DEVELOPING A METADATA DATABASE AT THE CENSUS BUREAU

Daniel W. Gillman and Martin V. Appel
Bureau of the Census, Washington, DC 20233

KEY WORDS: Survey data files

The Census Bureau is conducting research into developing a metadata database for census and survey data. Metadata are data items which describe the characteristics of data. They consist of information such as physical location, storage medium, description of survey, file layout, record format, code structure, etc. If the prototype repository being constructed proves feasible, it is envisioned that a metadata database will be the repository for the metadata of all finalized and some research data sets created by the Census Bureau. Researchers and analysts will have a much easier task of searching, locating, and identifying available relevant data for their work. This paper describes the concepts and models that have been developed to produce this prototype.

1. INTRODUCTION

The Census Bureau collects data about the population and economy through a series of censuses and ongoing surveys. Almost all of this information is processed and stored using computer technology. As time goes on, the quantity and complexity of the data sets created has increased. The responsibility for creating data sets for an individual survey lies with the organizational unit overseeing that survey. Since the Census Bureau conducts many censuses and surveys, this results in a widely decentralized system for maintaining data and files.

To make matters worse, information about which data sets have been created, what they contain, or where they are stored is not maintained in a standard way. Researchers, especially within the Census Bureau, who need some data for their work must spend substantial time to find the right person to contact. Often the documentation for the data set is scant or does not correspond to the data exactly. The situation is somewhat better for people desiring public-use files because there is a centralized phone-in facility. However, only descriptive information is available about each data set. Detailed information about specific fields must be obtained elsewhere.

To address this problem, the Census Bureau is conducting research into developing a metadata database (MDB) for census and survey data. This research involves modeling Census Bureau data and constructing a prototype. This paper describes what metadata are in more detail, gives a detailed vision of what the MDB should be, describes a model for Census Bureau data, and the progress achieved so far.

2. METADATA

Metadata are the data items which describe the characteristics of data. It is envisioned that the MDB will be a repository for the metadata of all finalized and some research data sets created by the Census Bureau. Researchers and analysts will have a much easier task of searching, locating, and identifying available relevant data for their work. They are information about data that allows for effective management, intelligent queries, and efficient retrieval (Almond, 1994 and Sumpter, 1994). Typically, Census Bureau metadata might contain information such as physical location, storage medium, description of survey, file layout, record format, code structures, etc. Metadata can be contained in files along with the data, such as in the "BOX File" format (Bean, 1991), or they can be stored in a separate location. We will assume for the remainder of this paper that metadata are also stored separately, because the MDB envisioned will be a repository of metadata, not the data they describe.

Metadata can be divided into three categories: system, application, and administrative. System metadata describe physical and logical characteristics needed for computer processing such as location, storage medium, record layouts, data dictionaries, date of creation, etc. Application metadata includes the survey universe, sample design, questionnaire, and other information which a data user needs to understand, query, or use the data. The third type of metadata concerns administrative information such as budgets, schedules, phone numbers, and office locations for persons who can produce the data files or provide subject matter expert advice.

The efficient and effective use of data is made possible by the organized storage and use of metadata. Data sets become much more useful when complete metadata descriptions are readily available. When metadata are centrally maintained for collections of data sets, users who need to determine which files are appropriate for their work can do so. Many types of questions can be answered through metadata queries. Some examples are:

a) Find data sets which contain specific information, such as yearly income;
b) Find data sets which share common information from which links can be made to form larger data sets;

c) Locate data sets by broad subject through pointers to specific items under those subjects;
d) Monitor data storage system usage by tracking file sizes;
e) Locate surveys with similar or a specific set of characteristics.

These and other types of queries will be discussed further in sections 3 and 4.

A metadata repository standard does exist.[1] However, it is not used at the Census Bureau. This standard describes the way information about data should be organized. Software systems exist which are compliant with this standard. The X3H4 standards committee is now working on a new international standard called PCTE, Portable Common Tool Environment. It will be several years before PCTE is finalized and accepted. IRDS will continue to be supported but will eventually be superseded by PCTE.

