THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

CENSUS BUREAU MICRODATA: PROVIDING USEFUL RESEARCH DATA WHILE PROTECTING THE ANONYMITY OF RESPONDENTS

No. 76

G. Gates
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

This paper was prepared for presentation at the annual meeting of the American Statistical Association in New Orleans, August 22-25, 1988. The views expressed are the author(s) and do not necessarily reflect those of the Census Bureau.

TABLES OF CONTENTS

INTRODUCTION..... 1

LEGAL CONSIDERATIONS..... 4

PUBLIC USE SOLUTIONS..... 4

The Microdata Review Process..... 4

Research on Microdata Disclosure
Risk and Reduction..... 6

Public Use Alternatives to Microdata..... 8

ADMINISTRATIVE SOLUTIONS..... 10

Non-Title 13 Surveys..... 11

Use of Special Sworn Employees..... 12

LEGAL OPTIONS..... 15

CONCLUSIONS..... 18

REFERENCES

Table 1. Objectives Of The ASA/NSF/Census
Research Program.....22-23

Table 2. Census Bureau Regional Offices..... 24

INTRODUCTION

The U. S. Census Bureau has provided public use microdata as a component of its decennial census data products since 1963 when we released a one-in-one-thousand sample file for the 1960 Decennial Census. Since then, microdata files have become an integral part of our decennial census and demographic surveys programs. As a result, researchers in other Government agencies and research institutes have been able to conduct important policy and planning studies that could not be answered through the use of published tabulations. Were it not for public use microdata files, the only way these studies could be done, if at all, would be by contracting with the Census Bureau for special tabulations. This is not the preferred solution for several reasons. First, these special requests are totally dependent on programming and computer support that is committed to routine Census work. Therefore, the time required to complete the work does not always satisfy user needs. Second, statistical analyses do not always turn out the way researchers intended. They may want to change the variables or the analytical methods after they see the initial results. Finally, in contrast to the costs of using available staff and micro-computers, the costs of using Census Bureau main-frame computers and programmers may exceed the available resources for the project.

The advent of public use files has eliminated many of these problems but has introduced some new ones both for the researcher and for the Census Bureau. Because of the flexibility available when using microdata files, the broad access to high speed computers, and the increased sophistication of data users, there has been an increased use of this medium in the 1970's and, particularly in the 1980's. With this increased use has come increased demand for detailed information that was excluded from or curtailed on public use files to protect the identity of survey and census respondents. The statute (Title 13) under which the Census Bureau operates requires that when we collect and publish data under this authority we not publish results that can be used to identify a particular respondent. Realizing that it may not be possible to release data from which it is absolutely impossible to identify an individual, we strive to ensure that the risk that the data will be used to identify someone on the file is extremely small. For example, current microdata disclosure protection criteria prevent the release of geographic identifiers for areas with small populations, extreme values for continuous variables, and information that is obtained from

---

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

or matchable to administrative records systems. These restrictions prevent survey sponsors from conducting some analyses, such as certain microsimulations; reprocessing the individual responses; or having their own administrative data appended to the survey results. They also inhibit the potential of surveys we sponsor (for example, the Survey of Income and Program Participation (SIPP)) for program policy research by other Government agencies.

Here are some recent examples of requests for demographic microdata that could not be satisfied because of confidentiality concerns:

- The General Accounting Office requests a file linking SIPP data to Social Security beneficiary records. This file is needed for a study related to a disparity in Social Security benefits between adjacent cohorts of retirees. The information from the Social Security records are match keys that could be used by the SSA to identify SIPP respondents.

- The Economic Research Service of the Department of Agriculture wants a file showing non-metro status of SIPP respondents in order to assess the economic well-being of non-metro residents in terms of their wealth, asset holdings, and participation in Government programs. These non-metro designations, in combination with the geography on the released public use files, reveal areas of fewer than 100,000 persons.

- The National Opinion Research Center (NORC) requests a special 1980 census public use file with records linked to tract and SMSA data. This study, linking people to their immediate neighborhoods (tracts) and the larger area in which they live (SMSA), is part of a three year study of racial segregation in the U.S.. Tracts and some SMSAs contain populations of fewer than 100,000 persons.

- Princeton University requests exact date of birth on a SIPP microdata tape in order to research the Selective Service draft lotteries held in the U.S. in the 1970s. (Lottery numbers were assigned to young men based on birth date.) Since date of birth is available on many administrative records files, it is an excellent match key and an additional risk to identifying SIPP respondents.

- The National Institute on Aging (NIA) wants to conduct a followup interview with respondents to the Longitudinal Retirement History Survey conducted in the 1970s. One condition for funding a followup survey is that a microdata file be made available for research studies supported by the NIA. Such a file would be potentially matchable to administrative records information maintained by the Social Security Administration.

- The NORC would like SSA earnings history data added to a SIPP microdata file to be used as a control group in an evaluation of the Job Training Partnership Act (JPTA) manpower training system. Data for the control group would be used to measure the impact on outcomes such as earnings, labor force participation patterns, and welfare recipiency of the JTPA program relative to a population of non-participants.

- The Bureau of Labor Statistics would like access to finer geography and certain longitudinal matching variables on a Current Population Survey (CPS) public use file. This survey is sponsored jointly by the BLS and the Census Bureau. BLS wants this additional detail in order to conduct statistical research, facilitate longitudinal analysis of the data, and develop small area estimates.

