CONSENT, MATCHING, AND RELEASE FOR PUBLICLY COLLECTED DATA

William P. Butz and Gerald W. Gates, U.S. Bureau of the Census*

Two years ago at these meetings I cautioned against greatly expanded exact match linkages of records from surveys, censuses and administrative data, for statistical purposes. I argued on technical, organizational and perceptual grounds. The issues for today's panel are related: What matching should be done with data? What data should be released? What should respondents be told about all this?

These remarks concern only data on individual respondents and their families and households, as opposed to institutional respondents like business firms and governmental units. Further, I will focus on the Census Bureau's practices concerning data collected under Title 13 of the U.S. Code, which authorizes most of the household data we collect. I begin with four organizing questions about data consent, linkage, and release, and end with four derivative issues.

Here is the first question: What uses can legally be made of information collected from people about themselves or their family members, other than the uses for which the information was explicitly gathered and about which the respondent was informed? Title 13 sets the legal limits. It says that information provided by a respondent cannot be used for any purpose other than the statistical purposes for which it is supplied.1/ This information must be held confidential and cannot be released to any individual or government agency in a form that would allow the identification of any individual.

Uses of existing records or other material available from other Federal agencies, the states or private persons are authorized "as may be required for the efficient and economical conduct of the censuses and surveys provided for in (Title 13)." In addition to this discretionary authority, the statute directs that, "To the maximum extent possible and consistent with the kind, timeliness, quality and scope of the statistics required, the Secretary (of Commerce) shall acquire and use information available from any source referred to in ... this section instead of conducting direct inquiries." Information acquired from these sources is provided the same protections as the information collected by the Census Bureau directly from respondents.

In addition to the Title 13 restrictions on the use of individual information, the Privacy Act of 1974 states that respondents are entitled to be notified of the principal purpose of the data collection; whether response to the survey is mandatory or voluntary; what penalties are imposed for not responding; any routine uses to be made of the data; and the legal authority to collect the data.

Together, these statutes provide rights and protections to respondents that are by any historical or contemporary standard extraordinary. With two qualifications, which I will expand on later, the Census Bureau operates within these limits. One qualification is that it may not be possible to issue data from which it is absolutely impossible for anyone ever to identify an individual. Other available data and analytical, computational, and statistical tools raise the probability of individual identification above zero. The second qualification concerns the meaning of "statistical purposes."

The second of my four questions is: What uses SHOULD be made of information collected from people, other than those uses for which the information was explicitly gathered? Frequently data, once gathered, are found to be valuable for statistical analyses other than the ones the data collector originally intended. Not to make use of this valuable information would limit important research and may result in the need for separate data collection with its added burden on respondents and cost to the government.

Each of these analyses will be affected, and often improved, by the steps taken to prepare the data for their ultimate release, either in tabulations or public use microdata files. What the Census Bureau does with these data is conceptually straightforward. We check them for internal consistency and completeness, and we sometimes change the reported values to correct omissions or obvious inconsistencies, so that the data can be used for the statistical purposes that motivated their collection. We also may use administrative records of other agencies to evaluate the quality of the data reported and determine the extent of coverage of the population being studied. We then distribute aggregated statistical analyses of the data, as well as the microdata tapes containing individual information, without identifiers, from many specific respondents. We take serious measures to delete information from these tapes that could allow a user to identify a particular person or family. A formal review procedure is in place to ensure that these tapes meet certain guidelines designed to protect the respondents' confidentiality.

While this is conceptually straightforward, it is usually complicated in practice. For example, the steps of correcting incomplete responses often entail inserting a "best guess" value into the record. Such information does not come from some other source referring to this individual; instead, it refers to other individuals "like" this one in relevant respects. It is an educated guess, but one that makes all the information useful for its intended purposes.