An attempt has been made to standardize the maintenance of some metadata at the Census Bureau. This system uses the "BOX file" format. BOX files contain metadata at the front and must be produced according to a specific format. The format is flexible so that any information considered important can be included. There is a core of information, such as field definitions, which must be included. The major advantage to the system is that metadata are automatically carried along with the data set in a file. So far, this system has seen only limited use. There is a data processing standard currently pending at the Census Bureau which addresses data archiving. This standard requires the use of the BOX or similar format for generating and maintaining metadata for files to be archived.

The Census Bureau maintains a large library of data for public use. Users receive information about available files through the "Census Catalog and Guide" or the "Monthly Product and Announcement." An automated information system called the Automated Reference Rack (ARRk) exists where callers describe to telephone representatives the data they seek and the representatives identify the appropriate data sets. ARRk is a hypertext information system built using Lotus's Smartext software. The Census Bureau also maintains an online bulletin board. Users can download data and software, access press releases, and exchange questions and answers. Lastly, there is a service called CENDATA, accessed through Dialog and Compuserve, from which users can download reports and press releases.

Three large scale data retrieval systems that have been developed for accessing either 1990 Census or Survey of Income and Program Participation (SIPP) data. They are SIPP-on-Call, CenSAS, and DAPS90. Access is both internal and external. There is also a database called Administrative Record Information System (ARIS) which is an information system for administrative records files which the Census Bureau use.

These examples are given to illustrate that good work has been done to organize information about Census Bureau data, but the attempts are either limited in breadth (i.e. cover limited types of data such as Census), limited in depth (i.e. cover a broad area without a lot of detail such as ARRk), or limited in acceptance such as the BOX file format.

3. MDB GOALS

The MDB envisioned is to be a repository for the metadata for all finalized survey, census, and some research data sets the Census Bureau produces. Links to the data files themselves may eventually be made, creating a full data warehouse, but that will be a future project.

Because the Census Bureau manages data in a decentralized and non-uniform way, the MDB will bridge the gap between the data and the users who wish to find them. On the one hand there is the need for the managers for each survey to create and manage their data in the most efficient way for their processing needs. On the other, there is a need for data users to be able to find and access data efficiently and effectively. The MDB will facilitate a solution for the data users while allowing the survey data managers to find a smooth transition to standard data management strategies.

There are many specific functions for which the MDB is being designed. Primarily, the MDB envisioned will be a standard tool for researchers and analysts to locate desired data. Every unique data dictionary, i.e. the list of fields and related information about a data set, will be stored with links to entities including file names which use the data dictionary, survey names of the surveys that produced the data sets, geography represented by the data in the data sets, and subjects covered by the data in the data sets. Queries will produce a list of files and locations related to the information pointed to through the links determined by the search criteria.

Some administrative functions can be performed using the MDB, too. An important function is to determine total data storage needs and current capacity. This can be accomplished through reports. Performing a full accounting of all the finalized data sets is a simple report, and identifying data sets which are duplicates of other data sets previously registered can be accomplished, though that task is more difficult.

A last but extremely important requirement of the MDB is naming standards and data element definitions. Many surveys define fields or attributes with the same name but with (slightly) different definitions. An example is the term "non-response". Survey designers define this term differently for different surveys. An analyst linking two survey data sets together from different sources on the non-response field would get into trouble. An aim of the MDB is to help people avoid this problem. If data elements are standardized across surveys, then confusion generated by the differences in meaning will disappear. Naming standards and conventions are also needed to remove the confusion. Once these standards are in place, two fields having the same name from different surveys will mean the same thing.

4. DATA MODEL

In previous sections we described the major functions we want the MDB to perform. Here we present the model for the Census Bureau's metadata around which the MDB will be designed. The model is still under development and is subject to change. What follows is a description of the model's main entities and the relationships among those entities. See figure 1.

CONTACTS: Entity lists the contact information for obtaining files.

DATA DICTIONARIES: Entity lists all the fields and related information for each data set described by the MDB. When two or more files have the same data dictionary information associated with them, these metadata are not duplicated.

DATA ELEMENTS: Entity lists the data elements used in data sets and data dictionaries. Links to *generic files* and *data dictionaries* tell where the data elements can be found.