Users of data from the Census Bureau economic surveys and censuses have a more basic problem when it comes to microdata. Namely, the Census Bureau has not released microdata on businesses because of the unique visibility of establishments, the availability of private sector data bases, and the effects such files would have on our ability to produce subsequent special purpose tabulations. Nevertheless, demand continues to grow for public use files on businesses; particularly those relating to the manufacturing sector. For example, the Census Bureau has developed a longitudinal file of manufacturers called the Longitudinal Establishment Data (LED) file. In a conference sponsored by the Census Bureau in 1984, more than 100 economists interested in the LED expressed their desire for a public use LED file. The only alternative they saw--submitting special requests for analyses to the Census Bureau--was totally unacceptable because of the limited utility of releasable products and the timing and cost factors, (Govoni-Waite, 1985).

Aside from the interests of our users, the Bureau of the Census must also be concerned about whether the protections afforded these public use files are sufficient. While high speed computers have made public use files more attractive, they have also increased public concern about potential abuses to individual privacy resulting from the creation of large integrated databases. In recent years, events in West Germany, Sweden and other European countries regarding government databases have highlighted this concern, (Butz, 1985; Dalenius, 1988). Moreover, widely publicized exploits of computer hackers have raised fears that, given enough patience, someone could defeat any scheme designed to protect confidentiality. On top of this we know very little about the true risk of someone identifying a respondent on a public use microdata file. Statisticians are just now beginning to quantify the disclosure risks associated with microdata, (Duncan-Lambert, 1987; Paass, 1985; Spruill, 1983). Perhaps instead of seeking ways to provide more detailed public use microdata we should be looking for alternatives that contain fewer unknowns.

With the growing demand for microdata products that cannot be made public under current guidelines and the lack of an acceptable quantitative measure of disclosure risk, the Bureau has undertaken to find solutions that provide our users with the data they want and our respondents with the data protection assurances they are entitled to. This paper describes our current plans in terms of public use microdata, publicly releasable alternatives to microdata, and administrative arrangements. I describe some applications of these solutions to recent requests. Finally, I discuss legal arrangements that have been recommended as ways of extending the obligation for protecting confidentiality to the users of microdata.

LEGAL CONSIDERATIONS

Release of individual data by the Census Bureau is restricted by Title 13, United States Code. Only sworn officers and employees of the Census Bureau are allowed to examine individual reports furnished under the provisions of this title. As needed, we have the authority by Section 23 to "utilize temporary staff, including employees of Federal, State, or local agencies or instrumentalities, and employees of private organizations to assist (us) in performing the work authorized by this title, but only if such temporary staff is sworn to observe the limitations imposed by Section 9 of this title." Section 9 (a) states that the Census Bureau may not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied" and may not "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

PUBLIC USE SOLUTIONS

Public use microdata are data products the Census Bureau releases for general, unrestricted statistical and nonstatistical use. As a result of our legal requirements, we must ensure that any microdata product we release to the public is anonymous (no individual identifiers) and that it will remain so. Consequently, individual characteristics on the file must be evaluated to determine if they can be employed to uniquely describe an individual in the population from which the sample was selected. This evaluation procedure involves making some assumptions about what external information is available, whether it is accessible, and the amount of effort required to retrieve it. When records are not available, we consider the visibility of persons (that is, things about them that are public knowledge and would be revealed in the file).

The Microdata Review Process

Prior to 1981, the Census Bureau's microdata disclosure reduction criterion consisted of a 250,000 minimum requirement for the population residing in sample areas that represent the finest geographic area to be shown on the file. Additional disclosure

reduction measures were established on a case-by-case basis by the Census Bureau staff responsible for releasing the file. In 1981, other criteria were established, including a new population minimum of 100,000 within sample areas; although a higher minimum could be set if the nature of the file warranted greater restrictions. At this time the Census Bureau also created a Microdata Review Panel (MRP) to review and approve all microdata files prior to release.

The Panel's membership included staff representing the Directorates for Statistical Standards and Methodology, Economic Programs, and Demographic Programs; and the Data Users Services Division and the Program and Policy Development Office. The MRP was given broad authority to require additional masking techniques to reduce disclosure risk. These include data grouping or aggregation, addition of random noise, rounding responses, and in some cases, suppression. In order to allow for a smooth transition and minimize the disruption to current microdata users, files that were released prior to 1981 were not recalled and surveys that were currently in the field were not subject to MRP review. Continuing surveys come under MRP review only after the sample is redesigned, the content of the questionnaire is materially changed or the content of the file is expanded.

A typical microdata review consists of the following steps:

1. The sponsoring Census Bureau division submits a formal request to release a file. This request includes:

   - tables showing population counts in identifiable geographic areas;
   - a description of the survey design, sampling procedures, and weighting scheme;
   - a checklist identifying potential disclosure problems with the file, including the existence of external files (e.g. administrative records) which contain data items similar to the proposed release;
   - proposed solutions to these disclosure problems including topcodes, recodes, and deletions; and
   - a data dictionary or annotated questionnaire for the proposed file indicating which items are to be recoded, topcoded, grouped or suppressed.

2. The MRP meets to review the request taking into consideration:

   - Disclosure reduction requirements imposed on previous releases (if any) from the subject survey.

   - If the survey is longitudinal, whether the proposed geography has been changed from the previous release? If it has, could the current and previous releases be matched on characteristics to reveal areas of fewer than 100,000 persons?
   - What information from the proposed file is available from

external files; including those available to the survey sponsor?

- If the survey sample was drawn from other Census Bureau surveys or censuses, were microdata files released from those programs and what information did they contain?
- The uniqueness and degree of visibility of characteristics on the file in conjunction with the proposed geography (for example, residence in a particular institution).

3. The MRP approves the file for release as proposed; requires specific modifications; suggests possible solutions that the division/sponsor may accept or propose an alternative; or rules that a microdata product is not possible given the requirements of the sponsor.

