U.S. Bureau of the Census Data Dissemination Programs

by William P. Butz[1], Associate Director for Demographic Programs and James R. Wetzel , Chief, Center for Demographic Studies U.S. Bureau of the Census

U.S. Census Bureau Data Dissemination Programs

I am pleased to join you today to discuss the data dissemination and public data user support practices of the U.S. Bureau of the Census. Our policy in this area is straightforward and very simple. We strive to provide maximum public access to and use of data we collect, while assuring the confidentiality of respondents and satisfying budget constraints. We maintain an extensive data user support organization, provide user training, and make data available in many forms-printed, microfiche, and all electronic media.

The Bureau staff is inventive and aggressive in its efforts to maximize data access. Among other firsts, we were the first statistical agency to release public-use microdata files from household surveys and censuses (in the early 1960's) and most recently, the first U.S. government agency to release files on CD-ROM. Today, I will describe our basic data dissemination programs and share with you some thoughts on our ongoing efforts to extend the use of our microdata archives by providing analytic access to longitudinal research files on the characteristics of manufacturing firms and microdata from household surveys and censuses for smaller geographic areas.

The Right Data in the Right Form

Our data dissemination policies are expected to simultaneously fulfill a broad spectrum of highly challenging objectives including:

> We are expected to provide low-cost, accurate data;
>
> Our data should be in an easy-to-use form;
>
> We are expected to ensure timely delivery;
>
> We need to provide technical support to users ranging from elementary school students to highly sophisticated, computer-based reprocessors of sample data; and
>
> Our systems must work for a broad range of data and must be flexible to deal with large fluctuations in our output.

To meet the needs of our broad spectrum of users, data are made available in six separate forms: printed reports, summary tape files, CD-ROM, microfiche, public-use microdata tape files, and customized special tabulations. All forms for output are rigorously reviewed and edited to ensure confidentiality of individual data.

Printed Reports. Printed reports are the dominant medium for the bulk distribution of statistics to users. They reach a broader audience than data in machine readable form, and our society still is better equipped to handle - by way of libraries, bookstores, and photocopying machines - the dissemination and maintenance of data in print. We publish between two and four thousand reports a year (ranging in length from one to 1,500 pages). We frequently come to the limit of what we can afford to put into print long before user demands for decision-relevant statistics have been fully satisfied. This problem, which derives in part from high printing and distribution costs, is exacerbated by financial disincentives built into federal printing policy. In brief, when we publish a report, we must provide from appropriated funds all of the capital costs of the publication and make copies available at no charge to various federal agencies (depository libraries, for example). We do not receive a penny from subsequent Government Printing Office sales of the reports. Thus, the more statistical reports we publish, the more it costs us, regardless of how well we may be meeting a need. And we clearly are meeting needs; otherwise the government would not sell 35,000 or more copies of publications such as the Statistical Abstract annually.

Summary Tape Files. Summary tape files (the second major medium for data dissemination) actually are a by-product of processing data for printed reports. These machine-readable data files contain frequency counts or aggregates similar to those in printed reports. In many cases, these files provides much more statistical detail than is available in final printed reports; the files often are more timely than printed reports; and, of course, they already are in a form conducive to automated analysis. The costs of data collection, and of designing, producing and thoroughly documenting the files are part of functions supported through our regular funding. Since there are no printing costs and few compulsory distributions, the disincentives which make printed reports so expensive for us do not apply to machine-readable data files.

To maximize the utility of data in this cost-effective form, we distribute selected, heavily-used files to fifty-four State Data Centers (including District of Columbia, Puerto Rico, Guam, and the Virgin Islands) to help make this resource widely available to users without computer facilities.

CD-ROM. Digital laser-disk technology is just beginning to be used by the federal government for dissemination of large files. CD-ROM (compact disk read-only memory) readers are now on the market that allow a microcomputer user to search and read data bases from a single disk holding over 500 million characters, equivalent to the combined storage capacity of at least three high-density tapes. Further, since these disks allow "random access," any piece of information in an appropriately indexed data base can be retrieved immediately, unlike data on "sequential access" tape, where every preceding piece of information must be read first before the desired data can be accessed. The price of CD-ROM laser-disk readers is likely to decrease further, but already they are well within the reach of business microcomputer owners.

