic Database, the Fertility Database, the Prevention Register (health data), the Social Research Register and others. The most popular database is the Integrated Database for Labor Market Research developed over a span of 9-10 years. Research institutions may also pay for the creation of databases (Anderson 2003).

Statistics Netherlands

Statistics Netherlands initiated a pilot, with the Dutch Ministry of Social Affairs and Employment as the principal user, to evaluate the feasibility of a remote access facility in 2002. Based on the pilot, the system was made available to all Dutch government ministries in 2003.

During the pilot, the Ministry of Social Affairs and Employment submitted queries by email in SPSS. The Ministry was given access to a microdata file with over a million

records with information on social allowances from 1997-2000. A sample of the data was prepared for users to test their program syntax and become familiar with the variables available. All SPSS commands were accepted but extracts of individual records were not permitted.

The pilot was purposely simple. E-mails were received and acknowledged by phone. Statistics Netherlands processed the SPSS programs and reviewed all output manually to determine the kinds of queries that were submitted and how they could be controlled for disclosure limitation. Output was returned to the users by email. Findings from the pilot will be used to implement automated filters, with emphasis on output filters. Disclosure rules and filters may be developed based on the types of variables requested, the sensitivity of certain variables, the possibility for identifying subpopulations, and the most recurrent types of analyses (Schouten and Jonker 2003).

The development of the output filter will be gradual starting with simple frequencies and contingency tables. Dutch legislation and Statistics Netherlands policies are being reviewed to determine what output is not allowed (Schouten 2003).

Statistics Netherlands may extend access to the system to nongovernmental researchers. It hopes to implement other statistical software in addition to SPSS, to automate query and results filters and to construct log files to evaluate disclosure.

Australian Bureau of Statistics (ABS)

The Australian Bureau of Statistics divides the means used to disseminate microdata into eight categories. Although some of these categories provide "safe data", the categorization recognizes that microdata is the source for all statistical output. The eight categories also provide an excellent framework for NSOs to consider the options for disseminating microdata.

According to Dennis Trewin, the Australian Statistician, the eight categories are:

> "1. Standard Statistical Outputs: The release of statistical outputs, usually in the form of tables, in printed and/or electronic form…
>
> 2. Datacubes: The release of detailed statistical matrices that have already been confidentialised. It is a more appropriate form of release when confidentiality protection can be automated, particularly for small cells (eg. population census)…
>
> 3. Special Data Services: The release of statistical outputs, not necessarily tables, at the request of researchers…
>
> 4. Confidentialized Unit Record Files (CURFS): The release of microdata files on a CDROM which have been amended so that the identification of an individual person or organisation is unlikely…

5. Remote Access Data Laboratory (RADL): Running jobs submitted by authorised users via the internet against CURFs held at the ABS, and returning analysis results after largely automated confidentiality checks…

6. ABS Site Data Laboratory: Similar to RADL except that no downloading of unit record data is available (this is possible in RADL for up to 30 records to support outlier detection, etc)…

7. Collaboration: Working collaboratively with a researcher to produce an output (often a published output) of relevance to the ABS…

8. In-house Analysis: - The ABS can engage persons as 'officers' if they are undertaking functions to support the ABS in its activities. In these situations they can access unit record data although subject to the same secrecy provisions of other ABS officers…" (Trewin 2003, quotes throughout paper).

The key areas of future development for ABS dissemination of microdata are 1, 5 and 7. Category 5- the RADL became available in April 2003 after it was developed by a special project team. The system will be modified and more files made available over time. The system will permit users to submit SAS and SPSS programs through e-mail. Output from the system will be reviewed by the ABS for disclosure avoidance and automatic triggers will be used to identify output that requires more thorough inspection. Downloading of unit data is possible up to 30 records to support outlier detection. Usage logs for confidentiality review will be kept. There will be sanctions against offenders. The ABS will encourage use of RADL when users want linked files and the data matching risk is present. The Information Services Division of ABS will maintain the system. Because the system is new, the ABS provided no information on users and usage of the system.

Statistics Sweden

Statistics Sweden currently does not have ongoing access to microdata. However, it is exploring the feasibility of implementing a system similar to that of Statistics Denmark. In the feasibility study, users are able to submit programs by email and obtain the results by email. The results must be in the form of tables (Nordback 2003). The user logs in from a predefined domain using an encrypted name and password. Access to data is subject to confidentiality procedures and clearance by Statistics Sweden. The feasibility study tested the CITRIX system in combination with RSA software and boxes. Users taking part in the study are researchers, users of regional statistics and users from other public authorities (Hjelm 2003).