The decisions of the MRP are partly subjective in that no quantitative measures of disclosure risk are available for each file. The panel members varied backgrounds within the Census Bureau tend to promote a balance in the review process which recognizes the needs of our users while emphasizing our obligations to respondents. In recent years, with increased demands for more detailed geography and administrative data appended to surveys, the Panel's seemingly conservative stance has come under criticism by users.

Research on Microdata Disclosure Risk and Reduction

In order to provide a more scientific approach to evaluating microdata disclosure risk, the Census Bureau has established a permanent staff to conduct research on disclosure risk measurement and reduction, (Greenberg, 1988). This Census Bureau Confidentiality Staff is currently undertaking "reidentification studies" for the Survey of Income and Program Participation and the 1990 Decennial Census sample files. These studies involve measuring (or quantifying) the risk of disclosure (identification of a respondent) and designing methods to reduce this risk. Reidentification studies for the proposed decennial census microdata files will be done using the 1980 Decennial Census five-percent public use microdata file and the entire 1980 Census file. The files will be matched using rules that incorporate knowledge of what information is available on external files. The SIPP study involves a similar investigation with a special focus on the effect of geographic detail on levels of disclosure risk.

A logical extension of this research is a methodological evaluation of various masking techniques. In the early planning stages, this work would involve designing methods to evaluate and optimize the effectiveness of various techniques with respect to reducing disclosure risk and maintaining the statistical utility of the data. The schemes we will look at include: 1) recoding responses into intervals; 2) rounding responses; 3) recoding responses into categories; and 4) adding random noise to the responses. We will evaluate the effectiveness of these techni-

ques to reduce disclosure risk and incorporate them, as necessary, depending on the results obtained in the study of disclosure risk, (Greenberg, 1988).

Some Applications to Demographic Microdata

There have been a few instances where we have developed special purpose masking schemes which involve the introduction of random noise. One case involved a microdata file from the Continuous Longitudinal Manpower Survey (CLMS) which we conduct for the Department of Labor to evaluate the effectiveness of the Comprehensive Employment and Training Act (CETA) of 1973. The public use files from this survey contain earnings data matched from SSA administrative records. Since this survey was in effect prior to 1981, the microdata files had not come under MRP review and had not been subject to the systematic analysis of risks involved with files linked to administrative records. Through the addition of random noise and data transformation, we were able to continue to release public use files that adequately protected the confidentiality of respondents, (Kim, 1986). However, we were not able to provide the full range of income data through these techniques.

On occasion, we have developed masking schemes in response to user requests for special purpose data files. An example is the previously mentioned request from the NORC for census tract characteristics on a 1980 census sample file. The Census Bureau Confidentiality Staff has developed a two part approach to this problem. First, they are developing variance-covariance matrices of the data, along with the means, based on the modeling that NORC has planned (see "Public Use Alternatives to Microdata" below). Also, we will prepare a microdata file containing tract characteristics to which noise has been added in order to reduce the risk of tract identification. This approach was developed in consultation with the NORC who determined that the noise would not unduly affect the utility of the data.

Potential Applications to Economic Microdata

The Census Bureau has recently explored the utility of surrogate public use files, involving data transformations, as a means of releasing sensitive economic microdata. To be useful, these transformed files must preserve the correct estimates of the true economic model; allow the analysis of subsets of the data cross-sectionally and longitudinally; and allow expansion of the file to include new economic variables and a link to outside sources, (McGukin-Nguyen 1988). Two types of transformations have been suggested: 1) stochastic transformations which involve adding random noise to the original data while preserving the mean and variance of the variables and the covariance relationships between variables (Kim 1986); and 2) non-stochastic transformations which provide for the release of the data in ratio form, (Monahan, 1986). Each of these methods has merit but each has limitations with respect to the types of economic research for which it will provide a suitable database. McGukin and Nguyen have described the disclosure issues involved in each of these

types of surrogate files and the usefulness of transformation techniques in providing correct estimators for a particular class of single-equation economic models. They conclude that:

> It is extremely important to develop precise criteria for evaluating the disclosure risk. Without such criteria, evaluating a microdata public use file in terms of disclosure is almost impossible. But, we emphasize that disclosure free files are not enough. Such files must be useful and we think the best hope for developing a public use file lies in focusing on surrogate data files which allow researchers to estimate common economic models. Finally, because current economic analysis often uses multi-equation economic models, further research into transformation techniques should take into account these models as well.

Public Use Alternatives to Microdata

There are occasions where traditional masking techniques do not allow for the release of microlevel information needed by policy makers concerned with both economic and social programs. In some cases the sensitivity of the data (for example, information on businesses) or the amount of masking required will prohibit the release of a useful microdata file. That is, the masking necessary to protect the file will destroy important relationships among the variables in the file. To handle these situations, we are experimenting with the release of data tapes containing summary statistics. In addition, we are considering the development of test files as a means of allowing researchers to interact with the internal microdata without having direct access to the files.

Tabulations of Summary Statistics

One category of products would include tabulations of summary statistics, such as microaggregations, whereby individual records are grouped according to specified criterion variables and responses are replaced with averages for the group, (Wolf, 1988). This approach, which is operationally straight-forward, has been suggested as a way to provide access to economic microlevel information, (Govoni-Waite, 1985). It is not a panacea, however, since certain useful properties of the individual data will be lost. One major area for investigation in this approach is to determine rules for grouping establishments. Some users will not be satisfied with the rules that are chosen and this inflexibility is a major limitation to this approach.