Microfiche. The needs of data users with bulk holdings, such as libraries, frequently are met through a third medium, microfiche. Microfiche are much less expensive than printed products; they display much larger quantities of statistics in a given amount of storage space; and retrieval of selected statistics can be accomplished without the technical sophistication required of a computer user. Most data from the 1987 Economic and Agricultural Censuses and the 1990 Decennial Census will be available on microfiche.

Special Tabulations. While the quantities of statistics available in print, microfiche, computer tapes, and other media have increased steadily for many years, there are inevitably specialized needs we cannot hope to anticipate with general-purpose products. To deal with such situations, the Bureau maintains a capacity to respond to user requests for customized statistical output, assuming, of course, that the user is prepared to reimburse the Bureau for production costs. I am not talking about the kinds of statistical compilations that could be performed by someone else using products in the public domain, but rather those tabulations which require use of the Bureau's confidential basic microdata on establishments, persons, or housing units.

All special tabulations are treated as an extension of our publication program by eventually making copies available at reproduction cost to anyone after delivery to the sponsor. Thus, they sometimes provide a way for us to prepare data products which we would like to have available to users, but which cannot be prepared within appropriated funds. For example, ZIP code tabulations from the 1980 census were included in our original publication plans but were cancelled in budget cutbacks. A private consortium of data users raised the funds needed to prepare the ZIP Code data. The consortium had first access to the data, but the general public eventually had access as well.

Public-Use Microdata Files. Policy analysts, researchers, and data reprocessors who want to create their own unique data sets, regression analyses, simulation models, and so forth can do so by using public-use microdata samples (PUMS) that present household sample survey data or census sample data in individual record form. Of course, certain geographic information and respondent characteristics data are suppressed on these files to ensure that the identity of any particular person or housing unit is not disclosed. Yet, they provide enormous detail. For example, a Current Population Survey public-use file may contain age in single years through age 90, occupation coded to more than 500 categories, and earnings in original detail up to an appropriate disclosure limit. The popularity of these files is illustrated by the growth of demand. In 1968, we sold a few dozen public-use files. In 1988, we sold more than a thousand.

Because of technological advances and the amazing power of low-cost microcomputers, it is now becoming feasible and cost effective for an individual to access and use the output of a large survey. A good example is the American Housing Survey. The microdata records from the last biennial survey were recently released on CD-ROM at a cost of $125. The disk includes about 250 megabytes of data, but it can be used for research purposes on a very low-cost microcomputer set-up. Of course, we are not talking about processing speeds such as one would expect on a minicomputer or mainframe. A CD-ROM is a relatively slow storage device, something like the speed of a floppy disk. Thus, it can take a very long time for a microcomputer to process a large microdata file on CD-ROM, reading and tallying one record at a time: perhaps a number of hours. Such timing is frequently not a problem, and where it is, a user may be able to extract a subset of the data onto a fast hard disk for more acceptable response time.

We are excited about the potential of CD-ROM, and a variety of releases can be anticipated over the near term. We are already looking into CD-ROM releases of the March Current Population Survey for 1988 and 1989, of the National Crime Survey for 1988, and of the American Housing Survey for 1987-88. These likely releases will be accompanied by full documentation (which can be purchased separately for $5), and we are now conducting CD-ROM user courses at select locations.

Relationship of Machine-Readable Data to Published Data. I have described briefly six ways in which we make data available: printed reports, microfiche, three kinds of machine-readable data files - summary tape files public-use microdata, and CD-ROM, and special tabulations (which can take any of the foregoing forms).

Although this audience is primarily interested in computerized data, I think it is worthwhile to remind you of the other media for three reasons:

First, some of your data needs can be met more quickly and at a lower cost with information from printed reports or microfiche.

Second, verification that you have accurately retrieved information from summary tapes or microdata files is easy when you have substantial information already tabulated to check against.

Finally, as you process machine-readable data, you frequently need to group data into categories. Review of frequency distributions in existing tabulations may save you several steps in data processing or model estimation.

Economic Programs Data

For the next 18 months, the 1989 Economic and Agricultural Censuses will be our most prolific source of new information. Hundreds of reports, summary tape files, and CD-ROMs will be released over the months ahead. Each of the summary tape files is associated with one or more publications series, so there are always some published numbers for verification purposes.