3. Monitored Remote Access Systems in US Federal Agencies

Eight US federal agencies were contacted to determine if they currently have systems for monitored remote access to restricted microdata for external users:

Department of Agricultural, National Agricultural Statistical Service and the Economic Research Service
Department of Commerce, Census Bureau
Department of Education, National Center for Education Statistics
Department of Energy, Energy Information Administration
Department of Health and Human Services, National Center for Health Statistics
Department of Labor, Bureau of Labor Statistics
Department of Justice
Department of Transportation, Bureau of Transportation Statistics

Among these eight agencies, three had implemented monitored remote access systems for external users:

National Center for Health Statistics (NCHS)
National Center for Education Statistics (NCES)
Census Bureau (CB)

National Center for Health Statistics (NCHS)

The Analytical Data Research by Email (ANDRE) system provides remote access to virtually all of the surveys sponsored by the NCHS. The Research Data Center (RDC) and the ANDRE were created in 1998 to serve data users who need data with smaller geographic areas (eg. State, county or lower) and other detail not available in the public use files. The effort was spurred on by funding to provide users with contextual data and small geographic areas with direct identifiers removed from the National Survey of Family Growth (Horm 1999). Other important surveys that can be accessed by ANDRE are the National Health Interview Survey and the National Health and Nutrition Examination Survey.

ANDRE allows users to submit SAS programs by email. SAS was chosen because it is widely used by researchers and it lends itself to review by an automated scanning process. Automatic scanning is used as a query filter to examine the input files and suppress or modify certain SAS commands for disclosure avoidance and for ease of automatic output scanning. Commands such as ADD, PRINT, OBS are suppressed and commands such as PROC MEANS, N MEAN STD are modified. Automatic scanning of SAS output wipes out extreme values and suppresses complete output lines with sample sizes less than the minimum standard value (Gambhir and Harris 2003). Although ANDRE is totally automated, questionable output identified in the automated output scan is routed to an RDC staff person for manual resolution. In addition all of the users' data requests, log files and results are maintained in electronic form.

Users of ANDRE must provide a research proposal to the NCHS RDC. If approved the user must sign a research affidavit of confidentiality. Approved users must submit a user identification and password to access the system. The researchers may request data from

multiple files or have NCHS data merged with their own data files. "In general each dataset is specifically prepared for the user. Such a dataset may include many variables selected from multiple internal data files of NCHS. User supplied data may also be merged. The user owns the dataset prepared for him/her and RDC serves as custodian to the dataset. No user is allowed to access the dataset of any other user" (Gambhir 2003).

All of the microdata files accessed by ANDRE have undergone confidentiality editing. Dummy files are sometimes created so the user can refine the SAS input files and documents are emailed to the user explaining the system. Personal assistance is also provided if required. The system is available 24 hours a day so users may submit requests, however, output is returned during working hours. The results are returned to the user within a few hours. The fee for use of ANDRE is $500 per month, making its use less expensive than visiting the RDC.

ANDRE has had 45 users and has run 10,000 SAS programs in the last 5 years. The most popular file requested is the National Family Growth Survey. "The main purpose of the 1973-1995 surveys was to provide reliable national data on marriage, divorce, contraception, infertility, and the health of women and infants in the United States. More than 250 studies in academic journals and NCHS reports have been published using NSFG data. Topics covered by researchers include fertility, family formation, marriage, cohabitation, divorce, contraception, sterilization, unintended pregnancy, HIV/STD and risk behavior, infertility, health, and health services. The National Survey of Family Growth was conducted again in 2002 and 2003 (Surveys of Men and Women, 2002). The interviews include questions on schooling, work, marriage and divorce, having and raising children (including contraceptive use, infertility, and parenting), and related medical care. The first statistical reports and public use data files and documentation should be available in 2004" (http://www.cdc.gov/nchs/nsfg.htm).

National Center for Education Statistics (NCES)

The Data Analysis System (DAS) provides remote access to Department of Education survey data. The DAS was developed in order to increase access to postsecondary data without having to license each and every data user. Another rationale for the creation of DAS was that NCES established that the amount of categorizing, top and bottom coding, and additional perturbations required to produce public use data files would render much of the content of the postsecondary sample survey data files useless to the average analyst (Seastrom 2003).