Another summary statistic approach we are considering for more general application is the release of variance-covariance or correlation matrices of the data, (McGukin-Nguyen, 1988). Such files allow the outside user to obtain information needed for

producing linear regression estimates based on the underlying microdata and provide excellent confidentiality protection since any given covariance matrix can derive from an infinite number of data sets. As with all summary statistics, the biggest disadvantage with correlation matrices is that they are relatively inflexible for general statistical use. Different users will require different matrices just as the same user may require new columns in his matrix as the analysis proceeds.

Remote Access or Test File Approach

Another public use alternative which we are considering resembles a procedure used by the Luxembourg Income Study (LIS) to provide worldwide access to the LIS database through a telecommunications network, (Rainwater-Smeeding, 1988). In the case of the LIS, certain databases were loaned to the Study by countries with severe privacy and confidentiality restrictions. Since no public use files were permitted, and due to the cost and inconvenience of traveling to Luxembourg to work with the database, an alternative had to be developed.

The solution involved the use of an electronic file transfer network over which users submit program jobs to be run by LIS staff on the database housed in Luxembourg. This process depends entirely upon a user package created by the LIS staff containing: 1) a technical description of the data file; 2) a description of the variables for each country's file including summary statistics; 3) a codebook; 4) recodes for income definitions; 5) a sample data file containing 200 records from each country; and 6) information on available software packages. With this package, the potential user can plan a study, program tabulations, and determine, to some degree, the utility of products created using the "live" data.

Important considerations for the data provider are 1) the degree of confidentiality protection afforded the test file; 2) the physical separation of the users from the live data through the intervention of the LIS staff; and 3) confidentiality measures applied to the tabular output. Important for the user are: 1) familiarity with required software (SPSSX); 2) the degree to which the test file resembles the complete data file and 3) the time required from submission of jobs to the receipt of output. Regarding the test file, LIS provides live records, without personal identifiers, that contain little or no geographic detail but no additional masking. In the absence of public use files containing geographic identifiers, these records should be relatively anonymous. Jobs that are received are held until released by LIS staff. Once submitted, LIS software checks the programs for consistency. Completed jobs are checked by other software for minimum cell size and to ensure that the individual records are not being transmitted. Turn-around time is not instantaneous but, given that nearly everything is automated, it can generally be measured in hours rather than days.

The application of this approach at the Census Bureau would introduce additional complications. First, the Bureau has a policy of not allowing direct telephone access to its mainframe computers, other than through "dedicated" lines. Even with encryption techniques, use of passwords, and operator intervention, we have concerns about the public perception that computer hackers could get into the live data. A second problem is that if public use files are also created, the test file could potentially be matched to the public use file revealing additional information (data suppressed or modified on the public use file) for those cases on the test file. Also, for unique cases that fall into the test file, removing the geography may not be sufficient to protect the identity of the respondent. Finally, we must be concerned with the possibility that although the individual tabulations are "safe", various combinations, taken together, may reveal unique characteristics about a respondent.

The Census Bureau has recently initiated a Data Resource Center (DRC) for the SIPP which will serve as a testing ground for this approach to disseminating microdata. The DRC was created about two years ago to serve researchers who cannot obtain the data necessary for their analyses from available Census Bureau data products. "It has been designed to serve as both a technical and an administrative link between non-Census Bureau researchers and the data contained on internal Bureau files, especially those files produced from the SIPP data set. Further, it has been charged to coordinate and produce special demographic, social, and economic data sets, tabulations, and analyses for non-Census Bureau researchers and analysts from these files." (Cavanaugh, 1987) Although the Data Resource Center has an ultimate goal of developing useful, and anonymous, test files, so far its primary use has been to provide research files from the SIPP wave data sets. (These research files have been modified to protect confidentiality but have not yet been made public-use because they require further research or evaluation.) Nevertheless, some work has been done on the development of anonymous test files that would be representative of the entire sample. Although much work is required before a Luxembourg-type program is in place at the Census Bureau, the DRC is working with interested analysts to help make it a reality, (Herriot, et.al., 1988).

ADMINISTRATIVE SOLUTIONS

Public use solutions, such as these, will provide benefits for the greatest number of users. They will not satisfy all users and, in particular, may not be the answer for statistical projects funded by other Federal agencies, including our survey sponsors. Many studies requiring the development of models, reprocessing of the raw data, or enhancement with various administrative data sources cannot be done using public use files or summary statistics. The nature of these studies requires use of information that may never be made public use.

Aside from the SIPP, nearly all of the Census Bureau's household surveys are fully or partially sponsored and funded by other Government agencies. The Census Bureau collects and processes the data under a reimbursable agreement and delivers a data product to sponsors (tabulations and/or public use microdata files). Under Title 13, survey sponsors are treated just like other non-Census Bureau employees and are not entitled to see individual records from the surveys they fund. This has presented problems for some of our sponsors--who in fact are primary survey users--and makes it more difficult to fulfill our mission to provide statistical information to a wide variety of users.

Non-Title 13 Surveys

Before 1976, Title 13 did not specifically authorize the Census Bureau to conduct surveys for other Federal agencies. Such work, however, was authorized by the Economy Act (Title 31) that allows one agency to perform work for another agency, or by Title 15, which authorizes the Secretary of Commerce (of which the Census Bureau is a component) to conduct special studies for other organizations. When conducting surveys under these titles, we cited the other agency's authority to collect the data but maintained that the data collected in this manner must be kept confidential when the sample from which the survey was drawn was developed under Title 13 (for example, addresses obtained in the decennial census). On the other hand, if the sample was drawn from lists provided by the sponsor or involved canvassing certain geographic areas (area sampling), the confidentiality, if any, was assumed to extend from the sponsor's authority and not from Title 13. Therefore, respondents were notified that we were acting as a collection agent and that the individual information would be turned over to the sponsor who would protect its confidentiality to the extent permitted by law. When Title 13 was amended in 1976 to give the Census Bureau explicit authority to conduct surveys for other agencies we began to use our own authority and apply the Title 13 confidentiality provisions to all reimbursable surveys conducted under that authority.

