IASSIST Newsletter, Vol. 3, No. 2 (Spring 1979)

A MODEL FOR USER'S SERVICE:

PROVIDING FOR INFORMATION AND DATA RETRIEVAL

FROM AN ARCHIVAL, USER, AND DEVELOPMENT LIBRARY

Harriet A. Dhanak
Christopher D. Brown
Michigan State University

Although the Newsletter has carred several articles dealing with the various operations of specific archives (The Roper Center, NORC, etc.), the organizations reviewed were establishments devoted primarily to providing archival and library services. Many of the members of IASSIST, in contrast, work alone or with limited assistance, within the context of an academic (or other) department. The article which follows was first presented at the 1979 IASSIST Annual meeting in Ottawa and is a description of the way in which a departmentally organized archive is confronting its work.--Editor.

INTRODUCTION

This paper will deal with the problems of information storage and retrieval associated with a machine readable data archive located within a teaching/research department, namely the Political Science Department of Michigan State University. A model, in the process of development and implementation, is described that hopefully will alleviate the problems we have encountered. The mode of delivery of copies of machine readable data files will be discussed as well as modes of user information storage and retrieval.

Some subjects germane to the topic, such as cataloging machine readable data files in a traditional library, have been dealt with extensively elsewhere and will not be discussed here. While there are generic problems facing the archives that deal with machine readable data files, this paper will focus on the ones that are central to our users and functions of the archive. Points to be developed and discussed are the characteristics of the user community and the Michigan State University computing system plus our tasks and needs and those of our user community.

LOCAL SITUATION

Because the Political Data Archive is located within a teaching and research department, our contact is with researchers, faculty and students, during the period that their work is in progress. Through the Politometrics Laboratory also located within the department, we will either conduct the computer runs or assist the users in all aspects of the computer applications. Therefore, we find it necessary to address their needs as well as needs of the archive.

The present situation on our campus is:

1. We are the only archive of machine readable data files of cross disciplinary material on the campus.

2. The main library is not in the immediate future planning to hold our types of files. At present the code books are not located in the main library, but are in a library located within the Political Science Department.

3. The archive also serves the College of Social Science and the university community as the supplier of data sets from the Inter-University Consortium for Political and Social Research.

4. The Michigan State University Computer Laboratory maintains a Control Data Corporation 6500, with an Inter Data that handles the terminal I/O. A CDC 6400 is available on a limited basis for batch work. A Hewlett Packard 2000 is maintained primarily for instructional purposes.

5. The Computer Laboratory offers consultation on system problems, computer language (Fortran, Cobol, etc.) execution problems and for analysis packages such as Statistical Package for Social Sciences and the MSU STAT package.

6. An interactive multipurpose computing package is not maintained by the Computer Laboratory.

7. Within the department, we need the use of special analytic programs that must be updated with computer system changes and modified for research needs.

RESEARCH PROBLEMS

One problem both for researchers and others using a computer system for information manipulation that differs from a traditional library use is that the functioning unit or system is under constant change, that is, the enviornment within which one works is constantly changing. The computer system itself may be changing with a new model or even worse a new vendor. The existing computer system is altered, hopefully upgraded, and is not necessarily upward compatible. New programs/methods are introduced on an existig system, and the changes may prevent previously executable program from running.

This situation presents computer-oriented researchers with an important cost in personal and professional time just to keep up with the changes. This time is above the time consumed learning to use computers as a research tool. For those who constantly use a computer the information is easily recalled and current. However, most researchers will use computers episodically and the development of our model is partially dictated by the characteristics and needs of this type of user. Those using a computer must familiarize themselves with some aspects of their use, at least analysis packages such as SPSS. We would hope to assist such users by enabling them to obtain a data set with a minimal knowledge of tape handling proce-

dures, assuming that the data set is on tape.

ARCHIVE PROBLEMS

Several authors have addressed some of the issues of special interest to us, with most noting common problems. White (1974) discusses and cites literature that compares the varyng functions between libraries and archives. The discussions deal with acquiring libraries and those concomitant problems, not with data management problems associated with research or secondary analysis. He cites the difficulties facing researchers in gaining information on the machine readable data files available.

The implementation of our system will expand the information on available data-sets, and will also archive user library services to the university community. Our aim is to break the pattern of information access noted by White that one needs personal contacts to gain knowledge of data set availability.

Ferguson (1977) notes the types of questions data file users ask of library and computing organizations. Their interest is not in partitioned responsibility between the library, computer center and archive, but in the availability, access and documentation of data files.

