UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2003/33
20 May 2003

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-first plenary session
(Geneva, 10-12 June 2003)

ACCESS TO MICRODATA – ISSUES, ORGANISATION AND APPROACHES

Supporting paper submitted by the U.S. Census Bureau[1] [2]

1. This paper describes the various issues and approaches the U.S. Census Bureau has considered to ensure that research microdata are both protected and made available for public policy use. There is no single issue to address or single approach that works for all research uses. Rather, the Census Bureau applies a data stewardship model that strives to achieve mission objectives while meeting legal and ethical constraints. The goal is to ensure data quality and maximize use while protecting privacy and confidentiality of respondents. Within this context, the Census Bureau strives to provide data users with options that meet specific research needs while protecting against confidentiality breaches and improper use of the identifiable records.

I. PUBLIC USE MICRODATA: IMPORTANCE, LIMITATIONS, AND THREATS

2. To conduct public policy analysis and research, federal, state, and local policy makers—and researchers from many disciplines—rely heavily on the Census Bureau to provide high quality information on the population and the economy of the U.S. The Census Bureau makes these data available to external users in the form of tables or public use microdata files that have been "disclosure proofed" to protect the identity and privacy of respondents.

3. A recent book and conference on confidentiality and data access brought home the Census

[1] Prepared by Gerald Gates, Patricia Doyle, Sam Hawala, Arnold Reznek and Shelly Wilkie Martinez.
[2] This paper reports on results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited Census Bureau review than official Census Bureau publications. The report is released to inform parties of research and to encourage discussion on work in progress.

Bureau's growing challenge to maintain its historical commitment to respondent confidentiality and still meet the American public's growing data needs (Doyle et al., 2001). The latest research suggests several reasons we will have a problem maintaining confidentiality (as defined by current legislation) in the future if we continue with our current disclosure and data dissemination methods:

- There is a growing wealth of individual- and business-level information available in the public domain.
- Data from other agencies that do not follow strict disclosure guidelines are publicly available.
- Technology to mine public information is increasing in sophistication.
- The general public has increasing concerns over privacy.

4. To meet critically important public policy and research needs, the entire federal statistical system faces increasing demands for more, better, and more recent data. The Census Bureau is responding to this situation with improved disclosure techniques, but the methods that reduce disclosure risk also reduce the level of detail and the quality of the data disseminated publicly. To remain the preeminent provider of data for public policy and research, the Census Bureau must be proactive in addressing the challenges posed by the simultaneous increase in stress on our system of maintaining confidentiality and increase in demand for our data.

5. A change from our current approaches to disclosure and dissemination would involve recognizing that disclosure risk is composed of both opportunity and incentive. Our disclosure practices to date have focused on minimizing opportunity, since we have had little control over incentive. However, to make substantial strides toward making public data more usable, without disclosure risk, we need to extend our focus to address the incentives users have for attempting to identify individuals in public use microdata. Activities needed to minimize both opportunity and incentive involve technological advances, legal strategies, policy enhancements, interagency coordination, new disclosure techniques, and privacy research.

6. Technical advances are those that allow methods like remote access to be more useful substitutes for public use microdata files. These also include new techniques to reduce opportunities for disclosure. Legal strategies are those that would provide shared data protection responsibility with the user or would severely penalize anyone conducting data linkages to identify individuals in Census Bureau data. Policy enhancements consist of formal guidelines related to confidentiality and privacy aspects of collecting data, controlling access to data, linking data, and providing data for research uses. Interagency coordination takes a Government system-wide view of disclosure risk, since the biggest threat to public use microdata tends to come from administrative data maintained by other government agencies. Finally, we need to pursue research in how to improve communication of our confidentiality procedures in a way that bolsters the confidence of our respondents (rather than calls their attention to an unrealistic potential for misuse of the data). While we consider how best to pursue each of these areas, we cannot lose momentum in the core approaches that depend on disclosure avoidance techniques.

II. DISCLOSURE AVOIDANCE TECHNIQUES: OVERCOMING NEW THREATS

7. The usual review and approval of a release of new microdata sets requires judgments by reviewers based on, among other things:

- the size of the geographic entity—either directly identified by the Census Bureau or indirectly identified by contextual variables (such as sampling information, area mean income, population density, or percent minority population);
- the proportion of the study population included in the sample;
- the sensitivity of individual data items;
- the age of the data.