With the increasing demand from current and prospective sponsors for identifiable data for use in conducting follow-up surveys or in merging a respondent's individual information with administrative record data, the Bureau established a policy in 1987 to conduct reimbursable surveys under Title 15, rather than Title 13, when the following conditions were met:

1. The sponsor has the legal authority to collect the information and to contract with the Census Bureau for the work.

2. The sample is not derived from Census Bureau records which are protected by Title 13.

3. The purposes, content, methods, or other aspects of the survey are not deemed objectionable by the Census Bureau.

4. The sponsoring agency will: sign an agreement binding the sponsor and its contractors and grantees to use the data only for statistical purposes; notify the respondents of the conditions under which the information is being provided; collect and maintain the data in accordance with applicable Federal laws; and prohibit redisclosure in identifiable form.

We have approximately 12 active surveys conducted under the sponsoring agency's data collection authority. The samples for the majority of these surveys were selected from administrative lists provided by the sponsor. However, we are doing the Health Interview Survey for the National Center for Health Statistics (NCHS) using area sampling and a Point of Purchase Survey (CPP) feasibility test for BLS using random digit dialing. Although we anticipate that we will continue to get requests to conduct surveys outside of Title 13, some sponsors will prefer to use census lists to select their samples because alternative frames are not available or too costly.

Use of Special Sworn Employees

As previously mentioned, the Census Bureau has the authority to use temporary staff to perform work authorized by Title 13. This includes employees of other Government agencies and private organizations. The Census Bureau, at its discretion, can appoint an individual as a Special Sworn Employee (SSE) when: 1) that individual is employed by an agency or organization for which we have a contract to provide services or are engaged in a joint project and the person has expertise or specialized knowledge that can contribute to the accomplishment of our projects or activities; 2) the individual is employed by an agency or organization performing a service for the Census Bureau under contract or provides information to the Census Bureau for statistical purposes; or 3) when Federal law requires an individual to audit, inspect, or investigate Census Bureau activities. As an example, we have sworn in employees of the Social Security Administration (SSA) to obtain information from SSA administrative files about respondents to the SIPP for a matching project that we are jointly undertaking. Also, during each Census of Agriculture, we swear in employees of the Department of Agriculture's National Agricultural Statistics Service to review county level summary data. These experts look for abnormalities in the data, based on their local knowledge.

In 1977, the Census Bureau instituted the ASA/NSF/Census Research Program, jointly funded by the Census Bureau and the National Science Foundation (NSF) and administered jointly by the Census Bureau and the American Statistical Association (ASA). Briefly, the purpose of this program is to promote methodological and substantive research involving Census Bureau databases; to provide hands-on experience for graduate students in the fields of statistics, economics, demography and related areas; and to help bridge the gap between academic and government social

science. The ASA Fellowship Program, as it is commonly called, has been instrumental in bringing improvements in Census Bureau operations--primarily by providing increased communication between Bureau staff and the users of our data. Between 1977 and 1987, 32 Fellows and 25 Associates have participated in the program.

The ASA Fellowship Program has fifteen specific goals to bridge the gap between government and academic social science (Table 1). An Evaluation Conference held in June 1986 found that "the program has been highly successful when assessed in terms of its objectives." Regarding Goal 1 (to provide academic scholars with the unique opportunity to have "hands on" access to Census data), Fellows and Associates have used data sets unavailable to researchers outside the Census Bureau for reasons of confidentiality. For example, several Fellows have used microdata from the SIPP and from the Longitudinal Establishment Data (LED) file. Some participants have used data sets constructed from Census Bureau data and data from other agencies. (ASA Grant Proposal, 1987). Through the various research activities conducted with these data sets many of the other goals of the program have been achieved.

In 1986, we instituted the Interagency Research Fellowship Program which was modeled after the ASA Fellowship Program. This new program, however, was designed to support projects funded by other Federal agencies. We believed that a larger program would expand on the successes already achieved; provide more visibility for the program; stimulate intellectual discussion between Census employees and Government researchers; and open up avenues for new approaches to our problems and procedures. As stated in the proposal for the Interagency Research Fellowship Program, it is intended to:

- foster and stimulate increased use of census data bases for methodological and substantive research which would benefit from access to individually identifiable data;

- provide a research environment emphasizing collaborative interests of the Census Bureau and the social sciences research community; and

- stimulate the exchange of substantive and methodological information between Census Bureau personnel and the academic communities.

To be eligible for this fellowship, a qualified person must have a project acceptable to the Census Bureau. In addition, the project must be funded by another Federal agency, state or local government, or an appropriate research funding source. The project must be accepted as having statistical merit, direct relevance to the Census Bureau mission, and be sponsored by a

component of the Bureau. Finally, the project must be approved by the Director of the Census Bureau who will make his determination based on the merits of the research as well as the long-run benefits and costs that the project may have on other Census Bureau programs. Appointments as Research Fellows are for a period of one year, with continuations of up to three years possible. As with the ASA fellowships, Research Fellows must commit to an extensive period of work at the Census Bureau facilities in Suitland, Maryland.

