

An Examination of Monitored, Remote Microdata Access Systems

Sandra Rowland

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

- Introduction
- Six Foreign Systems
- Three US Systems
- Research Projects in US
- Conclusions
- Appendices have a Summary

Purpose of the Paper

- Simple statement of known practices in NSOs
- Not an evaluation or assessment of systems or research
- Review of systems according to the references acquired
- Papers from conferences in Europe
- Direct contacts by e-mail

Monitored Remote Microdata Access Systems

- Computer systems that provide access to **restricted microdata files** on the Internet using methodology to limit or suppress output for disclosure avoidance.
- Does not include systems that access **Public Use Microdata Files (PUMFs)**.

Methodologies Employed in Systems

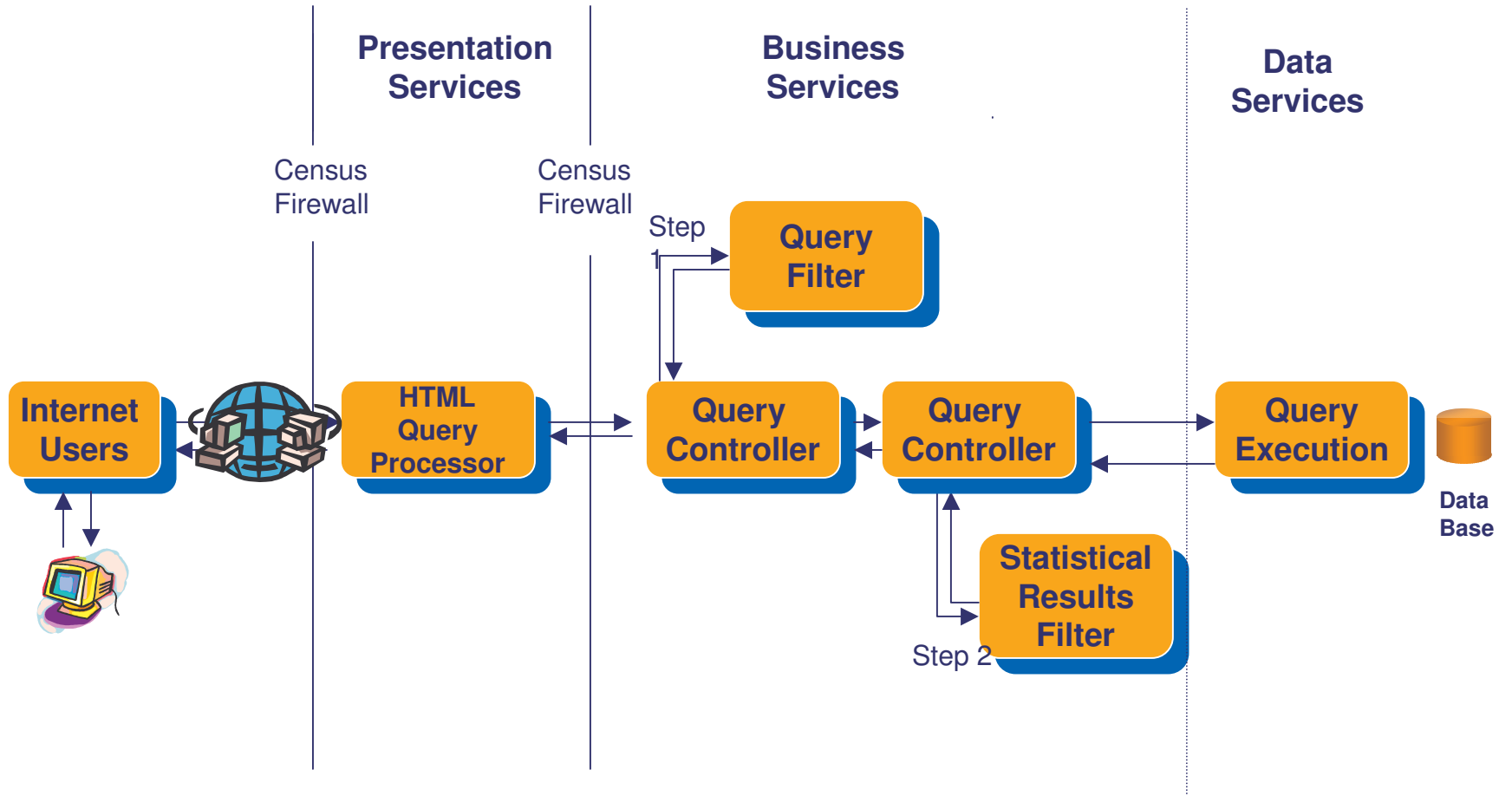
- Electronic authentication of users
- Email or web user interface
- Confidentiality edits to the base file
- Query and output filters
- Automatic and manual intervention
- Usage logs for review

