Issues in Accessing and Sharing Confidential Survey and Social Science Data

CODATA 2002, Montreal
October 3, 2002

Virginia A. de Wolf, Silver Spring, Maryland, USA
(dewolf@erols.com)

Outline of Presentation

- Provide brief background on U.S. Federal statistical system;
- Review the two primary approaches that U.S. Federal statistical agencies use to share confidentiality data collected from individuals and organizations;
- Highlight the contributions of three committees; and
- Conclude with suggestions for sharing confidential social science data based on experiences of the U.S. Federal statistical system.

The U.S. Federal Statistical System

- Is decentralized.
- Comprised of over 70 agencies.
- Agencies collect data from individuals and organization
  1. to inform policy decisions and
  2. for research.

The U.S. Statistical System (cont'd)

- With respect to the confidential information that they collect, agencies are "data stewards" and must balance two objectives:
  1. to assure that the responses of respondents are protected and
  2. to provide uses statistical information to data users.

Important to remember: There is no such thing as a "zero risk" of disclosure (parenthetically, the only way to have no risk is to not collect data). Federal agencies work hard to keep this risk as low as possible.

Presentation to Highlight Contributions of Three Committees

- Earlier committee # 1: Panel on Confidentiality and Data Access
  - Convened by the National Research Council's Committee on National Statistics.
  - Chair: George Duncan, Carnegie Mellon University
  - Work of Panel resulted in publication of *Private Lives and Public Policies* (Duncan et al., 1993).
  - Commissioned papers are contained in a 1993 special issue of the *Journal of Official Statistics*.

Highlight Three Committees (cont'd)

- Earlier committees # 2: Subcommittee on Disclosure Limitation Methodology (called "Subcommittee")
  - Organized by the Office of Management and Budget's (OMB's) Federal Committee on Statistical Methodology (FCSM).
  - 1994 Publication: "Report on Statistical Disclosure Limitation Methodology" http://www.fcsm.gov/working-papers/wp22.html

    Note: Chapter 2 of Subcommittee's report contains an excellent primer.

Highlight Three Committees (cont'd)

- Ongoing committee: FCSM's Confidentiality and Data Access Committee (CDAC)
  - Began in 1995.
  - Members are staff in Executive Branch agencies.
  - Over 16 agencies represented.
  - Products and related papers contained on its web site will be cited: http://www.fcsm.gov/committees/cdac

Panel on Confidentiality and Data Access

- Panel was first to provide generic labels for the two main alternatives that U.S. Federal statistical agencies use to protect the confidentiality of data that they collect. These are:
  1. Restricted data -- to restrict the content of the data prior to releasing it to the general public and
  2. Restricted access -- to restrict the conditions under which the data can be accessed (i.e., who can have access, at what locations, for what purposes).

Restricted Data Approaches by Type of Data Product

- Tables
- Microdata files

  Definition from Subcommittee's report: A microdata file is a computerized file that "...consists of individual records, each containing values of variables for a single person, business establishment or other unit."

  Notes: (1) Confidential data from organizations are rarely released as microdata because risk of re-identification is too high. (2) Confidential data from individuals are released as either tables or microdata.

Restricted Data Approaches: Tables

- If information is collected on a census, one way of preserving confidentiality is to only release tables based on a sample.
- Regardless of whether the data are a census or sample, the cells in a table should not be "too" small (some agencies require a minimum of 3 entries per cell while others require 5). This leads to the method of "cell suppression."

Tables (cont'd)

- Cell suppression:
  - Insert zero in cells containing "small" values.
  - After suppressing a value in a row, you must also suppress values in one or more other row(s) and column(s) so that the suppressed value can not be obtained by subtraction from the row/column totals.
  - Appropriate statistical methods must be used (see 1994 report by Subcommittee; especially see "primer" in Chapter 2).

Tables (cont'd)

- Sometimes the resulting "suppressed" table contains too many "blank" cells to be of value to data users. Policies have been developed to enable "small" cells to be published, e.g.,
  - National Agriculture Statistics Service (NASS) has a policy that allows its data providers to "waive" the confidentiality protection so that small cells can be published (data providers must sign waiver).
- NASS also produces special tables for data users and posts them on its web site.