Sometimes, research involves comparing respondents' answers with information from other data sources, in order to see, on the average, which questions elicit more accurate data. Knowing this information is important for those who do the statistical analyses that will fulfill the data's purposes. It is also important for the Census Bureau's efforts to improve data quality in subsequent collections. In these research evaluations, Census Bureau employees study and compare data about individual respondents, but the information that emerges describes only large groups of people. The most important point is that all this work serves the ultimate statistical uses that are to be made of the data. It prepares the data for these uses; it is not independent of them.

This practice is consistent with a guideline in the "Declaration of Professional Ethics" of the International Statistical Institute to make use of available data rather than embarking on a new inquiry. The declaration notes that: "Although some subjects may have objections to the data being used for a different purpose from that intended, they would not be adversely affected by such uses provided that their identities are protected and that the purpose is statistical, not administrative."[2/]

All of our procedures and uses for the data are strictly for statistical purposes. Most of these uses are well described by the examples given to respondents. Some uses differ, though they are still for statistical purposes. It is a fortunate characteristic of science that useful investigations sometimes end up far afield from where their original direction pointed; otherwise, much more data would have to be gathered from respondents and scientific progress on important policy questions would be slower.

My third question is: What do we tell survey respondents about the purposes and uses of the information they give? We tell all respondents that the information they provide will be used only for statistical purposes in a manner in which no information about them can be identified. Other statements given to respondents in a survey or census usually cover the intended uses of the data in general, but include several specific examples as well. This is illustrated by the following statements from the letter sent to prospective Current Population Survey respondents: "This survey provides the official Government figures on employment and unemployment issued each month....Our reason...is to find out what changes have occurred in employment, family size, school enrollment, and other important subjects." An additional example is contained in the brochure given to respondents in the Survey of Income and Program Participation: "The survey provides information on a wide variety of topics relating to the economic status of Americans... data on the types of jobs and other sources of income that people have, as well as the number and characteristics of persons who participate in various Government programs. These types of information will help in evaluating the economic status of the United States, show how things change over time, and help policymakers make better economic decisions."

In addition to informing respondents about purposes and uses of the data, we also tell them something about how we will prepare the data for these uses. For example: "We will combine data from the SIPP with data from other Government agencies to provide a comprehensive set of summary information about employment, income, and participation in various Government programs." This statement refers to our uses of other data both to correct for incompleteness and inconsistency, and to complete the statistical picture concerning aspects not asked about in this particular survey. A similar statement is now used in the CPS respondent letter mailed to each household.

We also tell respondents whether their cooperation is obligatory or voluntary. And we tell them that we protect the confidentiality of the information they give. Finally, in our interviewer-administered surveys if a respondent wants additional information on any of these topics, our interviewers give them more specific details and, if necessary, provide phone numbers where they can call for even more. Interviewers also carry recent news clippings containing results of the survey, to give respondents examples of the data in actual use.

The fourth question is: What SHOULD respondents be told about how the data they give us will be used? Even though we do not know everything beforehand about what we or others will do with the data, we certainly know many more details then we print in the information for respondents. It would require a long report to detail all the various statistical uses and procedures for any one survey, and it would be impossibly difficult to do this for the decennial census. Different respondents might like to know different things in order to make an informed response. In order to decide what to say, we are forced to decide what most respondents would like to know. We choose to focus on telling respondents why their participation is important and how the government and other researchers will benefit by the results. To provide much more, especially regarding complex procedures, would lose many respondents' attention and interest. When our standard informational practices fail to provide all that a particular respondent wishes, extra information is available from the interviewer or from Census Bureau officials.

While still meeting the requirements of the Privacy Act, we could provide either more or less information as a matter of course. Just as we do not detail all the ultimate uses to which the data might be put, we also do not detail our procedures for studying and improving the quality of the data. These procedures contribute to the usefulness of the data; they are not themselves a separate use. We continually assess how much information to provide about these matters and change it from time to time as the regular uses of data change, and based on suggestions from our interviewers and questions and comments from our respondents.