The NCES developed the DAS as a CD application in 1987. Users can use the CD at their desks but they cannot transfer files to their hard drives or over the Internet. The data are on the CD but the system was designed to prevent access to the microdata per se and permits only tabular results.

In 1997 the first DAS web application was designed and deployed. It permitted a download of the DAS software and gave tabular results, but data files could not be downloaded. Users could send their table requests through the Internet or through file

transfer protocol (FTP). The requests were processed in about 6 hours and the user could obtain the results in designated "pick-up bins" (Carroll 2003).

The third and current DAS application was designed and deployed in 2003. The application is available in windows and web-based formats. DAS Online is the web version of the DAS. The system allows users to create programming instruction files (DAS files) that specify the information they want to display in a table. There is a separate DAS for each survey data set and each one is usually built around an analysis report and includes the major analytical variables published in the reports. Users must learn the programming language used in DAS, however, all DAS applications have consistent interface and command structures. (NCES 2003, nces.ed.gov/dasol/index.asp).

The underlying DAS databases must include a series of Disclosure Review Board confidentiality edits. The edits are directed toward outliers and DAS applications with more capabilities require more perturbation edits. "However, every respondent/item is given a chance of being edited. In order for this to be effective with the least amount of edits, it is critical that how this is done and how much is done be kept strictly confidential" (Seastrom and Kaufman 2003, p. 3). There is a results filter for the tables that suppresses cells with less than 30 cases per cell and rows with proportions that have less than 30 cases in the denominator. Data files cannot be downloaded.

Aside from the use of the Internet, there have been other major advances in DAS over the years. The system can compute standard errors appropriate to the complex sample designs employed in the postsecondary surveys. It can compute a correlation matrix that can be used as input to run regression analyses and it allows users to recategorize the variables within the DAS. The batch processing feature of the system permits multiple data sets to be run in one job.

DAS has real-time processing of microdata. The web-based application results in tables delivered within seconds to minutes over the Internet. The system is available free of charge and unrestricted, 24 hours a day, seven days a week. Help is available on-line and personal assistance is given through email, as required.

Most of the survey files that DAS accesses are for postsecondary education analysis. There are eight of these surveys including: Baccalaureate and Beyond Longitudinal Study, Beginning Postsecondary Students Longitudinal Study, High School and Beyond Longitudinal Study, and National Postsecondary Students Aid Study. The use of DAS is a requirement in several NCES research contracts. For example, all estimates produced for the Postsecondary Education Descriptive Analysis Reports must come from DAS. Generally, analyses done are policy related. Examples of some reports produced based on DAS results include:

> How Families of Low- and Middle-Income Undergraduates Pay for College: Full-Time Dependent Students in 1999-2000. This report describes how the families of dependent students used financial aid and their own resources to pay for college, emphasizing variation by family income and type of institution attended.

What Colleges Contribute: Institutional Aid to Full-Time Undergraduates Attending 4-Year Colleges and Universities. This study provides information about recent trends in institutional aid receipt and then examines the relationship between such aid and the likelihood of recipients staying enrolled in the awarding institution relative to comparable unaided students,

Characteristic of Undergraduate Borrowers: 1999-2000. The report describes the demographic and enrollment characteristics of these borrowers as well as their risk for not persisting to completion of an educational program and the various types of loans and other financial aid they received.

Descriptive Summary of 1995-96 Beginning Postsecondary Students: Six Years Later. This report describes the enrollment, persistence, and degree attainment of students who began postsecondary education for the first time in the 1995–96 academic year. It covers the experiences of these first-time beginners over a period of six academic years, from 1995–96 to 2000–01, and provides information about the rates at which students completed degrees, transferred to other institutions, and left postsecondary education without attaining degrees.

Many more examples can be obtained at http://nces.ed.gov/das/reports.

The NCES 1999 Customer Satisfaction Survey is somewhat outdated and precedes the last revision of the DAS. However, it does give an order of magnitude of use and user satisfaction with the system. Overall, 42 percent of NCES' potential customers were aware of available databases and user tools and 12 percent had used them in the two years preceding the survey. Small percentages avoided user tools because they were too difficult. Four percent of respondents had used DAS applications in the two years preceding the survey and 84 percent of them were satisfied or very satisfied. Depending on the database, two to four percent of the respondents used the databases that have DAS applications and 78 to 91 percent were satisfied or very satisfied with the databases (NCES 1999).