8. Notwithstanding the fact that released data contain no direct identifiers (such as name, address, telephone number, social security number), statistical disclosure limitation (SDL) experts recognize that the release of "truly safe" microdata (or raw individual data records) is extremely difficult. Data releases do not preclude, by all means, the disclosure of the individual respondent's identity. However, data are released in such a way that attempts at re-identification would require investments in manpower, time, and other costs that would be unreasonably high. In light of rapid changes in the technological and data environment, there may be an increased risk that a data user could match microdata records to another file containing identifiable information with reasonable accuracy—leading to the discovery of identities or of sensitive information. To better understand these types of elevated risks of disclosure, the Census Bureau conducts re-identification experiments to attempt matching files with overlapping information.

9. Re-identification experiments can shed additional light on the particularities of a microdata set. Hence, before the Census Bureau releases a microdata set, the Disclosure Review Board may decide to consider some additional information on the nature of the data file. The information includes:

- the number and distribution of unique records;
- the amount of error in the data;
- the availability of external files with comparable data content[3];
- the resources that may be needed by an "attacker" to identify individual units.

10. Experience in re-identifying respondents from de-identified microdata sets show that the experiments should be run on a periodic basis to continually update SDL strategies. This is especially true for microdata sets published from recurrent large-scale sample surveys. Re-identification research is only one of the research areas the Census Bureau relies on to update SDL strategies. Research areas also target other aspects of dealing with disclosure risk such as measuring the risk, modifying of the data, and releasing synthetic (not observed) data.

11. Measuring disclosure risk for a microdata set usually entails the study of unique combinations of values in the data, and an assessment of whether an intruder can infer whether

[3] All forms of public or proprietary external files are considered: other microdata files, macrodata files (or tabular data), and databases allowing queries of microdata records.

given sample unique records are also population unique (Bethlehem, Keller, and Pannekoek, 1990; Feinberg and Makov 1998; Skinner and Elliot, 2002; Skinner and Holmes, 1998; Zayatz, 1991). Most work in this area assumes that there are no measurement errors in the data and that sub-sampling and other aspects of data releases are often not sufficient to protect against disclosure. Once records at risk of disclosure are identified, or a measure of disclosure risk for the entire file is calculated, traditional SDL strategies center on reducing the amount of information released. The Census Bureau considers statistical data as a public good and, therefore, does not want to rely on this as the best response to disclosure risk.

12. Methods of modifying the data include data swapping (Willenberg and de Waal, 2001) and adding noise (Kim, 1986). Records or blocks of records that are unique in their geographic area are sometimes swapped with partnered records or blocks of records that have identical characteristics but are in different geographic locations. The proportion of records that are swapped has a direct affect on the quality of the data. The Census Bureau modifies quantitative data—such as dollar amounts, travel time and dates—by adding small random quantities or noise, without affecting certain characteristics of the distributions of the original data. However, it is not possible to guarantee that the results of all analyses that can be done using the original data are reproducible using the perturbed data.

13. An alternative to releasing confidential observed data is the release of fabricated or synthetic data (Raghunathan, Reiter, Rubin, 2003, and Abowd, Woodcock in Doyle et.al 2001). The obvious advantage of this method is that releasing entirely simulated data guarantees protection of respondents' confidentiality. One drawback is that the quality of inferences from the synthetic data depends on the imputation models. The research in this area follows earlier, related but different, research efforts on masking microdata (Cox, 1994) to preserve confidentiality.

III. RESTRICTED ACCESS: THE CENSUS BUREAU'S CENTER FOR ECONOMIC STUDIES AND ITS RESEARCH DATA CENTERS[4]

14. Several modes exist for providing restricted access to confidential data while limiting the risk of their disclosure. The Census Bureau has adopted (and pioneered) Research Data Centers (RDCs). RDCs permit restricted use of confidential files at secure sites under Census Bureau control, using limited access to dedicated computing equipment and enhanced physical and computer security.

15. Protecting the confidentiality of the data and ensuring their appropriate use are paramount in establishing and operating RDCs. To accomplish this requires several activities: providing physically secure offices and secure computer systems; selecting projects that use the data appropriately, benefit Census Bureau programs (as required by law), and present low disclosure risks; imparting to researchers at the RDC the Census Bureau "culture of confidentiality;" putting in place policies and procedures that protect confidentiality in the RDC office; and releasing only research output that is within the scope of approved projects and that does not reveal confidential

[4] This section is adapted from portions of a report from the Confidentiality and Data Access Committee of the Federal Committee on Statistical Methodology. A copy of that report, Restricted Access Procedures, is available on the Internet at this address: <http://www.fcsm.gov/committees/cdac/cdacra9.pdf>.

information.

16. Each RDC has a security plan developed and approved according to established Census Bureau procedures. The RDC office is in a restricted access environment with locks and key cards that meet Census Bureau specifications. In response to increasing concerns about security (and to promote efficiency), the Census Bureau RDC system is now completing conversion from secure local RDC networks of PCs and Unix workstations to a centralized "thin client" environment. Under this arrangement, data are stored on secure servers at the Census Bureau headquarters. The RDCs are connected to the servers via dedicated T-1 lines. From the RDC offices, researchers use X-terminals ("thin clients") to access the data authorized for their projects. No confidential data are stored at the RDCs. Researchers are accountable for their computer use, through the use of passwords and system logs. Researchers have no access to any non-Census Bureau network (including the Internet) from within the RDC facility. They may not bring laptop computers or other portable mass storage devices into the RDC facility.

17. Access to an RDC facility is given only to Census Bureau employees or other persons with special sworn status (SSS) who are approved to use the facility—including researchers carrying out active, approved projects at the RDC. To be granted SSS, any researcher must have an approved project, must obtain a security clearance, and must sign the Census Bureau's standard sworn agreement to preserve the confidentiality of the data. Researchers are given access only to the confidential data needed for their approved projects. Persons with SSS are subject to the same legal penalties for revealing confidential information as are regular Census Bureau employees—up to a $250,000 fine or five years in prison. Another equally important legal requirement for SSS is that the researcher's project must benefit the Census Bureau's data programs. The Center for Economic Studies and its RDC partners have set up a formal project selection process to ensure that all approved projects satisfy these requirements.[5]

18. The Census Bureau stations a Center for Economic Studies' employee (the RDC administrator) at each RDC. Among the administrator's most important duties are to instill the Census Bureau's "culture of confidentiality" into the researchers and to train the researchers regarding the security and confidentiality restrictions. The administrator also examines any research output a researcher wishes to remove from the secure facilities – to ensure that the output is covered under the approved project and to prevent the release of confidential data. This examination of research output is called disclosure analysis. In carrying out disclosure analysis, the administrators use disclosure avoidance techniques.

IV. PERCEPTIONS OF CONFIDENTIALITY: THE LURKING THREAT TO MICRODATA

19. Beyond the quantifiable threats to microdata from intruder attacks and security breaches lies the little understood—but no less important—field of public perception (see Gates, 2001). Data collectors must not only be confident in their ability to protect data from determined intruders, but must also be confident that the public believes the collectors have taken all necessary precautions. In the past, the public (in its role as survey participant) was mostly

[5] For more details on the project selection process, see the CES Web site: <http://www.ces.census.gov>.

unaware of who used the survey results and how they used them. Today, with our ability to make data easily accessible to the masses through the Internet, the survey participant has become the survey user. That fact, combined with advances in data mining and data fusion methodology, creates a real risk that the public will not support the data access approaches that have served so well in the past. Our challenge is to ensure that the data we release are clearly labeled for what they are and what they are not.

20. As a result of declining mail response in the 1990 census, the U.S. Census Bureau has been concerned that individuals' concerns for privacy may be playing an increasing role in their decision to provide information in our census and surveys. Census Bureau surveys of public attitudes have attempted to measure what the public knows and thinks about our legal requirements and our practices. We have found that the majority of the U.S. population does not believe we keep their personal information confidential—even though we have legal requirements to do so and strongly convey this message to all potential survey participants (Gates and Bolton, 1998). The extent to which attitudes will ultimately influence an individual's decision to participate in a survey is not well understood. Nevertheless, just as we cannot take a risk that our data products are vulnerable to attack, we cannot take the risk that misunderstandings about data access and protection procedures will cause respondents and potential respondents not to respond to our survey.