All computerized releases under the 1989 Economic and Agricultural Censuses programs will be available on CD-ROM; and special ZIP Code files on retail trade, services, and manufacturing will be available only on CD-ROM and computer tape. I should add that these data sets will be the first major release of information that employs the new Metropolitan Statistical Area definitions. The economic data are all summary data and, when properly indexed, response time in normal use can be quite fast, even with large files.

Public-use files have not yet been prepared from economic surveys and censuses because of the unique visibility of establishments, the availability of private sector and regulatory data bases, and the consequent risk of disclosure of confidential data. Demand continues to grow, however, for public-use files on business, particularly those relating to the manufacturing sector. In a conference sponsored by the Census Bureau in 1984, more than 100 economists expressed their desire for such public-use files.

Recently, we developed a longitudinal file of manufacturers called the Longitudinal Research Data Base (LRD). At present, access to these files is limited to staff and selected visiting researchers (NSF-AAA-Census fellow, for example) who must work on-site at the Bureau and be sworn in as special agents of the Census Bureau so they are subject to all the disclosure rules and penalties that apply to the regular staff. By developing models and estimating relationships at the Bureau, however, these researchers are in a position to conduct powerful empirical studies in areas of major policy concern. At present, at least five economic research projects involving researchers from prominent universities are based in part on special LRD estimates prepared at the Bureau. We are exploring a system that permits researchers to submit special requests for computerized statistical analyses and get the results quickly. Also, the staff is working on the development 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 microdata estimates of the 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 the like from outside sources. Two types of transformations are being considered: 1) stochastic transformations that involve adding random noise to the original data while preserving, for example, the mean and variance of the variables and the covariance relationships between variables; and 2) non-stochastic transformations that provide for the release of the transformed release of data in ratio form. Each of these methods has merit, but each limits the types of economic research it will support and in its ability to mask the microdata. We are continuing to work on these knotty problems and expect to find solutions, thereby expanding the range of economic information available to the research community.

Demographic Programs Data

Nowadays, the Census Bureau releases a steady stream of household survey public-use files. Public-use files are edited to remove name, address, and selected geographic identification for smaller places, extreme values for continuous variables, and information that is obtained from or matchable to administrative records systems. There was a time when the user community was fully satisfied with our edited public-use files. That is no longer the case. There are increasing user demands to restore detailed information that was removed from public-use files to protect the identity of survey and census respondents or to add information from administrative records systems to the survey results. Often these demands are clearly in the public interest. That is, the request has a clear-cut federal government program or policy application. However, the Census Bureau's authorizing legislation - Title 13 of U.S. Code - clearly forbids release of data that can be used to identity a particular respondent. Thus, we must be concerned about whether the protections afforded these public-use are sufficient. High speed computers have made public-use files more attractive, but they have also increased public concern about potential abuses to individual privacy resulting from the creation of large integrated data bases. In recent years, events in West Germany, Sweden, and other European countries regarding government data bases have highlighted this concern. Moreover, computer hackers have raised fears that, given enough patience, someone could defeat any scheme designed to protect confidentiality.

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 to which they are entitled. We have established a permanent staff

to study microdata disclosure risk and avoidance. This staff is charged with finding "safe" methods of increasing access to microdata. Some options they are exploring include masking public-use microdata, creating publicly-releasable alternatives to microdata (e.g., tabulated summary statistics by class, correlation matrices of the data, provision of test files so programs can be developed by outside analysts then run by Bureau staff), and special administrative arrangements such as visiting scholar programs.

Our research effort is designed 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 are responsible for allocating billions of taxpayer dollars. At the same time, we will avoid the potential risk of identifying survey and census respondents from public-use microdata. Where public-user microdata are not possible given this risk, we will, in time, design alternative products and administrative arrangements, within the requirements of Title 13, that satisfy our users' statistical requirements. We will assure our respondents that the data they provide the Census Bureau for statistical purposes will not be used to make determinations about them as individuals but will be used for the fullest possible appropriate application in the conduct of the public business.

Assisting Data Users

Assisting data users is a major responsibility. We do not release machine-readable files without providing written information about each file. We do conduct training in the use of new technology (e.g., CD-ROM), and we maintain a centralized customer sales/service function.