FILE NAMES: Entity lists all the names and related information (such as number of bytes, creation date) of each file whose metadata is contained in the MDB.

GENERIC FILES: Entity names and lists all the sets of data sets with different data dictionaries. If two data sets have the same data dictionary they belong to the same generic class.

GEOGRAPHY: Entity lists the types of geography available in Census Bureau data. Data sets which are tied to types of geography will be linked to those types.

LOCATION: Entity lists the location where physical files of data sets are stored.

MEDIA: Entity lists the types of storage media in use for keeping Census Bureau data. Information includes average life span of medium, size, and access speed.

SOURCE: Entity lists the sources for each data set. These include survey names or census name and year. This entity can be distinguished from *generic files* by noting that surveys go through periodic changes and redesigns, thus changing the data dictionary but not the name.

SUBJECT: Entity lists the various subjects which are described by data. Links are created to data elements. They are composed of the subject and a representation for the data.

R1: Relationship between *geography* and *subject*. Relates geography types with subject types.

R2: Relationship between *source* and *generic files*. Relates data sources with generic file types. One of three the main search paths to specific files.

R3: Relationship between *geography* and *generic files*. Relates the types of geography with generic file types. One of three the main search paths to specific files.

R4: Relationship between *generic file* and *data elements*. Relates the data elements to the generic file types. Completes the path from the subject types to the generic file types.

R5: Relationship between *subject* and *data elements*. Relates subject types to the specific data elements under that subject. One of three main search paths to specific files.

R6: Relationship between *generic files* and *data dictionaries*. Relates each generic file type with each item in the data dictionary of that file type.

R7: Relationship between *data elements* and *data dictionaries*. Relates the data elements with actual data dictionary items.

R8: Relationship between *generic files* and *file names*. Relates the generic file types with each of the files of that type.

R9: Relationship between *file names* and *location*. Relates the file name to the location where it is stored.

R10: Relationship between *file names* and *media*. Relates file names to the medium on which the file is stored.

R11: Relationship between *file names* and *contacts*. Relates a file name to persons or branches responsible for handling a file.

5. CURRENT WORK

Currently, a proof-of-concept prototype of the MDB is being developed. We are using a subset of *InfoSpan's Open Repository* software system. InfoSpan is currently the only company with a workstation class repository system which has been certified compliant with the IRDS standard. A decision as to which

complete system to buy will be made later this year. Money for this purchase has been made available through a Pioneer Fund Grant from the Department of Commerce.

The proof-of-concept system will include data sets from the 1990 Census and SIPP. These data were selected because of the availability of data dictionaries.

If the proof-of-concept system is successful, a larger prototype will be proposed. Building the prototype will require the cooperation of many people within the Census Bureau. It promises to be a large project.

6. CONCLUSION

The Census Bureau maintains many data sets collected through a series of censuses and surveys. The administration of many of these data sets is not under a centralized control. As a result, finding appropriate data for analysis and research can be very difficult. There is a centralized system for finding appropriate data files among the public release files, but no such system exists for internal Census Bureau files. Also, the information system for the public release files, ARRk, has limits to the amount of information it can provide.

Metadata are data about data. Metadata are required for the effective and efficient use of any data sets. At the Census Bureau, metadata are often poorly maintained and difficult to find.

The MDB is a metadata repository which when fully implemented will alleviate many of the problems associated with finding and understanding Census Bureau data. Currently, the MDB is in the proof-of-concept stage of development. A model of Census Bureau metadata has been developed, and it is the blueprint for building the MDB. The model may get more complicated as new features and functionality are added to the MDB.

7. REFERENCES

Almond, J., "Ideas for Information Types and Metadata Attributes," Center for High Performance Computing, Feb. 1994.

Bean, E., "ASCII Box Files for Data Portability," Census Bureau internal document, 1991

Bretherton, F., "Reference Model for Metadata: A Strawman," University of Wisconsin, draft, 1994

Sumpter, R. M., "Whitepaper on Data Management," Lawrence Livermore National Laboratory document, 1994

1. Information Resource Dictionary (IRDS) (x3.138 and FIPS 156).