

# New Tools for Accessing Data While Still Protecting Confidentiality

ANCSDAAP Population Census Conference

Daniel H. Weinberg and Nancy M. Gordon  
U.S. Census Bureau  
March 2009

This presentation is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological issues are those of the presenters and not necessarily those of the U.S. Census Bureau.

# Traditional Methods of Data Access

## Tabulations

- Confidentiality of respondents protected by limiting the number of cells relative to the number of observations
- May use cell suppression to preserve confidentiality of responses, especially for business data
- Census Bureau uses the American FactFinder for Internet access to all tabulations from its censuses

# Traditional Methods of Data Access

## Public–Use Microdata Samples

- Confidentiality of respondents protected by omitting some information and modifying some of the remaining information
- Methods include top– and bottom–coding, re–categorization, adding noise, swapping, and geographic aggregation

# Three Recent Data Access Approaches

- Licensing – providing restricted data directly to individuals or organizations under a confidentiality protection agreement
- Statistical data enclaves – enclaves provided by National Statistical Offices for research purposes
- Remote access – submission of analysis requests (typically computer programs)

# Licensing

Aspects of the license agreement:

- Define the information covered by the agreement
- Specify the individuals who may have access to data
- Describe required disclosure avoidance and clearance procedures for research results
- List administrative requirements including IT security, unannounced inspections, penalties for violations, and advance approval of projects

# Statistical Data Enclaves

- The nine (soon to be ten) Census Bureau Research Data Centers (RDCs) are partnerships with academic and non-profit organizations
- Meet all physical and computer security requirements
- Researchers must pass fingerprint check and are subject to the same penalties as employees
- Applications must show how project will benefit the Census Bureau's programs

# Research Data Centers

Research increases the utility and quality of Census Bureau data products

- Encourages knowledgeable researchers to become familiar with an agency's data products and data collection methods
- Subjects current collection and processing methods to testing through additional uses of resulting data
- Leverages the value of existing data by allowing data linking not possible outside a secure site

# Other Statistical Data Enclaves

## Examples:

- U.S. Bureau of Labor Statistics (HQ only)
- U.S. National Center for Health Statistics (HQ only)
- U.S. National Institute of Child Health and Human Development (3 locations)
- National Opinion Research Center (a non-profit government-funded organization -- 2 locations)
- Canada, Germany, New Zealand, UK

# Remote Access

- Data files usually edited in advance to reduce the possibility of disclosures
- Employ automated and manual filters that block certain kinds of queries and certain results from being returned to researchers
- Must be monitored automatically or manually for disclosure avoidance

# Remote Access, continued

## Methodologies

- “Remote job execution systems” – an email interface that allows users to send programs; processing is usually done in batch mode
- Web interface with custom-built or custom-tailored software

# Examples of Remote Access

- Luxembourg Income Study
- Australia, Canada, Denmark, the Netherlands, New Zealand, Sweden
- In the U.S.: Census Bureau Advanced Query System (discontinued) and Microdata Analysis System, National Centers for Education and Health Statistics

# HotReports, DataFerret, TheDataWeb

- *TheDataWeb* provides a fast and reliable infrastructure to allow users to gain access directly to data stored remotely to support DataFerrett and HotReports
- *DataFerrett* provides powerful statistical tools for sophisticated analysts to use for their own work and to create displays in HotReports
- *HotReports* organize data into usable information for decision makers who are not statistical professionals

# Demonstration

- American FactFinder
- HotReports
- Use of Synthetic Data: *OnTheMap*

# Concluding Comment

- Threats to public microdata release are increasing
- Statistical agencies must respond to this threat in order to meet the needs of their users, both through traditional methods and innovation
- Statistical agencies have responded – through licensing, research enclaves, remote access, and new Internet-based tools