Desirable Methodologies

- Complementary disclosure avoidance techniques
- User-defined areas or automatic aggregation of nonstandard areas

System Overview

Source: Census Bureau Advanced Query



Systems Usage

- Authorization required
- Principal or permitted users
- Files available
- Metadata and assistance
- Turnaround time
- Cost
- Hours available
- Types of analysis
- User assessments

Foreign Programs

- Luxembourg Income Study System
1987
- Statistics Canada 2001
- Statistics Denmark 2001
- Statistics Netherlands 2002
- Australian Bureau of Statistics 2003
- Statistics Sweden 2003

US Federal Agencies

- National Center for Education Statistics
1997
- National Center for Health Statistics
1998
- Census Bureau 2003

Luxembourg Income Study System (LISSY)

- Grandfather of systems
- Used as a model by others either consciously or unconsciously
- LISSY as used in the LIS program
- LISSY is a stand-alone system used in other programs like LES and Eurostat

LISSY Methodology

- Remote job execution system
- Email user interface
- SAS, SPSS, STATA programs emailed by users
- Restricted files kept at agency
- Processing done off-line (batch)
- Output returned by email

LISSY Methodology

- Retrieves the email program from user
- Authenticates users
- Checks confidentiality issues
- Returns jobs not accepted
- Processes accepted jobs in batch
- Examines output file for confidentiality
- Returns acceptable results to user
- Sends suspicious output for manual review
- Maintains query and output logs for review

