Evaluation and Appraisal of University Administrative Computing Datasets

by Mark Conrad[1]
Data Archivist
Pennsylvania State University

Introduction

In September of 1990 the Pennsylvania State University Archives in cooperation with Management Services, a division of the university responsible for administrative computing services, began a two year grant project funded by the National Historical Publications and Records Commission (NHPRC Grant #90-095). The objectives of this project are to appraise, preserve, and make available electronic records created, stored and used in the Management Services Division of Penn State University; to develop ongoing procedures for the appraisal of administrative computing data in the future; to develop protocols for the use of the data by institutional and outside researchers while enforcing restrictions on access for privacy and confidentiality purposes; and to provide recommendations based on the project for the preservation of archival data from on-line administrative database systems.

The project was divided into four phases. The first phase lasting two months was used for the orientation of the data archivist to the operations of the University Archives, the University Records Management Program, and Management Services. Phase two, the phase we are currently operating in, is used for the appraisal of datasets. This phase is scheduled to last 18 months. Phase three and four are scheduled to last the final four months of the two years. In phase three recommendations will be made for the identification and preservation of future archival datasets and protocols will be developed for the research use of the datasets. In phase four reports on the project will be prepared and circulated. In actuality, much of the work scheduled for phases three and four has already begun and is being carried out concurrently with the appraisal process.

This paper will focus on the second phase of the project. I will discuss the appraisal process as it is currently being carried out, difficulties encountered and lessons learned to date.

The Appraisal Process

For the purposes of this project the appraisal process will be confined to a finite number of datasets from the university's administrative computing mainframe. We will not be examining electronic records from other mainframe, mini or microcomputer systems. Rather, the appraisal process will be limited to some 3,000 datasets recorded on "history" tapes. History files, in Management Services' parlance, are usually copies of master files at a particular point in time (often the end of a semester or academic year). The datasets would be very difficult to recreate if they were destroyed because they are copied from files that are constantly updated. These files are kept for possible reuse [2].

The datasets date from the late 1960's to the present. These datasets contain information for fourteen areas of administrative responsibility at the university. The areas are: Accounting, Payroll, Bursar, Student Aid, Agriculture, Planning and Analysis, Budget and Resource Analysis, Management and Systems Engineering, Admissions, Registrar, Testing Services, Development, Graduate School, and Physical Education.

Each area has a data steward who "develops the coding structure of the data, insures the data's accuracy, determines the frequency of updating, and establishes data use and protection requirements."[3] The data steward is usually a senior administrator, such as the Registrar, or that persons' designate.

The appraisal process begins by selecting a data steward's area of responsibility. The data archivist must have the written permission of the data steward to examine the datasets under his/her control. In some cases datasets are jointly "owned" by more than one data steward. In such instances the disposition of the datasets must be discussed with all interested parties. In starting the appraisal process I chose data stewards that had only a few history tapes to test the procedures developed for the appraisal process and the database system to track appraisal information.

Having chosen an area of responsibility, the next step is to identify those datasets that belong to the data steward from the (Management Services) Tape Library listing of the history files. Once identified the datasets are grouped by common dataset name. This is done because datasets with the same name usually contain the same types of data and can be initially evaluated as a group.

The next step is to locate as much information about the datasets as possible. I am to find the procedure and program that created or used the information, any documentation, a file description, a record description, record counts, or any samples of input or output. Records Management Program retention schedules are checked to see if similar records in another format have already been scheduled. If similar records have been scheduled for destruction and the datasets do not have some additional value by virtue of their being in electronic format and thus more manipulable, the datasets may be recommended for destruction. (Junk is junk no matter what the format.)

While the search is on for information about the datasets, some of the tapes are read and printouts are made of a sample of records from each file. This serves several purposes. Firstly, it verifies whether or not the tapes are still readable. Some of these tapes have been in storage for a very long time under less than ideal environmental conditions. Secondly, the dump can be used for comparing what's actually on the tape with what the documentation says should be on the tape. If the two don't match, other documentation must be located or the file may be recommended for disposal. A file is of no value if a determination cannot be made as to where one field ends and the next begins or as to what a particular value in a field indicates.