Census Bureau (CB)

The Advanced Query (AQ) system was developed to give users the ability to request Census 2000 variables they want in tabulations from the one hundred percent and sample microdata files. Remote access to microdata was originally planned as part of the American FactFinder system that disseminates Census 2000 predefined standard summary tables over the Internet. However, the Advanced Query became a stand-alone system that requires user registration. The system was developed and tested for one hundred percent and sample data in 2001 and 2002. It was made available for use in April 2003 to State Data Centers, Census Information Centers and State Legislatures. The Census Bureau will expand the user base on a flow basis and will try to accommodate as many users as possible.

The individual records (observations) in the base files for both the Census 2000 one hundred percent data and sample data are swapped. The variables are categorized and users may not define their own categories. Sensitive variables dealing with items such as income and costs are top coded (Zayatz, Steele and Rowland 2000).

The system has a query filter that limits what users may select in the user interface. Users may select up to 3 variables for a table. The geographic detail available in AQ is limited to the standardized areas from Census 2000 and is more curtailed than that available from the summary tables in American FactFinder. In AQ the census block group is the smallest area available from the 100 percent data and the tract is the smallest area available from the sample data. Each area selected must have at least 200 people. The statistical results filter checks for a minimum mean and median cell sizes and minimum percentage of cells with one observation. The minimum values in the results filter are applied to every geographic area requested in the table. If any area does not meet any one or more of the minimum values, the entire area is omitted from the table (Rowland and Zayatz 2001).

The Advanced Query uses commercial business intelligence software that was tailored for use in the system. The system is completely automated and no manual intervention takes place. User logs of each query are maintained that contain the variables and geographic areas requested by each user. The logs are examined periodically to determine what is most often requested and to identify possible disclosure risks. There is no provision for automated complementary disclosure avoidance.

Users must register with the CB, but there is no cost to use the system. Registered users must log in with a user identification and password. They must learn how to select the geographic areas and variables they want through the web user interface. The results are returned in the form of database tables and means and medians can be requested with the table. The software allows the users to reformat the table in many different ways and graphic results are available. The results are generated in real time and are returned within seconds to minutes on the web. Tables may be downloaded in several formats and/or printed. The system is available during 24 hours a day, 7 days a week.

After the system was developed to access sample data, the Census Bureau tested the system with users primarily from State Data Centers, Census Information Centers, and Census Bureau Regional Offices. The purpose of the test was to determine the utility of the system based on the results the users could obtain with confidentiality filtering. Eighty-two testers produced 1,186 tabulation from Census 2000 sample data. They were asked to fill in evaluation forms to determine the usefulness of the tabulations. They evaluated approximately 370 tabulations.

The objectives of the tabulations were fully met for over half of the tabulations and partially met for 90 percent of them. “The main reason that objectives were not met fully was the confidentiality filters. More than half of the cases specified failure of geographic areas to pass the filters and one third specified the cause of failure to be the subject detail requested. About 20 percent mentioned insufficient subject detail available in the AQ

recodes" (Schneider 2002, p.2). Users expected to use the results of the tabulations for research, to plan or evaluate programs, to define needs, to apply for funding and to implement programs.

The testers gave many examples of the expected use of the tabulations. Most have a direct or indirect government purposes:

Examples of expected uses for research
"We are conducting an immigration study for Kentucky…to tell what parts of the state had larger numbers of non-citizens by specific demographic characteristics."
" … determine the tax burden on homeowners by comparing real estate taxes paid to mortgage payments and income."
"…identify level of need for basic water services in the US Also, to identify range and variance in water rates in US counties."
" … prepare a report on the needs of special populations in New York."
" …assessment of the percentage of their salary that NYC citizens are using to cover rent."

Examples of expected uses for planning or evaluating programs
"…(answer) policy questions about changes in welfare programs and the types of jobs the working poor have, with a focus on women."
" … better understand …the economic characteristics of non-English speakers …to develop and plan programs to match employer needs…"
"… plan and evaluate programs for the elderly and identify areas of the state (PA) where there may be special needs."
"… assess the potential effect of legislation on …(participation in) the state's social services programs."
"… provide a benchmark against which to measure what percentage of the target population is reached by programs."