In the initial application of the Interagency Research Fellowship Program, we have brought in a full time employee of the Economic Research Service (ERS) of the U.S. Department of Agriculture to work with a SIPP file containing non-metro designations. As described previously, this detailed geography was needed to assess the economic well-being of non-metro residents--a study fully supported by the Census Bureau's statistical mission. In addition to the analyst, a programmer for the ERS was assigned for six months at the beginning of the project to create an extract of the specific SIPP file which could be used with SPSS software and also to assist in checking the initial tabulations. The research is proceeding quite well and we expect several reports will be published. Also, there has been a healthy exchange of ideas between the ERS researcher and staff in our Population Division which is supporting the work. Administratively, cost accounting has worked fairly well with a special account set up to draw from the $30,000 allocated for computer expenses. The primary administrative complication resulting from this program is the lack of adequate space in the Division for the Fellows to work. The lack of adequate space may limit the expansion of this program to a great extent, especially until after the 1990 Census.

The requirement of this program that all research with the individual microdata records be done onsite has been a significant limitation to some potential Research Fellows who do not wish to commit so much time away from their homes. Although Title 13 does not require that the data we collect be maintained at a specified facility, it is the Bureau's policy that in order to assure security and maintain the public's confidence, we generally require that the data be used in Suitland. To overcome some of the inconvenience to the Research Fellows and other SSEs, we are experimenting with a procedure to locate restricted data at our regional offices. These offices are located in twelve large cities (Table 2) throughout the United States.

We are experimenting with this program through a Joint Statistical Agreement (JSA) with Harvard University. The purpose of this project is to analyze the results of the Post Enumeration Survey component of the decennial census pretest conducted in Los Angeles, California in 1986. Since the file contains geographic identifications to the block level, a public use product is not possible. In addition, it would be quite inconvenient for the Harvard researchers to come to Suitland to process the data.

As a result, we are providing the individual data from this test to the Harvard researchers, who are SSEs, on a microcomputer located at our Boston regional office which is within commuting distance of Harvard. The computer was brought in by the researcher and the data were loaded from floppy disks. Interactive sessions are restricted to the regional office; however, the tabular output can be analyzed at Harvard. The work is to be done over a period of several months and, upon completion, the computer's hard disk will be scrubbed and the computer will be returned to the University.

In the long run, we would prefer a more centralized approach to this program. We envision dedicating a minicomputer at Suitland for this work and connecting it to each of our regional offices through the secure telephone lines which will support our decennial census activities. Terminals at the regional offices could access specific files for authorized projects. There would be no connections to the Bureau's mainframe computers and the files on the minicomputer would not contain any individual identifiers. Survey data matched with administrative records could also reside on the minicomputer. Staff in Suitland would provide technical support to the Research Fellows by monitoring the interactive sessions.

This regionalized approach will not satisfy those Special Sworn Employees who are great distances from a regional office city; nor will it satisfy some of those located in Washington or near a regional office city who are locked into their own machines due to software requirements or cost factors. However, as in the case of the Harvard JSA, there will be instances where it is preferable given the alternatives.

LEGAL OPTIONS

In addition to our public use and administrative solutions, there are legal options which would extend the obligation to protect confidentiality, and the resulting liability, to the data user. These options involve: (1) creation of statutory penalties for improper use of public use microdata, and (2) legal contracts or license agreements that bind the user of public use microdata to use the data only for prescribed statistical studies.

In support of statutory provisions, Robert Pearson of the Social Science Research Council wrote that: "Acceptable disclosure risks are neither easily nor precisely calculated, but such agencies as the Bureau of the Census and the Internal Revenue Service often require (or interpret the laws that govern the release of such data as requiring) that these levels equal zero. I reveal my prejudices here, if not before, in believing that the extended use of federal statistics per se is not inappropriate; but rather that (1) the value of these data are not fully realized and (2) most current statutes under which the release is governed are

inadequate because they recognize only the obligations of those who collect the information, not the obligations of those who may subsequently use them." (Pearson, 1986).

Similarly, Jelke Bethlehem of the Netherlands Central Bureau of Statistics, in a paper presented at the Census Bureau's Fourth Annual Research Conference, concluded that "...disclosure of micro data sets is possible and often difficult to prevent unless the information in the data set is severely reduced." "Therefore," he wrote, "if micro data are released under the conditions that the data may be used for statistical purposes only and that no matching procedures may be carried out at the individual level, any huge effort to identification and disclosure shows clearly malicious intent. In view of the duty of a statistical office to disseminate statistical information, we think disclosure protection for this kind of malpractice could and should be taken care of by legal arrangements, and not by restrictions on the data to be released." (Bethlehem, et.al. 1988).

There are only a few examples of legal arrangements currently being used by statistical organizations. In West Germany, the Federal Statistical Office assumes that there is a residual risk of disclosure in any release of public use microdata. Consequently, they have a means of releasing microdata to an institution under an agreement requiring that:

- The receiving institution pay the cost of modifying records for disclosure control prior to release;
- The receiving institution not try to reidentify records;
- Data may not be transferred to third parties; and
- Violation of the conditions of the release will result in a fine and exclusion from future access to microdata.

Recently, two laws (the Federal Law of Statistics of January 1987 and the Census Law of November, 1985) have had a significant impact on the release of microdata in West Germany. In the Census Law, Articles 17 and 18 specifically prohibit the reidentification of respondents from census data:

<u>Article 17</u>

(1) The characteristics, including the block side (Article 15, para. 4, sentence 3), recorded on the basis of this law will be used only for statistical purposes.

(2) It is prohibited to match characteristics pursuant to para. 1, or to combine such characteristics with data from other statis tical surveys, for establishing a reference to individual persons for other than statis-

tical purposes of this law.

Article 18

Whosoever, contrary to Article 17, para 2, brings together characteristics or data after the characteristics according to Article 17, para 1 have been transferred to data media intended for further computer processing, will be liable to a term of imprisonment not exceeding one year or to a fine.