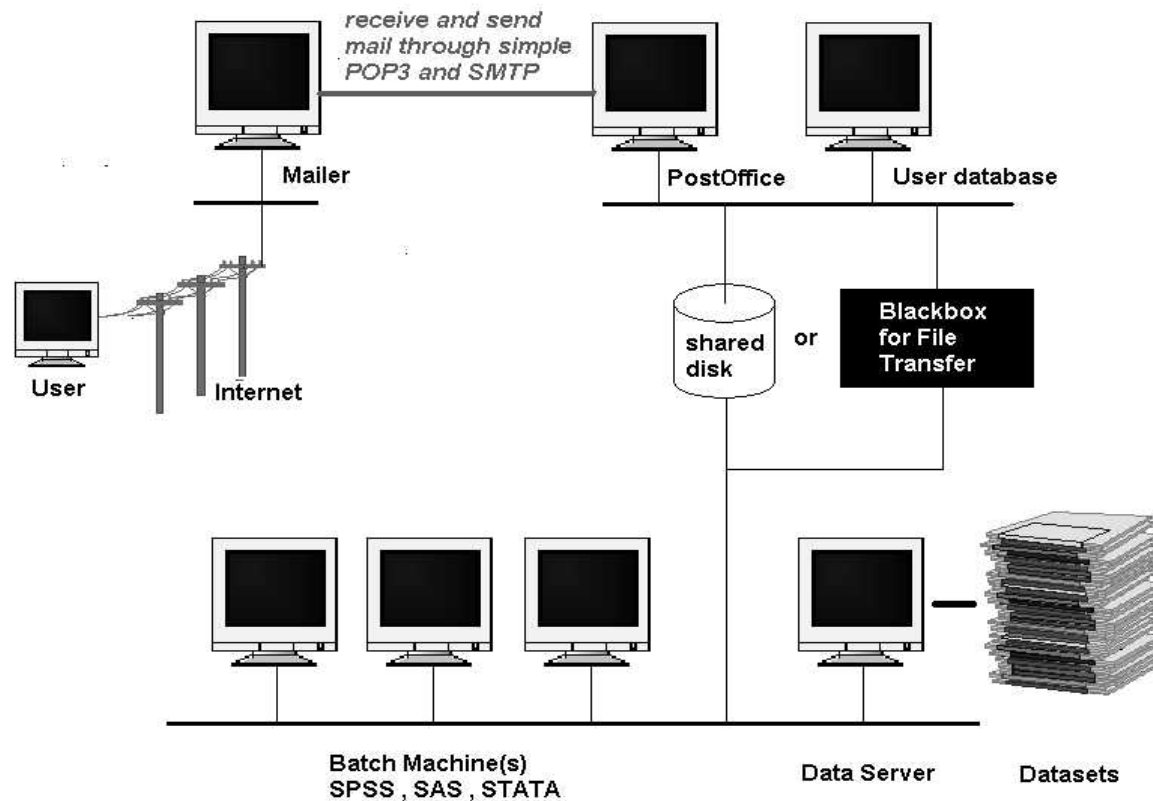
LISSY Methodology

- No complementary disclosure techniques used
- “Complete evaluation of multiple queries may be too complex, time consuming or restrictive to implement” (Source: Schouten and Cigrang 2003)

LISSY as Used in LIS

Source: Schouten and Cigrang (2003)

Remote Access Systems for Statistical Analysis of Microdata



LIS Usage

- International program that makes microdata from 66 household income surveys available from 25 countries
- Public use and restricted files
- Researchers submit a proposal and sign a contract and confidentiality pledge
- Academic users are majority

LIS Usage

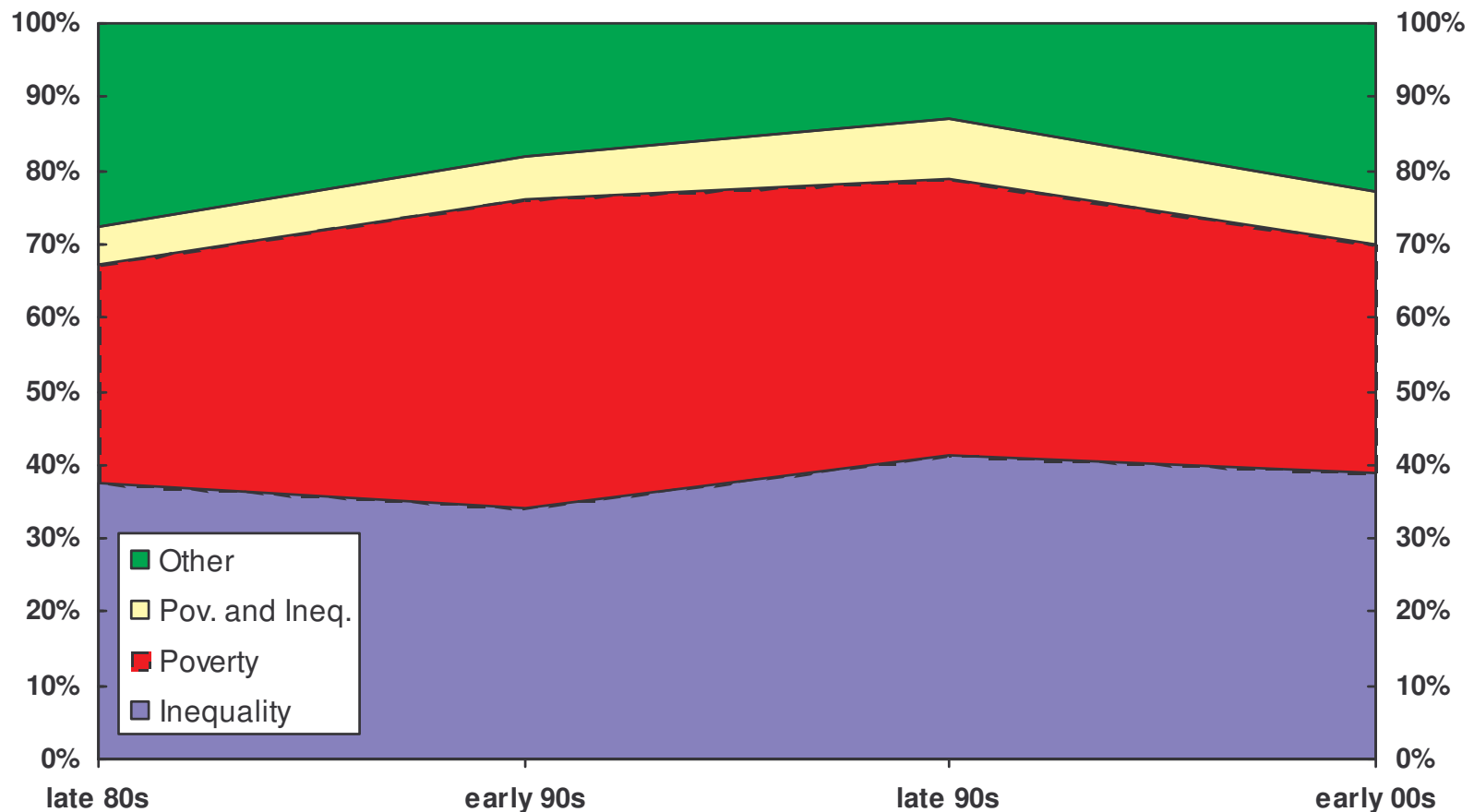
- Metadata: documented key variables, synthetic microdata files
- Workshops and help desk
- Annual country fee, but no cost to researcher
- Turnaround in minutes
- Available 24/7