Examples of expected uses for defining need, applying for funding
"Used by local government to assess number of householders 65 and above still having both a mortgage and a second mortgage."
"…identify need for low-income energy programs."
"… advocate for more ESL training for students…"
"… determine if workers in selected core occupations are able to locate affordable housing close to their place of work in major cities.."
"...understand the income distribution of households in which grandparents are responsible for minor children…"

Example of expected uses for implementing programs
"…identify the characteristics of the unemployed population to provide targeted job training, information, and services."
"…implement a program for mammography screening for women 50-64 who are in poverty."

"…estimate the number of income-eligible families for federal food stamp program in Minnesota in order to better target outreach to un-served eligible families."
"…need of county courts in selecting jury pools which include Hispanics."
"…analyze whether they meet affirmative action standards."
(Schneider 2003, anonymous quotes from testers, pp. 7-9)

In summary the test evaluation report found that, "testers included the Census Bureau's major intermediaries in distributing data, all of which are experienced in providing data to the ultimate users. Based on testers' comments, the AQS was efficient and friendly enough for these experienced users but could present difficulties to users who are not well versed in the use of census data." (Schneider 2002, p.1).

There are currently over 500 users registered to use the AQ system. The table below shows the usage statistics since May 1, 2003.

Advanced Query (AQ) Usage Statistics by Month

Month 2003	Number of users	Number of tabulations
May	72	886
June	54	947
July	75	611
August	119	762

Source: Census Bureau 2003

4. Sample of Research Projects in the US

There are a number of ongoing research projects in the US that have made contributions to practical applications or that collaborate with the Federal government to make microdata available remotely. These projects may have an impact on a wide spectrum of topics ranging from methodology, to hardware, to archiving and dissemination of microdata. Some use only restricted microdata and others do not. Many of the projects are funded by federal government agencies, the National Science Foundation, the National Academy of Sciences, the National Institutes of Health, etc. The examples reviewed below include just a few selected projects chosen because they touch on different aspects of remote access with implications for future development of remote access sites.

Digital Government

Continued research in monitored remote access to restricted microdata is important for future development of such systems by NSOs. Researchers from various institutions led by the National Institute of Statistical Sciences (NISS) have undertaken research in a number of areas of interest such as data swapping, confidentiality of tabular data, and

web-based systems that disseminate data and protect confidentiality (www.niss.org/dgii/techreports.html).

Researcher from NISS developed a prototype web system for the National Agricultural Statistical Service (NASS) of USDA to disseminate survey data on usage of fertilizers on farms. The principal purpose of the research was to develop a methodology to disseminate data for smaller geographical levels than states (preferably counties) and still protect the confidentiality of the farms in the survey.

The data consisted of almost 200,000 records of average fertilizer use per acre from 30,500 farms for the years 1996-1998. The goal was to allow dissemination of data at the county level, but data for half of the counties in the US were not eligible for disclosure using the prescribed confidentiality rules. Therefore, the researchers developed a methodology to aggregate geographic areas to the extent that data could be disseminated within the confidentiality rules. "Undisclosable counties are merged with neighboring counties (in the same state) to form disclosable "supercounties" (Karr and Sanil 2001, p. 1).

The system can be accessed through a web browser that allows the user to select the state, the crop and fertilizer type desired in the output. The output is in the form of a list of supercounties derived from the aggregation algorithm with related fertilizer application rates by crop by year and a list of component counties. The output can be shown on a map or in tabular form. A history of queries is kept in a database that can be used to monitor complementary disclosure.

The NASS prototype developed by NISS tackles both automatic aggregation of small geographic areas to avoid disclosure and examination of complementary disclosure from previously answered queries. These two automated techniques have thus far not been included in the remote systems examined in this paper due to lack of funding and feasibility. The NASS would like to implement the system pending availability of funding.

As part of the Digital Government program, researchers from Carnegie Mellon reviewed the confidentiality protection in the Advanced Query of American FactFinder (Duncan, Roehrig, and Kannan 2000). The examination of confidentiality protection in the system

> compared disclosure limitation used with that of agency best practice,
>
> determined if confidential data could be inferred using nonconfidential data outside the system or from data within the system itself, and
>
> assessed whether results from the system could be compromised using modern record linkage techniques.

Carnegie Mellon recommended a mechanism for controlling complementary disclosure from repeated queries through the use of a linear programming method to check for confidentiality before allowing queries. The method was too complex to implement.