In the United States, two research organizations, Ohio State University's Center for Human Resource Research and the University of Michigan's Institute for Social Research (ISR) are using or are planning to use contracts as a means of releasing more detailed microdata files. Ohio State University releases a public use file from the National Longitudinal Surveys Youth Cohort (NLSY) conducted by the NORC through funds provided by the Department of Labor. In addition to the public use file, a separate "geocode data tape" containing county codes, college identifiers, some administrative data, and limited information from the County and City Data Book are sold to institutions under a license agreement.

The OSU license agreement requires that: 1) results of the research be published only in summary and statistical form such that no individuals can be identified; 2) files will only be used for specified statistical research and will not be released to unauthorized persons; 3) no attempt will be made to identify an individual on the file; 4) the tape recipient may not hold OSU liable for claims resulting from release of the file; 5) the tapes are destroyed when the work is completed; and 6) the recipient agrees to all protections required by the Privacy Act of 1974. This type of agreement has been used since 1980 and there have been no known breaches of confidentiality or evidence of impropriety. Presumably, if a breach were to occur, the main recourse to OSU is to stop providing the guilty user with these kinds of microdata.

The ISR proposal involves the Panel Study of Income Dynamics (PSID) which ISR conducts with funds provided by the National Science Foundation and others. Currently, PSID public use files show geography to the county level (there are no restrictions on county size). To meet increasing demands for local area data, special research files are being created which identify records by census tract and by ZIP code. ISR plans to release these files to researchers whose institutions co-sign an agreement patterned after the Ohio State license agreement. The recipient institution would be required to provide a detailed proposal as to how they plan to protect the data while it is in their possession. If ISR agrees that the measures are appropriate, the researcher must post a $1000 security deposit before the files would be released. Upon completion of the work, the recipient

attests that all files or derivatives have been destroyed and signs a statement that no known breaches of confidentiality have occurred. The $1000 deposit would then be returned.

I know of no examples of statutory penalties or legally b ing contracts regarding the release of microdata currently in e by U. S. Federal Statistical Agencies. The National er Health Statistics (NCHS) does, however, require purchaser f public use microdata tapes to sign a statement in which t y agree to abide by the NCHS legislation which states that he data may be used only for the purpose for which they we ob- tained, i.e., for statistical purposes." (Mugge, 1983) his signed statement is in addition to established disclos protec- tion measures which are similar to those employed by t Census Bureau. Although not a means to provide greater access, the statement does help to sensitize the user to NCHS' co rn for the confidentiality promised to the respondent.

A legislative approach that could expand access to Census Bureau microdata involves creating a new type of appointment, similar to Research Fellowships, that would provide access to microdata only for general statistical research. Persons, so appointed, would not be Census Bureau employees and would not be restricted to doing research specifically tied to Census Bureau work. This would open the Research Fellowship Program to additional re- searchers and would remove the time limitations associated with temporary employees (SSEs).

Currently, contractual and legislative options such as user liability and research appointments are not available to the Census Bureau. Title 13 does not give us the option of sharing liability with microdata users or providing access to identifi- able records by non-Census Bureau employees. The Census Bureau will look at legislative changes as a means of supplementing or replacing our public use and administrative programs. If such solutions are deemed appropriate, we would need to carefully evaluate how our respondents would react to sharing our respon- sibility for protecting their data with others before we recom- mend any modification to Title 13. In addition, we would have to consider the sensitivity of the information on the file and the consequences of a possible breach on our ability to gain the voluntary cooperation of our respondents in the future.

CONCLUSIONS

This year marks 25 years of producing public use microd files. When we originally conceived the idea, we thought that t users would want to receive the information on computer punch ds, (Zeisset, 1988). Things have changed a lot in these 25 rs--to the point where over one-half billion bytes of informatic can now be stored on a single CD-ROM disk. Now, many private e- searchers and the staffs of nearly all Government agencie have

the ability to process large databases and to apply sophisticated analytical and modeling techniques. The potential social and economic benefits resulting from this research are enormous.

On the other hand, public concerns for protecting individual privacy and confidentiality have been heightened by the vast databases maintained by Government agencies and the trend toward matching files across agencies. These practices, along with the ease with which the data can be handled and analyzed, may cause survey respondents concern that the Government may use their responses, that were provided voluntarily, against them in some way. Businesses, on the other hand, are concerned that competitors will take advantage of information they may glean about their financial situation or marketing strategies. These concerns, if substantiated by a misuse of statistical data, could reduce participation in our censuses and surveys. Although this result may not affect the immediate short-term goals of any individual researcher, it would certainly be a long-term tragedy for the entire statistical community and should provide sufficient incentive for researchers not to abuse the trust respondents placed in the Census Bureau when they provided the information. However, even an innocent misuse or carelessness may be all that is required to markedly reduce public participation in our programs. Also, public use products are not restricted to bonafide researchers and others may not be so motivated.

In this environment, the Census Bureau, as a service organization, must continue to provide the best possible service to our users--especially Federal users who depend on our data to make policy decisions that affect the quality of life for millions of Americans and, at the same time, are responsible for allocating billions of taxpayer dollars. In addition, we must continue to examine and evaluate the potential risks of identifying survey and census respondents from public use microdata and we must establish criteria for acceptable levels of risk. Where public use microdata are not possible given this risk, we will consider alternative products and administrative arrangements that satisfy our user's statistical requirements. Finally, and perhaps most importantly, we **MUST** ensure our respondents that the data they provide the Census Bureau for statistical purposes will **NOT** be used to make determinations about them as individuals.

Acknowledgement

I wish to thank Sherry Courtland who provided guidance throughout the development of this paper and particularly with its organization. I am also indebted to Brian Greenberg and Sang Nguyen for their thorough review and helpful recommendations, especially in the areas of disclosure risk and reduction and public use alternatives to microdata.