LIS Analysis

- 20-year history
- 01/01-06/03 average per year
 - 213 users submitted 36,280 programs
 - US researchers submitted 10,047 per year
- Major research on inequality of income and poverty

LIS Working Papers

Source: Forster and Vieminckx (2003)
Inequality and Poverty contributions of LIS



Statistics Canada Methodology

- User authentication
- Email interface
- Varies by survey: SAS, SPSS, STATA, Foxpro, flat ASCII
- Disclosure control varies by survey and availability of PUMFs
- No automatic filters - disclosure avoidance manual
- Techniques to prevent linkages

Statistics Canada Usage

- Users contact agency to create files for use
- Users must be familiar with record layout for some surveys but not others
- Government and university major users
- Available 8/5
- Turnaround in 1-2 working days

Statistics Canada Analysis

- **National Population Health Survey**
 - 2001 - 99 programs per month
 - 2002 - 40 programs per month
- 242 articles in 91 journals
- Cancer and diabetes prevalence
- Utilization of health services

Statistics Canada Analysis

- **Survey of Labor and Income Dynamics**
- 05/02-07/03 - 160 requests for SLID data
- Research in employment and labor dynamics, life-cycle labor transitions, job quality, life events and family changes, dynamics of low income, combining work and school

Statistics Denmark

- Experience from 2001 to date has been good with no breaches in confidentiality
- Will replace on-site access with remote access to restricted microdata
- Remote access granted to only authorized institutions
- 03/01-03/03 - 43 authorizations granted

Statistics Denmark

- Users communicates on Internet
- Users submit SPSS, SAS, STATA programs and may **work with the data freely creating new databases from the originals**
- Batch processing done at agency
- Output examined by agency staff
- Output returned by Email

Statistics Denmark

- Register-based samples
- Research databases link information from various registers
- Demographic database
- Fertility database
- Health register
- Social research register
- Most popular is Integrated Database for Labor Market Research
- Custom produced registers

Statistics Netherlands

- 2002 pilot with Dutch Ministry of Social Affairs
- Submitted SPSS programs by email
- Data on social allowances
- Output manually reviewed to automate filters especially an output filter
- Disclosure rules developed based on pilot

Statistics Netherlands

- Made available to all government agencies in 2003
- May extend access to nongovernmental researchers in future
- Future system will use more types of analytical software, automate filters and construct log files

Australian Bureau of Statistics

- Remote Access Data Laboratory (RADL) became available in April 2003
- Key area of future development
- ABS will encourage use of RADL when users want linked files

Australian Bureau of Statistics

- SAS and SPSS programs by email
- Output reviewed by agency staff
- Automatic triggers for closer inspection
- Downloading of unit data possible for up to 30 records to support outlier detection

Statistics Sweden

- Exploring feasibility of a system similar to Statistics Denmark's
- Researchers and public authorities took part in feasibility study
- Programs submitted and results obtained by email
- Results in form of tables

US Federal Agencies

- Agriculture, NASS and ERS
- Commerce, Census Bureau
- Education, NCES
- Energy, EIA
- Health, NCHS
- Labor, BLS
- Justice
- Transportation, BTS