Restricted Data Approaches: Microdata

- Creating a public use microdata file is as much an art as a science since
  - the methods used to protect confidentiality are varied and
  - often depend on the type of data that underlies the microdata files.
- First step: remove all personal identifiers. Difficult question: What is identifiable? See CDAC's paper "Identifiability in Microdata Files."

Microdata (cont'd)

- Second step: use methods to lessen the chance of re-identifying individuals from "unique" combinations of variables, e.g.,
  - Releasing a random subsample;
  - Limiting geographic detail;
  - Reducing the number of "unusual cases" (examples of methods used include rounding, recoding categorical responses, using ranges for age rather than exact age or date of birth); and
  - Increasing the uncertainty associated with data (i.e., data swapping, adding random noise).

Microdata (cont'd)

- Computationally intensive statistical methods are also used, e.g., multiple imputation (Little and Rubin, 1987). The Federal Reserve Board's Survey of Consumer Finances uses multiple imputation as a disclosure-limiting technique.
- In the next presentation Jack McArdle and David Johnson will discuss several statistical techniques to reduce the potential of inferential disclosure.

Microdata (cont'd)

- Because of the expansion of data available via the internet it is critical to conduct "re-identification assessments" that attempt to ascertain the identify of individuals. Some agencies have hired "hackers" under contract to do this; some do it in-house. Needs to be done
  - prior to the release of all microdata files and
  - on earlier microdata data releases: important to determine whether or not microdata files which were once deemed "protected" can inadvertently be re-identified.

Assessing the Level of Protection for Tables and Microdata Prior to Release

- Prior to releasing a restricted data product, agencies assess the level of protection afforded the confidential information; this is done through a formally or informally designated unit called a Disclosure Review Board (DRBs).
  - For information on DRBs, see CDAC's web site for panel session on DRBs presented at the August 2000 Joint Statistical Meetings.

Assessing the Level of Protection (cont'd)

- CDAC's "Checklist on Disclosure Potential of Proposed Data Releases": based on the practices of several agencies and contains three subsections:
  - one for microdata files and
  - two for tables (one for data collected from individuals, the other for data collected from organizations).
- Completed Checklists should be submitted to the Disclosure Review Board for review.
- Organizations should modify the Checklist as needed. (Note. Checklist is on CDAC's web site.)

Restricted Access Procedures

- Administrative procedures to enable research use of confidential data.
- Agencies place restrictions
  - on the use of the data (for statistical purposes but not for regulatory, judicial, or other administrative purposes);
  - conditions of access (e.g., location, cost);
  - whether or not data can be linked (and if so, who does the linking); and so forth.

Three Examples of Restricted Access Procedures

- Research Data Centers
- Remote Access Systems
- Licensing or Data Use Agreements

Research Data Centers (RDCs)

- The Census Bureau pioneered RDCs
  - which were first used to enable researchers' access to economic microdata.
  - The National Science Foundation was involved in establishing this Census Bureau program.
  - There are six RDCs at this time.
- Other RDCs
  - National Center for Health Statistics
  - Agency for Healthcare Quality and Research
  - Statistics Canada initiative

Research Data Centers (RDCs) (cont'd)

- "Typical" RDC characteristics:
  - Researchers access the data at a site controlled by agency and staffed by employees;
  - Research projects must be approved by the agency;
  - Researchers enter into a formal agreement with the agency and often cover costs associated with the work (e.g., computer charges, rental of space);
  - Use of "stand alone" workstations that do not have floppy disk drives or CD readers and are not connected to the internet or any agency network;

Research Data Centers (RDCs) (cont'd)

- "Typical" RDC characteristics: (cont'd)
  - Restrictions on linking data (in general if a linkage is approved it will be done by agency staff);
  - Inspection of all materials removed from the RDC;
  - Limitations on the types of analyses; and
  - Disclosure review of researchers' output.
- For information on RDCs see
  - CDAC's "Restricted Access Procedures" paper.
  - Statistics Canada web site: http://www.statcan.ca/english/rdc/index.htm

Remote Access Systems

- National Center for Health Statistics' (NCHS) system is handled by its RDC and has two components:
  - After a proposal is approved, RDC staff develop a "pseudo" data file which has the statistical properties of the actual data file. This fictitious file is then sent to the researcher who uses it to debug computer programs.
  - Researcher sends NCHS debugged files by email:
    - All programs are automatically scanned upon arrival for non-allowable commands (certain SAS procedures are disabled).
    - The output is reviewed before it is emailed back to the researcher. (For information: http://www.cdc.gov/nchs/r&d/rdc.htm)

Licensing or Data Use Agreements

- Licensing or data use agreements that allow researchers to use non-public data at their home institution.
- Note. Seastrom's paper (2001) is an excellent summary of the current status of the use of licenses in a wide number of U.S. agencies.
- Following example is from National Center for Education Statistics (NCES).

NCES's License

- Application must include
  - Formal letter of request (e.g., who will use the data, a description of the planned statistical use of the data, specification of the time period for the loan of the restricted data file);
  - License documentation (i.e., a legal agreement signed by the researcher, a senior official at the researcher's institution, and NCES's commissioner);
  - Security plan at the home institution (NCES has specified a list of requirements); and
  - Affidavits of nondisclosure to be signed by each data user.

NCES's License (cont'd)

- Once licensed, researchers
  - Must follow NCES publication requirements when publishing results from restricted data;
  - Agree to unannounced and unscheduled on-site inspections by NCES's contractor, and
  - Return restricted data files to NCES once the project is completed.

Suggestions for the Social Sciences

- Ideas for Professional Associations
- Ideas for Educational Institutions

Professional Associations

1. Sponsor short courses that focus on "restricted data" and "restricted access" approaches.
   - Involve CDAC members; have it tailored to your discipline.
   - Involve association members with expertise.
2. Provide resource materials (e.g., on the association's web sites) including
   - Relevant laws and regulations that affect your members, e.g.,
     - Changes to Federal regulations governing grants (OMB Circular A-110)
     - Certificates of Confidentiality which prevent compelled disclosure in a court of law. Note. These are available from the Department of Health and Human Services irrespective of the source of funding for the project.
   - Information on restricted data methods; and
   - Information on restricted access procedures.

Profession Associations: Information on Restricted Data Methods

- Include links to Federal resources (ex., CDAC) as well as web sites from other countries, e.g., Canada, Eurostat, and Statistics Netherlands;
- Provide examples that are "relevant" to the discipline; and
- Encourage members to conduct "re-identification" assessments prior to releasing a new microdata file as well as doing such checks on microdata files that were released at an earlier point in time.

Profession Associations: Information on Restricted Access Procedures

- Include links to Federal examples (such as Census and NCHS); and
- Provide examples from Federal grantees subject to OMB Circular A-110 about restricted access approaches that are being used, e.g.,
  - the Health and Retirement Survey at the University of Michigan's Center on Demography of Aging has restricted access agreements and also supports a data enclave.

Educational Institutions

1. For data funded by grants and governed by OMB Circular A-110:
   - What are other disciplines doing?
   - Check with you legal office. Ask if it has a developed a plan of action if faculties' data are subject to a Freedom of Information Act based on use of grant data by the Federal government.
2. Create a cross-disciplinary DRB to review tables and microdata created from confidential data collected from individuals and organizations. DRB would make recommendations to researchers about the level of protection. Use/adapt Checklist.

Educational Institutions (cont'd)

3. See if your university's Institutional Review Board (IRB) has formalized a process for review of output from data collected under a pledge of confidentiality. If not, then perhaps a cross-disciplinary DRB could serve as an ad hoc committee to make recommendations about release to the IRB.
4. Create a cross-disciplinary Research Data Center on campus.

   An open question: Can the institutions that fund most of the social science research (National Science Foundation and National Institutes of Health) provide grants to establish such Centers?