REFERENCES

ASA Grant Proposal submitted to the NSF, March 1987; "On-site Research to Improve Government Generated Social Science Data Base;" Proposal SES-8713643.

Bethlehem, J. and W. Keller and J. Pannekoek (1988), "Disclosure Control of Micro Data;" paper presented at the Census Bureau's Fourth Annual Research Conference, Arlington, Virginia; March 1988.

Butz, W.P. (1985), "Data Confidentiality and Public Perceptions: The Case of the European Censuses," 1985 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 90-97.

Cavanaugh, F.J. (1987), "SIPP As An Initiator of a Data Resource Center At the Census Bureau." 1987 Proceedings of the Section on Statistical Computing, American Statistical Association, pp. 149-154.

Dalenius, T. (1988), Controlling Invasion of Privacy in Surveys, Statistics Sweden, Garnisonstryckeriet Stockholm, 1988.

Dalenius, T. (1988), "The Debate on Privacy and Surveys in Sweden," Chance: New Directions for Statistics and Computing, Vol. 1, No. 2; Spring 1988; pp. 43-47.

Duncan, G. and D. Lambert (1987), "The Risk of Disclosure for Microdata," Proceedings of the Third Annual Research Conference, U.S. Bureau of the Census, April 1987.

Govoni, J. and P. J. Waite, (1985), "Development of Public Use File for Manufacturing," 1985 Proceedings of the Section on Business and Economic Statistics, American Statistical Association, pp. 300-302.

Greenberg, B. (1988) "Disclosure Avoidance Research at the Census Bureau," paper presented at the Census Bureau's Joint Advisory Committee Meetings, Arlington, Virginia; April 1988.

Herriot, R. and C. Bowie, D. Kasprzyk, S. Haber (1988), "Enhanced Demographic Data Sets with Employer Characteristics," paper presented at NBER Conference on Research in Income and Wealth, Washington D. C. May 12-14, 1988.

Kim, J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," 1986 Proceedings of the Survey Method Research Section, American Statistical Association, pp. 370-374.

McGuckin, R.H. and S. Nguyen, (1988), Use of Surrogate Files to Conduct Economic Studies with Longitudinal Microdata"; paper presented at the Census Bureau's Fourth Annual Research Conference, Arlington, Va.; March 1988

Monahan, J. (1986), "Development of Microdata Public Use Data File from the Longitudinal Establishment Data File." Internal Memorandum, Center for Economic Studies, U. S. Bureau of the Census.

Mugge, R.H. (1983), "Issues In Protecting Confidentiality In National Health Statistics," 1983 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 592-594.

Paass, G. (1985), "Disclosure Risk and Disclosure Avoidance for Microdata," working paper, Institute for Applied Information Technique, Gesellschaft fur Matematik und Datenverarbeitung, Sankt Augustin bei Bonn, Federal Republic of Germany.

Pearson, R.W. (1986), "The Confidentiality and Extended Use of Federal Statistics," paper presented at the American Statistical Association Meetings, Chicago, Illinois, August 1986.

Rainwater, L. and Smeeding, T. (1988), "The Luxembourg Income Study: the Use of International Telecommunications in Comparative Social Research," The Annals of the American Academy of Political and Social Science, January 1988.

Spruill, N.L. (1984), "Protecting Confidentiality of Business Microdata by Masking," The Public Research Institute.

Wolf, M. (1988), "Microaggregation and Disclosure Avoidance for Economic Establishment Data," paper for presentation at the Joint Session of the Business and Economics Statistics Section and Section on Survey Methods Research, American Statistical Association Meetings, August 1988.

Zeisset, Paul (1988), Interview at Bureau of the Census, Suitland, Md.; May 11, 1988.

Table 1

OBJECTIVES OF THE ASA/NSF/CENSUS RESEARCH PROGRAM

1. To provide academic scholars with the unique opportunity to have "hands on" access to Census data.

2. To provide increased opportunity for accomplished social scientists to work on important problems in a non-academic environment, where production and research needs are often different and can conflict.

3. To stimulate methodological and substantive research in academia on the problems of collecting and analyzing data that provide the basic information for making decisions that can have broad impacts on society.

4. To increase exposure of Census Bureau social scientists to outside expertise, and hence to broaden their perspectives regarding the ultimate analytic uses of the data they produce.

5. To bring about an improvement of the quality of the data collected and disseminated by the Census Bureau.

6. To further specific scientific advances in methodological and substantive areas related to the data collection activities of the Census Bureau.

7. To provide an opportunity for graduate training and doctoral dissertation research using the problems of governmental data collection agencies.

8. To develop a resource group of personnel for future recruitment of statisticians and social scientists to help fill governmental research needs.

9. To provide a large variety of usable real data, as well as computer software programs for their analyses, for teaching and research at academic institutions.

10. To conduct seminars and conferences jointly sponsored by a group of agencies and academic institutions.

11. To increase the interaction and collaborative research and education among agencies and between agencies and academic institutions.

Table 1 (Cont.)

12. To improve the quality of the statistical analysis of Census Bureau data.

13. To suggest important new analyses of existing data that can and should be done.

14. To generate a positive impact on curriculum development at academic institutions.

15. To develop a cadre of people experienced in problems of data methodology and data use who will submit high-quality proposals to NSF to pursue basic and applied research based upon novel ideas and approaches.

Table 2

CENSUS BUREAU REGIONAL OFFICES

Boston, Massachusetts

New York, New York

Philadelphia, Pennsylvania

Detroit, Michigan

Chicago, Illinois

Kansas City, Kansas

Seattle, Washington

Charlotte, North Carolina

Atlanta, Georgia

Dallas, Texas

Denver, Colorado

Los Angeles, California