Cornell Restricted Access Data Center (CRADC)

CRADC was created at the Cornell Institute for Social and Economic Research (CISER) to give selected external researchers access to restricted use microdata, research tools and inference-valid simulated microdata files under development. CRADC uses a windows user interface that shows the files available according to the access rights of the user. CRADC makes available research tools such as EXCEL, SAS, STATA, Matlab, Fortran V6, GLIM, Genstat, Gauss, etc. Available data sources include the Longitudinal Employer - Household Dynamics data that can be accessed only by state partners participating in the project.

Inference-valid simulated data files will be accessible through CRADC. Researchers will create inference-valid simulated microdata files by scientifically producing replacement values for actual microdata. The simulated microdata files are meant to render useful analytical results. Research will be carried out to determine if the use of the simulated microdata renders valid results when compared to the actual microdata. A working group will prepare the inference-valid simulated files and compare results to the actual microdata using the Survey of Income and Program Participation as a test database (Lane 2003).

Integrated Public Use Microdata Series (IPUMS) – International

IPUMS - International was developed by the Minnesota Population Center to "inventory, preserve, harmonize, and disseminate census microdata". Data from seven countries is available (China, Colombia, France, Kenya, Mexico, the United States, and Vietnam between 1960 and 2000) and data from Central and South America will be available soon. The census data are microdata samples and most files are public use files. Users must apply for access and sign an authorization form agreeing to abide by regulations for using the data. The use of the data is restricted to scholarly and educational purposes. (http://www.ipums.org/international/release_dates.shtml).

Internet 2 and Next Generation Internet

Over two hundred universities are participating in a nationwide project known as Internet 2 in collaboration with industry and government. Over one hundred of the most important computer and telecommunications corporations are involved. "The primary goals of Internet2 are to:

- Create a leading edge network capability for the national research community
- Enable revolutionary Internet applications

- Ensure the rapid transfer of new network services and applications to the broader Internet community"(http://www.internet2.edu/about/).

Internet2 has developed and deployed a "10-Gigabit-per-second national backbone supporting high-performance connectivity and Internet innovation within the US research university community" (http://www.internet2.edu/about/).

The Next Generation Internet (NGI) initiative is a multi-agency Federal research and development program that developed advanced networking technologies and applications demonstrated on testing environments that are 100 to 1,000 times faster than previous capabilities. Its goals, similar to those of Internet 2, have been completed. Federal agencies are currently coordinating advanced networking research programs under the Large Scale Networking (LSN) Coordinating Group (http://www.ngi.gov/).

USGenWeb

USGenWeb is a volunteer supported effort to present actual transcriptions of public domain records on the Internet. The files contain historical census records, marriage bonds, wills, and other public documents organized by state. It represents an effort to make microdata available for genealogy research to the general public. Although such files may not be considered to be statistical data by some, they are a type of microdata of great interest to the general public. The effort reflects the importance of the Internet for disseminating all types of microdata remotely and reflects the fact that the general public, not just sophisticated researchers, want and have access to remote access technology and microdata. (http://www.rootsweb.com/-usgenweb/)

5. Conclusions

Remote access to restricted microdata is still rarely implemented by NSOs due to their desire to protect confidentiality and the difficulty and expense of implementing systems that disseminate results with adequate disclosure avoidance. Some of the problems such as protecting the confidentiality of single outputs have been tackled more successfully than others such as avoiding complementary disclosure.

Generally speaking, the use of systems for remote access to restricted microdata by researchers is much less than their use of PUMFs on CDROM and RDCs, to date. This is because users are more familiar with CDROM and RDCs, and are more comfortable using them. Additional reasons for less use of remote systems are that NSOs purposely restrict access to the systems and because the data accessed by the systems may have less detail or more distortion than that available in an RDC.

Researchers, however, continue to express interest in using restricted microdata to obtain the types of analyses necessary for their work. NSOs appreciate these arguments, especially when the results are required by official agencies for necessary policy development and implementation. Therefore, the quest to make restricted microdata

available has resulted in research and experimentation with remote access systems. The challenge is to translate research into practical, affordable applications.

NSOs have shown an increased interest in the development of remote access systems for disseminating restricted use microdata in the last few years. The LIS program pioneered the technology with the LISSY in the late 1980s, and most systems developed or under development since then have employed similar methodology. The methodology can be referred to as "remote job execution systems" because the programs are executed off-line usually in a batch mode that takes anywhere from minutes to days to return results to the user. The amount of time required to return the results is largely determined by the degree of manual intervention used by the NSO to review the results for disclosure avoidance. The user communicates with the systems by sending programs from popular statistical software and receiving results through e-mail. The advantage of this type of system is that it allows users to work with their favorite research software permitting them to write their own programs and obtain the results they need within the confidentiality constraints imposed by the system.

