Comments on the Data Access and Dissemination System

by Lisa J. Neidert[1],
Data Archive, Population Studies Center
University of Michigan

The Data Access and Dissemination System (DADS) will be the vehicle for dissemination of census data in the year 2000. Information distributed in published census volumes in 1990 will be accessed from the internet for future censuses. Complete data files will no longer be written to media for redistribution to users. Instead, users will access DADS and pull off the tables they need. The advantages to the U.S. Census Bureau and its customers are quicker turnaround for release of files, cost-effectiveness, and increased access.

Several factors have probably motivated the Census Bureau to make this move. First, all federal agencies are responding to Al Gore's call for internet access by January 1, 1996. Second, changes in technology, such as the development of the internet, high speed computers, and low-cost storage have made this method of distribution feasible. In the past year, we have witnessed an explosion in the number of websites that distribute data (e.g. PSID, HRS/AHEAD, National Longitudinal Surveys, IPUMS, Milwaukee Parental Choice Program, Wisconsin Longitudinal Study, Russian Longitudinal Monitoring Survey, World Fertility Surveys, Malaysia Family Life Survey, Survey of Families and Households.) Finally, distributing information via DADS is a cheaper alternative for the Census Bureau, particularly when compared with the cost of printing.

As with any change, however, there are probably people who were better served by the old dissemination methods than they will be by DADS. It is clear that the Census Bureau wants this system to serve all users. However, there are some shortcomings that should be solved in upcoming renditions of DADS if the Census Bureau is to reach that goal.

The Data Access and Dissemination System doesn't exist yet. It is still a concept. However, I will use the "Data Access" page on the Census Bureau's Web site provides a good working model of DADS; and much of it is likely to be incorporated into the future operational DADS. It is also likely that many features of this current system will be remodeled, so some of my comments may be "old news" to the inner circle of DADS developers.

The current configuration of DADS needs three important improvements. First, DADS should provide the same information that one could get using the old dissemination methods. The data may be in a different form than they were in the past, but the content of the data must not be compromised. When the PSID changed from a family/individual file with a reocrd length approaching 32,767 to its new form of family records and individual records, the same information was still available. It takes new knowledge to work with the data, but users can still create the same sorts of tables they could create in the past. In contrast, the current configuration of DADS does not allow users to create all the tables they could in the past. Second, DADS should accommodate the users who have access to high-speed computers and large amounts of disk space. Some users would like to have the data on their own systems, rather than requesting tables and extracts from DADS. The FTP access of raw files is weak in the Data Access site. If FTP access to raw files proves to be impossible because of the need to protect respondent confidentiality, then the extraction system should be improved. Finally, and perhaps foremost in the minds of data librarians and archivists, there is the question of whether DADS will meet the archival needs of future users of census data. Expertise on and access to state and federal records tends to be fairly short-lived. Thus, it is essential that the archival needs of future researcheers be considered in the development of DADS. What sorts of records will be turned over to the National Archives and in what form?

Loss of Information with DADS

Most users of summary tape files (STF) find the summary-level sequence charts confusing. However,the current configuration of the Data Access system does not make it clear that all the choices available in a typical summary tape file are available in the new system. Users can get tabulations for states, counties, metropolitan statistical areas (MSAs), tracts, and blocks—the most typical choices. But can they get tabulations for central cities of MSAs (summary level 340) or for any of the American Indian Reservation categorizations (summary levels 210-221)? What about county-specific zip code statistics (summary level 820 versus summary level 810)?

The way the Data Access system is currently configured some items in the geographic identification section are not accesible. Occasionally users need the longitude and latitude or land area of census tracts or blocks for the computation of a summary measure such as a residential segregation index. However, users can't select these items, or other items such as consolidated city population size code, place class code, or place description code from this section.

DADS works best when users are making a request for a small number of tables for a small number of geographic units. For example, this system works well for a user who wants to know the population size (a single cell) for all counties in Michigan or the distribution of income in Houston, Texas. However, many analysts need perhaps 10 or 12 tables (which might mean 200 or even 1,000 cells) for all zip codes in the nation. To get these data from DADS in its current configuration, one must list all the relevant zip code(s). Typing in over 10,000 zip codes is not a very practical alternative! In a typical STF request based on data stored on-line, one would select the appropriate summary level for zip code (820) and would get all the tables for all zip codes with the execution of one job. The configuration for block groups and tracts is similar, but one only has to highlight the tracts or block groups rather than typing them out. DADS can handle these requests for summary statistics for all zip codes or census tracks in the U.S., but it is a very labor-intensive task for the requester. The analyst who makes this sort of request is not just mining data that are never analyzed. The analyst is reducing 31,000 columns of information into 200-1,000 columns and then making the request for a unit of analysis that might range from around 3,000 for counties to more than 100,000 for block groups.