Grandon (1978) discusses archive development stages in general and then details the system under development at the Social Science Data Center/Roper Center. His paper describes functions in which all machine readable data file archives must participate to a greater lesser extent. The one aspect that we will focus on is his suggestion of a machine readable index of holdings. He details the SSDC method of keeping tape file information on a card image file. We have already used that method, and found that it was not adequate for our needs. The card image file could be augmented when new files were created for a study, but basically it was laborious and could be inaccurate.

All archives, large or small, with all our variations, are facing a similar problem: The basic problem is the storage and retrieval of various levels and types of information. We have probably all gone through the same process of first producing a hard copy list of holdings -- typed at first, then later placed in machine-readable form. Many archival holdings and most user data sets were on cards and were laboriously carried around with notes and information written on the cards or across the top of the deck. When data sets were/are stored on tape or disk, a user can no longer visually inspect the data and an uneasy feeling sets in. This is the period that technological change in data storage outstripped the means used to document files and studies.

This detailing of developmental problems in archiving does not even begin to confront the enormous task of indexing variables across studies. Since we do not have the resources to deal with these problems, we will move on to those that are presently manageable. We hold about two hundred and fifty studies, and have a library of about three hundred tapes that probably contain about two thousand files. We also have the usual need for the dissemination of information on the studies available and have a program that lists them as card images

by subject and author within subject. Archivists and researchers are faced with the problems of increased complexity of machine readable data files, and with multiple files associated with one study title.

A survey might consist of one rectangularized file of one to two thousand card images, or may contain tens of thousands. The latter presents only problems of bulk. A survey may be updated (new editions) by cleaning or for other reasons. Panel studies present additional difficulties by having the problems already noted as well as the addition of new waves. To this list of complexities and problems is difficulty of hierarchical files.

At present our biggest problem is to store, retrieve, and have available information on tape files and a machine retrievable document for each study with a description of each file. Turtle (1978) also notes that file documentation was deficient in the library at the Tennessee Valley Authority and proposed procedures to improve the existing documentation system. We differ from her in that she states "a computer tape library ... contains only one physical format for information". A serious problem for our archive and users is that we may have multiple formats for files of one study. These may include forms usable by a batch-oriented analytic package while the file may also be stored diffecently for on-line terminal use.

Clerical procedures, for tracing files and problems with tape, (such as parity errors or the need for cleaning), require an inordinate amount of time and become insufficient and redundant. File documentation for studies with only one file that may be updated is not difficult and a method is suggested by Carter and Roistacker (1976) that is enforced by the Social Science Quantitative Laboratory at the University of Illinois. However, levels of information and documentation plus a prohibition of duplicate names (a search is conducted as each entry point for dups) will give our users more flexibility.

MODEL

The model that we are developing:

1. assumes minimal computer-associated knowledge by users.

2. will be a method for information storage and retrieval of documentation and/or numeric data.

3. will be a process that will allow modification of the stored indexes* and the programs executing number two above without having to redo the stored sets of archival information.

This model can be viewed as a multipurpose system that will serve users who:

1. are browsing for information by author, title subject (general) and

---

*This refers to archival documentation records as opposed to codebooks and other hard copy documentation describing a specific data set.

geography;

2. are searching for a specific author, subject, or geographical area;

3. may want to retrieve a file that they specify from tape or disc;

4. have the need to create their own library;

5. wish to hold information and files in the developmental section of the library for creating temporary files and the documentation associated with them.

The system will enable archival maintenance and updating of all material, and logging and accounting records of all activities.

After trying some of the approaches previously mentioned, it became obvious that we need a comprehensive approach to the management of information on studies and their associated files, that would enable us to actually simplify our procedures. One needs the efficiency of only doing a job once, not multiple times, and when using student help one needs an ongoing system that does not require more explanatory time than executing time.

Frequently not enough time is spent in designing a system, which leads to implementation problems. We have spent much time, thought, and planning for a model that will serve well in the context within which it is anticipated it will operate.

The development or use of a system such as RIQS (1975) with tutorials would be very helpful to new users, but at the present time it is beyond the scope of the system we will be implementing. A CAI is under development at our institution, and will be used if suitable to our application. We plan to develop a simple querying system, not a sophisticated one.

Our interest in a developmental library and in allowing users to create their own libraries comes from working with researchers and classes and attempting to keep track of all the tape and file activity. As with many aspects of research work, documentation of work in progress is one of our biggest problems. Although the problem of classifying, cataloging and documenting studies has been dealt with in the appropriate literature, there is not, to my knowledge, any (or much) literature on systems designed to deal with the information problems associated with developmental file systems for users in an academic setting.

Those who have large grants may be able to hire their own staff to keep records, but most university researchers must keep their own records of their work as it progresses to completion through various files. At a later period the information on files could remain on a permanant record and accessible to them or it could be scattered.