A couple of remote access systems used by NSOs permit users to communicate with the system using a web browser or windows application. These systems use custom built or use tailored commercial software and deliver tabulations. Users cannot submit their own programs but the results are executed while the user is on-line and are usually returned within seconds or minutes because there is no manual intervention.

Currently, popular statistical programs such as SAS and SPSS are not web-enabled. Users may submit programs though email but cannot use them with a web browser. It is possible that such software will become web-enabled in the future, thus making their use of a web browser possible. Business intelligence software that is web enabled will increase the use of statistical measures thus becoming more like the favored statistical packages. Both of these possibilities foresee the merging of email and web systems.

Commercial software programs, be they statistical or business intelligence packages, in some cases have or will have the capacity to stop queries going in and out of the systems thus providing automatic query and results filtering. Programming is then required only to create the disclosure avoidance rules desired by the NSO.

The use of research results in practical applications is evident, and the cumulative effect is important. Continued research by statisticians is necessary to resolve the remaining problems of automating complementary disclosure avoidance and developing aggregated and/or user defined areas. The conduits for such research are often the mathematical statisticians from the NSOs who work closely with statisticians in universities. NSOs that have developed systems and test systems have gained and shared experience that is necessary to keep up with modern methods and technology. Existing systems contribute to the body of knowledge about remote access that can be drawn upon to build future systems.

As time goes on, developments in technology and familiarity with Internet applications among users will continue to grow, preparing fertile ground for applying research to further development of systems. Progress is inevitable due to advances in computer hardware, software, and Internet communications. Faster database and processing software and cheaper hardware will make it more feasible to implement the research.

NSOs and researchers will continue to reach compromises on what data can be made available according to the laws of each country. Hopefully, advanced methodology and technology will improve the ability of NSOs to make restricted data available for research with the confidence that confidential data will not be released.

Each generation of users will become more sophisticated in the use of Internet applications. Users have adapted as the media for data dissemination has gone from printed tables, to tapes, to diskettes, to CDROM, to remote access. Each decade's new media becomes the next decade's most popular media, as it was with public use files on CDROM, so it may be with monitored remote access to restricted data.

References

Anderson, Otto (2003) 'From On-Site to Remote Access - The Revolution of the Danish System for Access to Microdata', Paper presented at the joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg.

_____ and Lars Thygesen (2003) 'The Danish System for Access to Microdata;from on-site to remote access', Paper presented at Swedish Workshop on Microdata, Stockholm, www.micro2122.scb.se/papers.asp.

Berker, Ali and Susan Choy (2003) 'How Families of Low- and Middle-Income Undergraduates Pay for College: Full-Time Dependent Students in 1999-2000', National Center for Education Statistics Reports, http://nces.ed.gov/pubsearch.

Berkner, Lutz, Shirley He, and Emily Forrest Cataldi (2003) 'Descriptive Summary of 1995-96 Beginning Postsecondary Students: Six Years Later', National Center for Education Statistics Reports, http://nces.ed.gov/pubsearch.

Carroll, Dennis (2003) Emails Carroll to Rowland.

CEIES (2002) 'Opinions of the European Advisory Committee on Statistical Information in the Economic and Social Spheres (CEIES)', 19th Seminar on Innovative Solutions to Providing Access to Miocrodata , Lisbon, Contributed paper submitted by Eurostat, www.unec.org/stats/documents/2003.04.confidentiality.htm.

Cigrang, Mark (2003) Emails from Cigrang to Rowland.

_____ and Barry Schouten (2003) 'Remote Access Systems for Statistical Analysis of Microdata' Methods and Informatics Department, Statistics Netherlands.

Clinedinst, Melissa E., Alisa F.Cunningham, and Jamie P.Merisotis (2003) 'Characteristic of Undergraduate Borrowers: 1999-2000', National Center for Education Statistics Reports, http://nces.ed.gov/pubsearch.

Drolet, Gaetan (1999) 'Sherlock: A Web Magnifying Glass for Microdata Files' in International Association of Social Science Information Service and Technology Quarterly, Summer Issue, pp. 15-18.