I'm certain that the future DADs will allow access to all summary tape file information. However, so far, only summary tape files 1 and 3 were released in CD-ROM form. Thus, none of the race-specific tabulations from STF2 and STF4 are available under the current Data Access system.

One would hope that DADS would allow the Census Bureau to eliminate the distinction between summary tape files and public use microdata files. It would be very useful for researchers to be able to define their own tables rather than being restricted to the limited number of tables supplied by the Census Bureau in its summary tape files. We had some researchers at Michigan recently wanted to look at disability according to race and sex. However, our researchers needed an age breakdown other than the typical 18-64 and 65+ groupings. Because the table they needed was not available in an STF file, they created one using PUMs files. Using PUMs, however, gave them a smaller case base, and the census geography could not be perfectly duplicated. In general, if analysts make a table or summary statistic based on a small number of variables, they should be able to get the tabulation for any level of geography. However, if they want to use 15 variables to define a summary statist!ic, then the level of geography becomes much more restricted (state, MSA, or PUMA). Thus, another advantage of eliminating the distinction between summary tape files and public use microdata files would be the increased sample size for the public use microdata files. Small populations such as male clerical workers, female pilots, 50 to 54 year-old women with own children under 5 years of age, or persons born in Guam could be studied better with a 16% count rather than the 5% files available with public use microdata. (On the other hand, I shudder to think that some of our researchers would be tempted to try to swallow the 16% count of white prime-age males when even the 5% count proves to be fairly cumbersome.)

FTP Access and/or Improving the Extraction Engine

Researchers who have excellent computing facilities may not want to get in the DADS queue every time they want access to census data. If the demand for the system is great, the Census Bureau may want to allow for FTP access so that users who have the capacity can bypass DADS except for quick exploration and for FTP access to the original raw files. If for reasons of confidentiality, the Census Bureau cannot provide access to the raw files—perhaps because confidentiality is built into the DADS software rather than into the data (via sample size or census geography)—then the DADS system needs to make improvements to the existing extraction engine.

The systems developed by CIESIN (Ulysses) and Public Data Queries (Explore) are extremely fast. One of the reasons they are so fast is that they produce is tables instead of the cases and variables used to produce the tables (or summary statistics). Any time one writes out individual records rather than tables or summary statistics, response time slows precipitously. If many users want micro-level extracts, as opposed to exploratory tables or even output from a summary tape file request (which is always a good example of data reduction), the response time will begin to discourage and irritate users. If a user has the capacity to handle the raw files, the Census Bureau should allow the user to do so, and thereby free up time for users who need the CPU.

The creators of the Integrated Public Use Microdata Samples (IPUMS) have found that their data cannot be used by all who might be interested in them, partly because of the sheer size of the files (125G) and partly because their primary audience (historians) traditionally has had limited access to powerful workstations. Thus, the IPUMS creators developed an extraction system that allows a somewhat disenfranchised user to create a work file. (These users are not completely disenfranchised as they do have access to the internet.) However, response time will not be quick with the IPUMS data extraction system because, at least for the short run, all extracts will be executed on a single Sparc20. Clearly, a user with access to large disk storage and a powerful processor will be better off running the extract at his/her own desktop. Of course, the calculus needed to figure out whether the extract should be executed by the IPUMS workstation or a local workstation is complicated by the fact that other products, such as an extract codebook and SPSS cards are created along with the IPUMS-based extract.

Researchers at my site, the Population Studies Center at the University of Michigan, have made countless extracts since the release of the 1990 PUMs files using an in-house program that rectangularizes the hierachical structure of these files. Turnaround time is relatively quick (45 minutes - 3 hours) depending on the sample being used (1%, 3%, 5%, 8%), number of states requested, the size of the file being written out, and the load on the system. We purchased most of the microdata from the Census Bureau for $4,800 (5% - $4,000 and 1% - $800). I don't have a count of the number of extracts performed over the past few years but conservatively it has been 500 which works out to be $10 an extract and is more likely to be over a 1000. DADS will not be able to provide this quick and cost-effective system for our users although many users will be ecstatic about the system that DADS will provide.