These four questions give rise to several issues that merit continuing discussion and assessment. Here are four:

1. Are the microdata files disseminated by the Census Bureau too restricted or not restricted enough? This issue surfaces one of the fundamental tradeoffs we face. On one hand, our job is to inform the American people about themselves and their institutions. We do this by giving back as much of the information originally given to us as we can. Government agencies and other data users legitimately seem to want more and more detail for their statistical analyses. On the other hand, we must always protect the confidentiality of information about individuals. With continuing innovations in computing and statistical procedures, and with increasing availability of data on individuals from other sources, outside users may now have more ability than before to identify individuals from data sources. We deal with this conflict between being a provider and protector of data by staying aware of developments in each area and, where there is a difficult choice, opting for greater care in pro-

tecting confidentiality.

2. What is meant by "statistical purposes" in translating the intent of Title 13 into practice? I would restate the concern enunciated in my remarks at these meetings two years ago: "An oft-stated principle is that data collected for administrative purposes should be available for statistical purposes, but data collected for statistical purposes should not be used for any administrative purpose. Certainly this is true, but it may be too general to be useful...This danger is particularly clear when program agencies and private organizations can combine aggregated statistical data with their own administrative records to estimate characteristics of specific groups of individuals and to identify outlying individuals for review of compliance or for directed marketing attention...Indeed, just what it is that actually is being kept confidential is a distinction that is becoming more difficult." The problem has not abated in these two years.

3. How much information should be given to respondents about the uses to be made of other data about them, and at what point in the process should the information be given? I have already given my opinion that additional information is unnecessary if the procedures serve the stated purposes of the data we collect and if they protect the respondent's identity.

The Privacy Act requires that agencies publish annually in the Federal Register each "routine use" of the records contained in their system of records, including the categories of users and the purpose of such use. This only applies, however, when individual data collected for a specific administrative purpose may be used by the agency that collects the data or by another agency for purposes other than for which they were explicitly collected. Data collected by the Census Bureau are not considered to have "routine uses" since they are not disclosed to other agencies in individually identifiable form and cannot be used for any administrative purpose.

If an agency should decide that individuals need information on uses of the data, beyond that provided under the Privacy Act, I believe this information should be provided at the source of the other data, by the agency that collects them.

4. If a respondent refuses to participate in a survey or to answer a particular question, should the Census Bureau substitute other information about that person, if it is available from other sources? We do not do this. It is sometimes possible, and doing so would undoubtedly make the total information from a survey more useful for its stated purposes. It may be in the future that the usefulness of making this kind of data insertion will grow to the point that we will seriously consider it. If so, we will also have to decide what to tell the respondents about this procedure and how to accommodate any objections to it. If these issues emerge for choice, we must confront them directly and seriously.

To conclude, I think that the Census Bureau's uses of household survey data are appropriate. I also think that what we tell respondents about these uses is close to the mark. Both are within the legal constraints and, in my opinion, ethically proper. At the same time, it is also my opinion that the Census Bureau should continue to assess its data linking, data release, and respondent information policies in light of changing technical possibilities and ethical sensitivities. That is to say, we must continue to assess the tradeoffs between informing Americans statistically about themselves, on the one hand, and protecting their privacy, on the other.

NOTES AND REFERENCES

1/ In our frequent reference to statistical vs. non-statistical uses of data, two important points need to be made: 1) statistical uses are independent of an individual's identity whereas non-statistical (administrative) uses cannot be accomplished without some form of identification; and (2) there is no commingling of statistical and administrative functions by the Census Bureau--we do not use individual information, from whatever source, for administrative purposes.

2/ "Declaration on Professional Ethics," International Statistical Review, Vol. 54, No. 2, Page 227-242, August 1986.

*Authors are Associate Director for Demographic Fields and Administrative Records Program Officer, U.S. Bureau of the Census. These remarks were delivered by William Butz at a panel on "Ethical Issues in Access and Linkage of Publicly Collected Data" at the Annual Meetings of the American Statistical Association, August 17, 1986, in Chicago.