A user creating their own library will be given the flexibility of entering all information and documentation that is standard for the system and will have the choice of limiting or permitting public access to any or all levels of information. Ferguson (1977) has noted the reluctance of researchers to enter information on their studies into library or archive records. Our system will allow them the privacy they require and a

method of documentation they need and a history of their studies previously entered.

The information to construct the indexes of information resides on several files. The files contain information in a hierarchical structure with imbedded fields to link them together. This allows us flexibility both with the files (used as data) and programming, that is, one or both can be modified independently.

By using the hierarchical file structure of the holding files, we plan to include a list of studies availabe to us, but not located in the archive. This will enable the university community to know all the ICPSR studies listed in the User's Guide and could be easily updated when we receive periodic notice of new holdings. If we obtain a membership in the Roper Center, that information would also be entered.

At the on-line querying stage any user will be given the information needed to submit a batch job to obtain a copy of the required data. Tape numbers and their necessary information (tracks, density, character mode, etc) and file position will not be needed by the user to request a file. File location, whether disc or tape, will be retained in the index and the file retrieved with a simple command followed by pertinent information. For various policy reasons, the computer laboratory at MSU will not allow a request for a tape mount to be executed when one is on-line at a terminal. Although it is limiting when one needs a small file, it is not as restrictive as it appears because when one is working with a large data set, batch mode is generally preferred due to both time and cost factors.

Users running under archive supervision could have a job submitted for them to retrieve a copy of a tape file. We will also have a pool of tapes available for storing their analysis data and the indexing information entered while the job is execution, thereby relieving them of the problem entering minimal information to the index at a later date. At this time (1979) we have four files of information linked in a hierarchical manner, plus a subject file, an authorization file and a documents file.

The holdings file contains information on the type of system (Archive, User Library, Development). A twenty character acronym will be the index reference for a study. Authors, title, subjects, geography, and date are information that will be used for searching and printout. Read and alter passwords can be set at this level.

The mapping file contains information on the version (edition, set or subset), and as many as necessary can be established for a study. This will cover multiple files, updates, etc. This level will also offer a restriction code if needed. Date, document, and other user information will be held here.

The access file has information on the form of a file (coded, binary, card deck, book, SPSS file, etc). The file location (tape reel number and file position and position of reel in a multiple reel file), and number of accesses. There will be as many forms of a version as is necessary. General users will be restricted (except with permission) to accessing only archive files that are in standard format that match the codebooks. Other form information will be for archival information only.

The tape file will contain the standard tape information, e.g. tracks, density, label, mode, owner, plus a tape history of accesses, cleaning date, and number of mounts.

The authorization file will contain access level, accounting information, logging information and scheduling action if necessary. This file will provide information for report generation. The scheduling algorithms will be particularly helpful for instructional files.

The document files contain the standard information on each study: number of cases and variables, sampling design, abstract, issuing archive, published material, etc. We are still formulating a design for a version document section. We plan to have documentation accessable in several modes and are still designing this phase.

Those studies under archive entry and control will have all of the specified information entered. To store a file a user may enter all of the above information for a permenant file, or enter only the acronym, version and form for temporary files. In this case default information will be entered by the archive system.

CONCLUSION

This paper has described an information system still under development that will solve some of our problems of maintaining a machine readable data file archive in an academic setting. Much planning and programmming effort has been expended to date on this relatively small problem. If the College of Social Science, Computer Center, and main library choose to build an infrastructure similar to those units found at Stanford, Wisconsin and Northwestern among others, we believe that the existing files of information in combination with the executing program structure will serve as a base for expanded holdings and services.

One is left with a frustrating feeling that there are developments occuring elsewhere that would be of use to us and coversely, others might use our efforts to their benefit if we could surmount the major problem of program transportability to other vendor systems. We seem to be "reinventing the wheel" at many institutions.

References

Carter, M. C. and Roistacher, R. C. Statistical and Data Support to a Heterogenous user Community. First International S. A. S. Users Meeting, 1976.

Ferguson, D. Social Science Data Files, The Research Library and the Computing Center. Drexel Library Quarterly 13(1977), 70-79.

Grandon, G. M. An Archive Development System: Specification and Progress. 1978 IASSIST Annual Conference.

ittman, B. and Borman, L. Personalized Data Base Systems. RIQS--Remote Information Query System (Los Angeles: Melville Publishing Co., 1975), Chapter 2.

Turtle, M. and Smart, C. W. The Role of the TVA Libraries in Digital Data File Documentation: A Demonstration Project. 1978 IASSIST Annual Conference.

White, H. D. Social Science Data Sets: A Study for Librarians. Ph.D. Dissertation. (Berkeley, California: University of California, 1974), Chapters I, II, and III.