Improving the Extraction Engine

Researchers often need access to more information than the tabular data provided through the STF data extraction system. The need for exploratory analysis can be met with tabular data and summary statistics; and more time spent exploring data before analysis often means less information actually ends up being extracted because the user has a much better idea of what is needed for the actual analysis. However, researchers often need access to microdata so that they can estimate equations. The systems developed by CIESIN (Ulysses) and Public Data Queries (Explore) allow researchers and policy analysts to get means and crosstabs from PUMs data in a matter of seconds; but not all statistical needs can be met with simple means and crosstabulations. The Census Bureau is aware of all of this and provides a Data Extraction engine for microdata. However, in the case of hierarchical files similar to census microdata (CPS), the interface for extraction is quite awkward.   The interface needs to be improved, particularly, if for reasons of confidentiality, access to microdata is limited to the Census Bureau extraction engine.

Currently, the extraction procedure requires users to extract the records separately by record type (household, family, person) even though almost all users want a rectangular product. Although, the user certainly can merge the household, family, and person records to create a rectangular file there does not seem to be a rationale for adding this extra step to the procedure. In addtion, merging across record types increases the possibility that a novice user will end up with an erroneous file and not realize it. Novice users would also benefit from features such as variables that provide counts across the household, (e.g. the number of children under 4 or the number of earners) and the option to rectangularize the record based on something other than record type (e.g., rectangularize by household relationship for husband/wife or a mother/child file). Another common request is to select all person records if any person in a household meets a certain criteria, such as foreign birth, unemployment, age 60+, or interstate migration. Of course, the more bells and whistles that are added to the extraction engine, the more likely it is that people will use it for data management rather than just data access.

Archival Issues

The final question that a system like DADS invokes is how it can be archived. How will the Census Bureau unpack DADS so that they can turn over raw data and a codebook to the National Archives? Will there even be a codebook if the Census Bureau intends not to disseminate raw files and technical documentation as it did in the past? If the confidentiality firewall is built into the software, how can this information become part of the raw data so that confidentiality requirements continue to be fulfilled? Much of DADS sounds dynamic, which suggests that the system will be updated to include more data and perhaps that variables will be recoded to meet the demands of users. At what point will DADS be stabilized so that there is an archival record?

Table 1 has a list of questions that can help provoke our thinking and serve as guidelines in determining who should be responsible for making an archive out of DADS. Given the complexity and enormity of DADS, we may be tempted to allow the Census Bureau to be the archive for the census of 2000 and for all future data products, particularly because the Census Bureau looks like the archival expert when compared to the National Archives on many of these questions. However, it is important to remember that most state and federal data producers have poor long-term memories about old data (sometimes the definition of old is just a few years) and that there has been a lack of institutional memory within the Census Bureau about previous data losses due to poor archival policy. An article by Dollar nicely summarizes the historical record of federal data producers and the National Archives. In the decision on whether to archive summary statistics versus microdata from the 1940 census the reasoning was "if the Government agency that created the records for statistical purposes did not fully exploit them, it is hardly likely that anyone else will." (Dollar, 198x: 79). Thus, 1940 microdata were expendable.

Table 1

Who Should be Responsible for Data

(1)Is there expertise in the creating agency that can explain the context, technicalities of the subject area, or the idiosyncrasies of the data which would not be available if the records were transferred to an archive? Will that expertise remain available for all electronic records, or only for those in active systems.

(2)What functionality of the system used to create the records is necessary to meet the needs of archival users? Can the archives provide the necessary degree of functionality, or is the creating agency the only economically or technologically feasible place to preserve the data in a usable format?

(3)Will the creating agency guarantee equitable access within freedom of information and confidentiality policy guidelines?

(4)Do the records have continuing value to the creating agency so that it has an interest in and need to maintain the records beyond an external requirement?

(5)Will there be a duplication of effort if the archives acquire electronic records that have continuing value to the originating office?

(6)Where will the risk of loss or destruction be minimized?

(7)Can the creating agency guarantee that it will stabilize and not alter the archival record?

(8)Do regulations prohibit transfer of records from the custody of the original agency?

(9)What is the total cost to the organization to maintain electronic records for accountability and research purposes? How can these costs be reduced for the institution as a whole, without eliminating services to users?

Source: Hedstrom, Margaret. 1991. "Archives: To Be or Not to Be: A Commentary." Archives and Museum Informatics, Technical Report, Number 13.

References

Dollar, Charles. 1979. "Machine-Readable Records of the Federal Government the National Archives.", *Archivists and Machine-Readable Records: Proceedings of the Conference on Archival Management of Machine-Readable Records*. Edited by Carolyn G. Geda, Erik W. Austin, and Francis X. Blouin, Jr.

Hedstrom, Margaret. 1991. "Archives: To Be or Not to Be: A Commentary." *Archives and Museum Informatics, Technical Report*, Number 13.

[1] Paper presented at the annual meetings of IASSIST, May 15, 1996, Minneapolis, Minnesota.