Duncan, G., Roehrig, S., and Kannan, K. (2000) 'Final Report on the American FactFinder Disclosure Audit Project for the US Census Bureau', prepared under contract to the US Census Bureau.

Forster, Michael and Koen Vieminckx (2003) 'Inequality and Poverty Contributions of LIS', presented at the Luxembourg Income Study 20th Anniversary Conference, Luxembourg.

Gambhir, Vijay (2003) Emails from Gambhir to Rowland.

_____ and Kenneth Harris (2003) 'CDC/NCHS Data Center', Powerpoint presentation given at US Bureau of Transportation Statistics Seminar Series on Confidentiality, Washington DC.

Hamilton, Elizabeth and Chuck Humphrey (2002) 'Data Access and Data Use: The Missing Link', http://admin.acadiau.ca/library/DLI2003/session%201.2_pumfs/pumfs%20pumped.ppt.

Hjelm, Claus-Goran (2003) 'Remote Access to Microdata at Statistics Sweden', paper presented at the Swedish Workshop on Microdata, Stockholm, www.micro2122.scb.se/papers.asp.

Horm, John (1999) 'National Center for Health Statistics Approaches to Protection and Release of Microdata', contributed paper for Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki.

Horn, Laura and Katharin Peter (2003) 'What Colleges Contribute: Institutional Aid to Full-Time Undergraduates Attending 4-Year Colleges and Universities', National Center for Education Statistics Reports, http://nces.ed.gov/pubsearch.

Karr, Alan F., and Ashish P. Sanil (2001) 'Web-Based Systems that Disseminate Information but Protect Confidentiality', dg.o 2001: Proc. First National Conference on Digital Government Research, pages 159-166. Digital Government Research Center, Marina del Rey, CA.

Lane, Julia (2003) 'Synthetic Data and Confidentiality Protection', paper presented at the Swedish Workshop on Microdata, Stockholm, www.micro2122.scb.se/papers.asp

Luxembourg Income Study (2003) Summary of Numbers of Users and Jobs Submitted to LIS by Country, 2001 – 2003, Luxembourg.

_____ (2003) Number of Working Papers Published per Year 1985-2002, Luxembourg.

Nordback, Lars (2003) Email from Nordback to Rowland.

Rowland, Sandra and Laura Zayatz (2001) 'Automating Access with Confidentiality Protection: The American FactFinder', in Proceedings of the Social Statistics Section, American Statistical Association, Alexandria.

Schneider, Paula J (2002) American Fact Finder Advanced Query System - Assessment Report on Stage Two (Sample File) Beta Testing, prepared under contract to the US Census Bureau, Washington, DC.

Schouten, Barry and Jan Jonker (2003) 'Remote Access at Statistics Netherlands', paper presented at the Swedish Workshop on Microdata, Stockholm, www.micro2122.scb.se/papers.asp.

Seastrom, Marilyn (2003) Email from Seastrom to Rowland.

_____ and Steven Kaufman (2003) 'NCES Disclosure Risk Procedures', paper presented at the American Statistical Association Meetings in San Francisco.

Statistics Canada (2002) Survey of Labor and Income Dynamics (SLID) Workshop Presentation, Montreal, http://www.ciqss.umontreal.ca/Documents/acetat_e_2002.ppt.

Tambay, Jean-Louis, Gustave Goldman and Gerry Potter (2003) 'Providing Researcher Access to Data for Analysis at Statistics Canada', paper presented at the Swedish Workshop on Microdata, Stockholm, www.micro2122.scb.se/papers.asp.

Trewin, Dennis (2003) 'Access to Microdata – Issues, Organization and Approaches' paper presented at the Conference of European Statisticians, Geneva.

US Department of Commerce, Census Bureau (2003) Advanced Query User Guide for 100 Percent Data and Advanced Query User Guide for Sample Data, Washington, DC.

US Department of Education, National Center for Education Statistics (2003), NCES Handbook of Survey Methods, Appendix C: Web-based and Standalone Tools for Use with NCES Survey Data, http://nces.ed.gov/pubsearch.

_____ (1999) NCES Customer Satisfaction Survey Report, Section IV. Questions about NCES Databases and User Tools, http://nces.ed.gov/pubsearch.

Zayatz, Laura, Philip Steel, and Sandra Rowland (2000) 'Disclosure Limitation for Census 2000', in Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria.