Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 595–598

Discussion
Statutes and Administrative Procedures

Harold W. Watts[1]

1. Introduction

The papers in this volume have ably considered disclosure limitation and data access in terms of philosophy, psychology, statistics, computer systems, and law. The inevitable conflict between productive use of microdata and fulfilling the promises made to survey respondents has been clearly marked out, and the current terms of mediation between these objectives is what is termed the status quo. My task is to consider other possible environments for working out the compromise between confidentiality and access. I urge a major change in the locus of responsibility for maintaining the anonymity of survey responses that could make the risk of disclosure less sensitive to changes in technology and the expansion of external data bases.

As an academically based researcher it must be clear that I am interested in maximizing research access to microdata files that are of genuine scientific interest. Obviously, I see my own analytic intentions as benign and in no way dependent on violating the anonymity of individual respondents – citizens or organizations. And I believe that viewpoint is widely shared, if not unanimous, among my colleagues who analyze microdata collected by public agencies. But I also recognize that the continuing feasibility of data collection is dependent on public cooperation, which is in turn, related to public perceptions of the integrity behind pledges of confidentiality.

With that disclosure of my own orientation, I comment first on what is encouraging about the status quo, relative to my first experiences with public microdata access. Second I consider how existing trends might be reinforced or augmented. Finally, I offer some more general remarks and questions about the determinants of public cooperation.

2. The Status Quo Is Not Bad

There has been an enormous change over the past 25 years in the availability of microdata. I recall the early struggles to make a public use file from the 1963 Survey of Economic Opportunity, and what a wrenching departure that was from the traditional way of doing things at the U.S. Census Bureau. Partly due to technological change, which has been breathtaking, and partly due to change and accommodation in agency practice, there is now a great deal of (masked and censored) microdata available for general use. There is also an enormously expanded capacity to use (or misuse) such data, both in terms of computing power and statistical expertise.

Public use files, with top coding to mask outliers and severe censoring of geographi-

[1] Columbia University, Department of Economics, Room 1402, New York, N. Y. 10027, U.S.A.

cal codes, have become part of the routine production process for household surveys. These files are increasingly available on portable CD-ROM disks as well as on magnetic tapes. While limited for some analytic purposes, these files have enabled many scholars and analysts to expand our understanding of social and economic conditions and the consequences of a wide range of policies. So far, the public use files have been used to disseminate individual surveys, rather than for combined files based on matches between surveys and other public data such as Social Security Administration or Internal Revenue Service files. It is worth repeating here that no violation of individual anonymity has been discovered or even claimed as a result of this access, despite convincing proofs that it is theoretically possible by combining data bases (many of which are privately disseminated).

Another avenue of access is provided by way of sworn agency affiliation. By assuming the same obligations as regular agency employees, analysts have been granted access (usually on-site) for analysis of uncensored data. The results of such analyses are reviewed for potential disclosure prior to publication or other extra-agency release. In contrast to public use files, this kind of access cannot be secured by anyone who simply has an interest in the data, and the resources to acquire and manipulate them. Some evidence of serious and informed analytical purpose must be shown and some support for computing charges must also be found. In any case the agency has full discretion in allowing or denying access on these terms. Again, to my knowledge, no breach of trust has resulted in loss of anonymity for any respondent.

Remote access to central data bases offers yet another means of enabling researchers access to data that are substantially masked but, for legal or other reasons, cannot be openly released. This has been pioneered by the Luxembourg Income Study, for access to Current Population Survey like files from a number of countries. Screening of users is possible by the applications required for use of the system. Even more important, the analysis programs and their output are reviewed by technicians to detect conscious or inadvertent selection of cases and variables that would permit identification of individuals. Remote access to public use Survey of Income and Program Participation files is also being tried by the Census Bureau, but no access to uncensored files is now contemplated.

The combination of data bases and software on a public use CD-ROM with an obligatory and legally binding pledge not to engage in activities to identify survey respondents is being tried by National Center for Education Statistics. This appears to be a promising approach that may enable the continuation of widespread access to public use files in a context where the growth of external data bases and shrinking computing costs are increasing the risk of identification.

The status quo, it seems to me, offers a substantial amount of access for legitimate research analysts, and there are encouraging movements toward further expansion. Moreover the careful relaxation of earlier prohibitions has not, so far, violated anyone's anonymity and the sense of privacy that provides. It is important to maintain the level of access now enjoyed and expand it where possible in a context that threatens more restrictions and even more severely censored and otherwise impaired data releases.

3. How Could It Be Made Better?

It is time to acknowledge that no data base can be made impervious to determined

efforts to identify (some) individuals. On the other hand, much can be and has been done to make that kind of activity unattractive relative to more direct spying and ferreting. Efforts to maintain the difficulty of breaking into the data files while minimizing the loss of information through masking and censoring should be continued. But the goal of a completely ersatz data set with no possibility of linking *true* information with a real respondent, as proposed by Rubin, seems to me too distant and not so satisfactory even when possible. Briefly, I am not sure that the half-truths and misidentifications that could be drawn from such files would be any less damaging to the respondent or the agency even though the culpability could be legally evaded.

I am encouraged by the interest of lawyers in these topics. Eight to ten years ago, lawyers showed little interest and there was little recognition that they might be useful. My optimism comes from what seems to be a growing recognition that the responsibility for protecting the anonymity of persons or organizations should be shared between the agency or agencies that collect or assemble the data and the analyst who is granted access to those data.

Such a change in the distribution of responsibility should be accompanied with an altered pledge to respondents. Such a pledge would still assure a priority for maintaining their anonymity and freedom from prejudicial use of freely given private information. It would add that authorized users of name-and-address-free data are subject to criminal penalties for violations of security procedures and to civil penalties for harm to individuals whose anonymity is violated.

This approach is being tried by the University of Michigan and Ohio State University for files that include relatively fine-grained geographical detail, and their experience should be exploited in designing new procedures for government agencies. The reward is that the nation will get much more value out of the human capital invested in scholars and policy analysts if they are enabled to use more fully the valuable and costly data resources that exist in government agencies.

We do not allow citizens to use automobiles without liability for possible harm to others. Despite the best efforts of consumer activists, we have not produced an automobile incapable of harm, so we strive to make cars safer and require that they be used safely. The same strategy is more realistic for sensitive data than the attempt to make it non-sensitive.

4. Public Cooperation versus Data Availability

There is, to be sure, some relation between the amount and kind of access allowed research analysts and the willingness of citizens to cooperate with official (and unofficial) data collection efforts. But it is easy to overlook the long list of *other* determinants of that state, as well as the several steps involved in the trade-off between access and response rate.

It is, first of all, the *perceived* integrity and benign intentions of the agency that will weigh heavily in a citizen's decision to submit to an interview. Those perceptions and beliefs could be damaged by a well-publicized instance of individual identification of a survey respondent. But the consequence would probably be just about as bad if the accusation was false and there was a predisposition to distrust the specific agency or government in general. On the other hand, an agency (or government) that enjoyed a good reputation for competence and integrity might find that an isolated and unavoidable (perhaps punishable) identification would have only a marginal and temporary effect on response rates.

I think more effort is needed to draw a sharp line between data collected for administrative and enforcement purposes, and those that are collected *or* assembled (possibly from administrative records) for statistical description or analytical purposes. The latter should be given full protection of anonymity and prohibition of use for administrative and enforcement purposes, with criminal penalties and civil remedies for violation. They also should be immune from both judicial and Congressional subpoena. Such policies, if implemented and understood by the general public, many of whom now think data swapping is routine, would do much more for response rates than stopping *all* access to census files by outside analysts.

Speaking personally, I have not yet been included in a government sample survey, so I am not certain how I would respond. I am, however, frequently pestered by nongovernmental surveys, some of which are part of commercial marketing endeavors. If I were contacted by a census interviewer immediately after exposure to marketing surveys, I suspect my willingness would be somewhat impaired.

Perhaps a determined effort to improve response rates should consider the extent to which the public's capacity to tolerate intrusion is strained. Maybe the Department of Commerce should ask whether commercial survey and canvassing efforts are damaging the ability of government surveys to operate successfully.