21. Some examples of possible misperceptions that could result from new access tools and methodologies include:

- "finding oneself" on a public use microdata file (a relatively easy matter);
- questioning the occurrence of a cell of size one or two on a table where data may have been swapped or perturbed;
- being able to use published data to isolate and profile sensitive population groups;
- learning that data miners can combine data from diverse sources with new technology and methodological tools;
- questioning the agency's commitment to confidentiality when researchers are permitted access under special agreements.

22. These examples can potentially lead to negative reactions and signal the need to better understand how activities that seem so reasonable and appropriate may create misunderstandings. Once we understand these concerns, we need to develop education and awareness programs to address them. Fortunately, we have new avenues to interact with the public. In the past, our only contact came at the time of interview. We have always provided our respondents with basic information on our authority to collect the data, the purpose and uses for the information, and our pledge to keep the information confidential. Today, we have reestablished contact with the respondent in his new role as data user. That fact creates both the problem and the solution.

23. By way of the Census Bureau's Internet dissemination tool, the American FactFinder, we reach millions of novice data users who now can access the entire decennial census data files and request tables of their choosing. The process is fast, easy, and free. Our challenge is to take this opportunity to reinforce the messages we provided at the time of collection and to address any misperceptions that may arise. With this technology, we can target the messages to specific users

and their specific concerns by providing general information (at first) and progressing to more specific details (if desired).

24. The challenge is not so much in how to deliver the message, but rather in what messages to convey. The Census Bureau has approached this in two ways: public opinion surveys and cognitive research. Through public opinion surveys, we have learned what relevant beliefs about privacy and confidentiality are most widely held. Since attitudes are affected by personal experiences and societal events, it is not sufficient to measure attitudes at only one point in time. Surveys need to be conducted periodically and trends monitored. Results will identify key areas of concern that may translate into changes in behavior (for example, reluctance to participate in surveys).

25. Armed with this information, we are able to develop and cognitively test messages that are clear, understandable and relevant. As research has shown, what may be intuitively appropriate is not always the best option. For instance, work done by Singer shows that overemphasizing the confidentiality promise at the time of data collection can have the unintended consequence of raising concerns that were previously not expressed (Singer, Hippler, Swartz, 1993). Cognitive interviewing and focus groups will offer insights into where these perceptions lie and how to best alleviate them.

V. DATA STEWARDSHIP APPROACHES TO CONFIDENTIALITY AND DATA ACCESS

26. In the last few years, the Census Bureau has introduced a data stewardship approach to making decisions about how to collect and provide useful data: balancing data quality and access on one side of the scale and privacy and confidentiality on the other. The concept of "stewardship" is borrowed from environmentalists, the objective being to create a sustainable balance that supports one's needs over the long term.

27. In June 2001, the Census Bureau established the Data Stewardship Executive Policy (DSEP) Committee. The DSEP Committee is composed of top agency executives who are charged with identifying and developing policies related to data stewardship. This executive decision-making body is staffed by the Policy Office and supported by the analyses and recommendations of four staff committees, including the Disclosure Review Board (Potok and Gates, forthcoming).

28. One goal of the DSEP Committee is to ensure that strategic goals, corporate ethics, policies, controls, and operational practices are integrated and consistent. This means that strategic goals are shaped by corporate ethics and drive policies. Policies in turn drive the creation of organizational controls, and these controls incorporate practices that ensure compliance.

29. The Census Bureau has considered a number of sources for guidance in strengthening its data stewardship approach. We conducted a benchmarking exercise, a literature review, and an evaluation of the DSEP structure; and we drew on a U.S. General Accounting Office report published in 2001. From these sources, we gained an understanding of four pillars needed to

strengthen our data stewardship program:

- culture and tradition;
- awareness and outreach;
- an integrating authority, such as a Chief Privacy Officer;
- technical and administrative tools.

30. The final item includes providing safe settings (such as RDCs), releasing safe data (by applying disclosure avoidance methodologies), as well as introducing automated tools that restrict access and limit uses within the organization. Finally, it includes ongoing research to ensure that these tools remain up-to-date.

31. At this writing, the Census Bureau is deliberately working toward full implementation of an enhanced data stewardship framework, based on the four pillars listed above. In so doing, the Census Bureau is also responding to new U. S. Office Management and Budget requirements for privacy impact assessments. These requirements offer an opportunity to integrate principles and policies into ongoing reviews throughout the lifecycle of data collections and supporting systems—allowing proactive planning to minimize risks (including those that are disclosure related). A key component for these assessments will be to build on a set of four privacy principles and sub principles that the Census Bureau identified as the ethical basis for the data stewardship structure. The principles cover mission necessity, informed consent, protection from unwarranted intrusion and confidentiality.

32. It is important to note that developing and maintaining a viable data stewardship structure requires a significant commitment and investment of resources from an agency. Nevertheless, this more structured approach to data stewardship is integral to striking a balance between the tensions inherent in meeting data user needs and honoring the privacy and confidentiality commitments to its respondents. In the end, privacy and confidentiality—which are typically perceived as business constraints—can actually enable an agency's mission and business objectives by establishing the public's trust and cooperation as respondents.

REFERENCES

Abowd, M.J. and D.S. Woodcock (2001). Disclosure Limitation in Longitudinal Linked Data, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), Amsterdam: Elsevier Science B. V., 215-277.

Bethlehem, J.G., W.J. Keller, and J. Pannekoek (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, Alexandria, VA: American Statistical Association, 85: 38–45.

Cox, L. (1994). Matrix Masking Methods for Disclosure Limitation in Microdata. Survey Methodology, Ottawa: Statistics Canada, 20, 165-169.

Doyle, P., J.I. Lane, J.J.M. Theeuwes, and L.M. Zayatz, eds. (2001). Confidentiality, Disclosure

and Data Access: Theory and Practical Applications for Statistical Agencies, Amsterdam: Elsevier Science B. V.

Fienberg, S.E., and U.E. Makov (1998). Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data, Journal of Official Statistics, Stockholm: Statistics Sweden, 14: 385–397.

Gates, G. (2001). A Holistic Approach to Confidentiality Assurance in Statistical Data. Statistical Journal of the United Nations Economic Commission for Europe, United Nations, 18: 299–307.

Gates, G., and D. Bolton (1998). Privacy Research Involving Expanded Statistical Use of Administrative Records, Proceedings of the Government Statistics and Social Statistics Sections of the American Statistical Association, Alexandria, VA: American Statistical Association: 203–208.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, Proceedings of the Section on Survey Research Methods, Arlington, VA: American Statistical Association: 370–374.

Potok, N., and G. Gates (forthcoming 2003). Federal Committee on Statistical Methodology. Statistical Policy Working Paper 35, Washington, DC: U.S. Office of Management and Budget.

Raghunathan, E.T., P.J. Reiter and B.D. Rubin (2003). Multiple Imputation for Statistical Disclosure Limitation, research report, Washington, DC: U.S. Census Bureau.

Singer, E., H. Hippler and N. Swartz (1993). The Impact of Privacy and Confidentaliity Concerns on Survey Participation, Public Opinion Quarterly, 4: 256-268.

Skinner, C.J., and M.J. Elliot (2002). A Measure of Disclosure Risk for Microdata, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64: 855–867.

Skinner, C.J., and D.J. Holmes (1998). Estimating the Re-identification Risk Per Record in Microdata, Journal of Official Statistics, Stockholm: Statistics Sweden, 14: 361–372.

U.S. General Accounting Office (2001). Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information, Report Number, Washington, DC: GAO-01-126SP.

Willenborg, L., and T. de Waal (2001). Elements of Statistical Disclosure Control, New York: Springer.

Zayatz L (1991). Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample. Statistical Research Division Report Number: Census/SRD/RR-91/08, Washington, DC: U.S. Census Bureau, <http://www.census.gov/srd/papers/pdf/rr91-08.pdf>.