The Census Bureau has established a substantial reputation for comprehensive documentation, including not only record layout information but glossaries of concept definitions and statements regarding limitations of the data. Documentation is included in the cost of each tape or CD-ROM purchased or can be obtained in advance for a small charge. Acquisition and availability information can be obtained from three census publications:

Major new data files are listed in Census and You, the Bureau's official monthly newsletter for data users. Listings are relatively timely, but only selected files are covered.

Monthly Product Announcement (MPA) complements Census and you by giving titles, prices, and ordering information for every publication, microfiche, and data file issued, all with the least possible delay. The MPA does so, however, at the expense of not including any descriptive information beyond the title.

The Bureau of the Census Catalog also presents ordering information, as well as abstracts. Each annual issue covers products issued since 1980.

Our Data User Services Division conducts training programs to help users become acquainted with Bureau products. Of special current interest is the one-day course on CD-ROM. Another component of the user training program is our College Curriculum Support Project, where we develop materials for instructors to use in teaching students about census data. More information on both software and training programs is available from the Data User Services Division.

Our centralized Customer Services Branch in the Data User Services Division handles tape, documentation, and some report sales. Copies of computer tape data files (public-use files and summary tape files) are priced at a standard $175 per reel of tape, documentation included. The CD-ROM of the American Housing Survey is available at a cost of $125 as are earlier CD-ROMs, one containing the 1980 Decennial Census ZIP Code file and the other containing the 1982 Census of Agriculture data base by county and the 1982 Census of Retail Establishments by ZIP Code. This summer, we expect to release separate CD-ROMs of geographic data from TIGER (Topologically Integrated Geographic Encoding and Referencing) systems, county business patterns, and the city-county data book. To facilitate distribution, we have developed a system that permits users to establish a deposit account, so they can call in telephone orders and get prompt shipment. Otherwise, it is necessary to send a check or charge to a VISA or Mastercard with your order. The same office serves as a centralized inquiry service on the availability of all Census Bureau data. Its phone number is 301/763-4100. Each of the 12 regional offices of the Census Bureau also handles inquiries, although they do not accept orders.

State Data Center and Clearinghouse Program. Wherever possible, we foster development of downstream data services in the private sector. One of our most successful efforts is the State Data Center program. This cooperative federal-state effort builds on the capacity of state governments to serve the data needs of their constituents. The Census Bureau provides designated state organizations in all of the states with basic data products and training. The state provides staffing and supports a wide variety of data retrieval, promotion, and training activities for users in the state. State Data Centers increasingly are becoming the focal point for local as well as Census statistics. For example, most State Data Centers have vital statistics and other data bases generated within their states. Lists of the State Data Centers and local contact points are available on request from the Data User Services Division.

We also maintain a list of private, public, and academic institutions in our National Clearinghouse for Census Data Services. Each organization registered with the Clearinghouse determines the services it will provide, sets its own fee structure, and is neither franchised nor certified by the Census Bureau. Many organizations on the list provide specialized services, and a few do provide time-shared access to census data on their remote data base.

Evaluation of Data Products

There is one last topic I would like to cover: evaluation of our data products and their uses. Our data users are not usually reticent about letting us know what they want and need. A great many individuals as well as organized groups of users provide excellent feedback; and we have taken steps, many of which I have described here, to improve our products and their accessibility. But we need to know how effective these improvements have been, whether other products warrant improvement, and whether we are spending scarce resources (staff time as well as money) correctly. To answer these questions, we plan to do a Bureau-wide evaluation. This will be a systematic evaluation, perhaps done by an outside contractor, covering all aspects of our data presentation and dissemination programs.. We expect to learn much about how effective the presentation of our myriad data products is, how useful they are to our customers, and what changes ought to be made.

Conclusion

The burgeoning demand for statistical data along with rapid advances in the technology to store and access data present wonderful opportunities for the Bureau of the Census to expand its data production and dissemination programs. We have worked hard to meet increased demands for data by releasing more detailed data products in a variety of formats. By providing data on CD-ROMs, by designing new ways of presenting data such as longitudinal research files, and by developing ways of doing special statistical analyses on request, we are striving to meet user needs. Always mindful of our absolute guarantee to protect the confidentiality of our respondents, we continue to explore new ways to improve access without jeopardizing that trust. ❑

[1]Presented at the IFDO/IASSIST 89 Conference held in Jerusalem, Israel, May 15-18, 1989.