ASPECTS OF DATA MANAGEMENT

by Ilona Einowski, Data Archivist
State Department Program
University of California, Berkeley

Like the traditional library, the data archive performs many different functions to meet the needs of users. These functions include data acquisition and cleaning, development of conventions and standards for description of the data, data processing and analysis, dessimination of information about the data, storage and maintainance of data tapes, development of an inventory system and inventory controls as well as a data retrieval system, diffusion of the data, training for archive users and program development.

Data Acquisition

Obtaining new materials for archive holdings from some continuing sources of supply requires establishment of both formal and informal arrangements with institutions, departments or bureaus that produce data on a regular basis in order to obtain some or all of their productions. It is also necessary to establish priorities for the kind of data to be acquired. Since the cost of processing and maintaining a data set is often greater than the cost of acquisition, selection must be made with great care. Ephemeral or frequently replicated data sets should be acquired only when there is a concrete need for them since there is a high probability of being able to obtain popular data sets elsewhere if a local need develops. The cost of acquiring, cleaning, indexing, and maintaining a data set should be considered in relation to:

(a) the likelihood of there being multiple users;
(b) the possibility of acquiring at a later date if the need should arise;
(c) its availability at a reasonable cost and with little delay from some other source;
(d) the amount of overlap with the existing collection;
(e) the intrinsic significance of the data.

The form in which data arrives varies from supplier to supplier and from study to study. This can result in a great amount of time being spent figuring out just what it is you have received. The ultimate answer is to have funding sources or institutions conducting the survey require that arrangements for archiving be made prior to the actual funding or conducting of the research. This way, the archive can be involved from the beginning and provide guidelines and standards for researchers. A formal way to handle this would be the development of an institutional policy on minimal standards for data to be turned over to the archive. A policy statement of this type would insure that the datasets turned over to the archive meet the criterion of methodological adequacy. It has been the case that an archive decided to pass up a study of great substantive interest which appears to have been done in such a poor manner, utilizing such sloppy and shoddy techniques of data gathering or documentation that, despite the interest of the subject matter, the data set is not worth acquiring. General archive operating policy should include a list of criteria which studies should meet if the data set is to be considered acceptable. At very minimum the following documentation should be available for each study:

-continued

-continued from page 23

(a) complete and accurate codebook or description of the data structure;
(b) description of the data format;
(c) illustrations of structure and format;
(d) total size of data set;
(e) complete and accurate description of the organization of the files for the medium in which the data is stored;
(f) precise definition for each data element;
(g) complete explanation of all codes used;
(h) sample of documents used in data gathering;
(i) description of sampling procedures employed, with intended and resultant sample size;
(j) summary of training provided field-workers and coders;
(k) description of data collection procedures;
(l) name and current address of study director.

Data Cleaning

In its most simplified form, data cleaning involves processes aimed at placing data into a format that is easily handled by computers. These processes include identifying and correcting possible discrepancies between the actual format of the data and the descriptions of that format. Many archives employ specialized staff members who do this type of data cleaning.

Development of Conventions & Standards

In order to facilitate the process of utilizing data initially prepared by others it is necessary to establish conventions for coding and standards for describing the data themselves in order to:

(a) permit combining information from different collections in some reasonable way;
(b) combine samples from different studies in order to increase the number of cases;
(c) make comparisons among data sets;
(d) facilitate later analysis.

Data Processing and Analysis

The data processing and analysis function of the archive provides for the manipulation of data for the user's purpose. This may entail the reformatting of data for use at the user's local facility or providing specially prepared subset of cases or variables rather than a simple copy. Some users may need a frequency distribution for the variables (if not provided in the codebook) or simple cross-tabulations. Other users may need more detailed statistical analyses.

Dissemination of Documentation

The most important documentation produced by the archive is the codebook describing the dataset. Archive staff also prepare abstracts of data sets for inclusion in a catalog and for advertising purposes. Production of some type of archive catalog is almost mandatory since it provides not only an in-house listing of current holdings but is also the best way for a user to browse the contents of the archive. Advertising the availability of the data can take many forms. Some archives prepare and distribute their own newsletter announcing new acquisitions while others include a special data announcement section in an existing institution newsletter. Archives should also strive to maintain a collection of published material related to the data sets in order to provide examples of how the data have been analyzed already and clarify ambiguities in the interpretation of the data.

Storeage and Maintenance

Internal procedures must be established to identify the current storage location of all materials. Magnetic tapes must be stored in a controlled temperature environment and protected from magnetic flux and

physical shock. They must be recopied on a periodic basis in order to assure their continued utlity and, where usage is heavy, to protect against deterioration due to machine-induced wear. (See Patricia Reslcok's article for a detailed discussion of tapes.)

Inventory

As with any collection, it is necessary that the archive maintain a catalog or index of holdings. The archivist might consider maintaining a "public" catalog and an annotated "private" catalog with additional information. The "private" catalog would include abstracts of studies added to the collection since the last published catalog update. It is also necessary to develop an internal inventory system to keep track of the current status of all studies in the archive including studies "on order" or being processed. Other internal inventory materials would include a catalog of tapes by tape or storage number and a catalog of studies by study number.

Retrieval

Requests from users for access to data relating to their particular topic of interest often requires the archive to search not only its own holdings but those of other archives as well and where necessary, to obtain from other archives those materials required to serve the needs of the user. For this reason it is advisable for the archive to maintain a collection of catalogs from other archives and to become familiar with the general class of holdings at other archives. Most archivists find it helpful to maintain personal contact with other archivists through the network established by professional associations (like IASSIST) in order to facilitate the exchange of information about data holdings and to keep abreast of technological developments in this field.

Diffusion

The data archive specializes in copying its own collection and making it available to the user at his convenience, in the form most suitable to his purposes. However, the archive must still maintain control over access to their materials in accordance with any wishes of the original donor. For this purpose the archivist usually develops a form letter which the user signs agreeing to archive terms. Once the data have been copied, the archivist fills out a standard form to send along with the data tape which describes the files on the tape...number of files, logical record length, blocksize, number of records...as well as general tape characteristics...tracks, density, format, character set, and internal labeling. Shipping data on magnetic tape requires that the tape be adequately packaged to prevent damage in transit and labeled on the outside of the package as being a MAGNETIC TAPE with a warning to keep the package away from magnets and electric motors which could destroy the data set stored on the tape. Tapes sent through the mails are often insured for the cost of recopying the data and have a return receipt included with mailing.

Training

Archives can perform a training function by teaching users how to make a query; where to make a query so that the appropriate data can be obtained; how to utilize the data once it is obtained; the devices available for processing; the strategies to be employed for analysis; and the kinds of interpretations that can be made from such analysis.

The archivist may wish to prepare a user manual for distribution to potential users documenting how their archive is organized and including information on archive services and locally available

-continued on page **29**