NCHS Methodology

- Analytical Data Research by Email (ANDRE) 1998
- **Web component under development**
- User Id/password authentication
- SAS programs

NCHS Methodology

- Certain SAS log commands not permitted, users' programs may be slightly modified
- 90% automated filters, 10% output randomly checked
- Cell and row suppression in output
- Usage logs for confidentiality review

NCHS Usage

- Registered subscribers, proposals received and approved by NCHS
- Anyone may apply
- 45 users in past 5 years
- 10,000 SAS programs run
- Cost \$500 per month
- Available 24/7
- Turnaround in hours

NCHS Usage

- Virtually all NCHS files available
- “In general each data set is specifically prepared for the user. The data set may include many variables selected from multiple internal data files of NCHS. User supplied data may also be merged.” (Gambhir 2003).

NCHS Analysis

- **National Family Growth Survey**
- 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

NCES Methodology

- Data Analysis System (DAS)
- No authorization required
- Custom built application accommodates various surveys
- Tables with standard errors and correlation matrices
- Must be enough cases to allow calculations
- Row suppression

NCES Methodology

- Web user interface
- 100 % automated, no manual intervention
- Output in seconds to minutes
- “Real-time” processing
- “On-line”

NCES Usage

- Free
- Available 24/7
- Principal users are policy analysts and researchers
- 1999 survey: 4 % of respondents used DAS in two years preceding the survey and 84 percent were satisfied or very satisfied

NCES Analysis

- 8 surveys dealing primarily with **post secondary education**
- National Post Secondary Student Aid Study is most popular
- Postsecondary Education Descriptive Analysis Reports require use of DAS

NCES Analysis

- How Families of Low- and Middle-Income Undergraduates Pay for College: Full-Time Dependent Students in 1999-2000
- Characteristic of Undergraduate Borrowers: 1999-2000
- Descriptive Summary of 1995-96 Beginning Postsecondary Students: Six Years Later

Census Bureau Methodology

- Advanced Query
- Census 2000 100% and sample data
- User name and password
- Web interface
- COTS business intelligence software tailored for use
- SQL processing

Census Bureau Methodology

- 100% automated query and results filters
- Tables only
- Whole table suppression for geographic areas that don't pass confidentiality filters

Census Bureau Usage

- Registered users
- State Data Centers, Census Information Centers, State Legislatures
- 500 registered users as of August 2003
- 600-900 tabulation per month
- Results in seconds to minutes
- Available 24/7, free

Census Bureau Usage

- Pilot test on Census 2002 sample
- 82 testers produced and evaluated 370 tabulations for utility
- Main reason objectives not fully met was the confidentiality filters
- Half specified failure of geographic areas to pass the filters and a third specified failure of subject detail to pass

Census Bureau Analysis

- Analysis for direct or indirect government purposes
- Research, plan or evaluate programs, define needs, apply for funding and implement programs

Research in US

- **Digital Government**
- NISS developed a prototype system for NASS/USDA to aggregate areas not passing filters and to avoid complementary disclosure
- Carnegie Mellon evaluated AQ and recommended LP algorithm for complementary disclosure avoidance

Conclusions

- Systems are still rare
- Systems not heavily used due to restrictions
- Difficulty and expense of implementing systems that disseminate results with adequate disclosure avoidance

Conclusions

- Confidentiality of single output tackled more successfully than avoiding complementary disclosure
- Research necessary to resolve the remaining problems of automating complementary disclosure avoidance and developing aggregated and/or user defined areas

Conclusions

- Analysts interested in more detailed data and avoiding expense of visits to RDCs
- NSOs showing more interest in remote access in last few years
- NSO experience in developing systems is accumulating

Conclusions

- Email, “remote job execution systems” most common in foreign counties
 - permit user to write own programs
 - have manual intervention
- In US, web systems under development or in use
 - faster
 - no manual intervention

Conclusions

- Future may have Web enabled statistical software packages
- Business intelligence packages will include more statistical measures
- Packages will have automatic filtering capability

Conclusions

- Progress is inevitable due to advances in computer hardware, software, and Internet communications
- Each decade's new media becomes the next decade's most popular media