Having gathered as much information about the dataset as possible, the next step is to interview the data steward and/or a contact designated by the steward about the datasets under his/her control. A standard list of questions has been developed to help the data archivist gather all the information necessary to make an informed appraisal decision.

Those questions are:

> Do you have documentation for these files?
>
> Do you have samples of input and output for these files?
>
> Where did the data come from?
>
> What was it used for?
>
> Is it still being used?
>
> Is it updated?
>
> How often?
>
> Are the records maintained in another format?
>
> Has the other format been scheduled for retention or disposal?
>
> Are there any requirements for the retention of this data that you are aware of?
>
> Are there any restrictions on the use of this data that you are aware of?

At this point the data archivist should have enough information to begin making the appraisal decision. The decision-making process is not that different from the process for more traditional records. It is certainly not very different from the process used by archivists working in the electronic records programs in government archives.

> Does the dataset have legal, evidential, or informational value?
>
> Is this dataset unique?
>
> Is it the most desirable format for keeping the information?
>
> Is the data hardware/software independent?

If the answer to all of these questions is yes, all datasets that share the common dataset name and structure will be recommended for accessioning by the Archives. Retention schedules are developed for all datasets, regardless of their status, in concert with the data steward, Management Services, the Records Management Program Staff and the University Archives/Records Management Advisory Committee.

Once the decision has been made that a group of datasets are archival, each dataset is read and a data dump is obtained. As with the sample of datasets read previously this is to verify that each dataset is readable and the data is valid. File structures and record descriptions change over time so the data archivist must insure that data from each dataset is adequately documented so that a researcher or the archives staff can use it. Assuming the datasets are readable and understandable, copies are made of each dataset and the relevant documentation. The datasets are accessioned by the University Archives and the data archivist turns his attention to the next set of files.

Problems Encountered

You may have already gathered that one of the biggest problems has been locating adequate documentation for many of the datasets. The record description for many files seems to change on a regular basis. Often the documentation is not updated to reflect these changes or conversely when the documentation is updated previous versions are discarded despite the fact that files still exist that were created using the previous documentation. It is not unusual to find a number of files with different file structures, the same name, and one set of documentation that may not match any of the files. In talking with other archivists working with electronic records, I have been assured that Penn State is not alone in this predicament.

Management Services is currently exploring the possibility of recording documentation for a dataset directly onto the first label of the tape a dataset is recorded on. As long as the documentation is copied along with the dataset whenever the dataset is transferred to new media, the proper documentation should be available for the life of the dataset.

Another related problem occurs when trying to appraise an older dataset. Often there are no employees still working in the office that used the file who remember what the file was used for. Sometimes the office itself no longer exists! The turnover on a university campus and the restructuring of administrative units can make it difficult to find someone who can tell you how a file was originally used or what dataset replaced the one you are evaluating. The only solution to this problem is to carry out the appraisal early in the life cycle of a dataset.

Lessons Learned

One of the most important lessons we have learned is that the shorter the time lapse is between the creation of a dataset and its appraisal the easier it is to identify archival datasets and insure their preservation. The archivist can interview all the players involved in the creation of the records to better understand why the records were created and under what circumstances. Documentation can be evaluated to insure it adequately explains the data so that it will be useful to future researchers. Datasets that are identified as archival can be marked for special handling to insure the data will still be readable 10, 25, or 100 years from now.

Another lesson we have learned is that the cooperation of the administrative computing center is essential to the success of an electronic records program. The archivist needs to understand how data is manipulated at the center to meet the informational needs of the institution. The administrative computing center personnel must have an appreciation of the potential value of the data beyond the purposes for which it was originally created. Any archives considering implementation of an electronic records program would be well advised to begin building relationships with their institution's administrative computing center(s) now.

The most important lesson we have learned is that more records are being stored in electronic format all the time. If we do not identify and preserve the archival datasets a large portion of our institutional memory will be lost. At Penn State we have begun the process of insuring these valuable records will be preserved, we encourage other institutions to join us, and we are happy to share information about our project.

Footnotes

[2] Pennsylvania State University. Management Services Division. Standards and Procedures Manuals (on-line manual).

[3] Pennsylvania State University. Administrative Policy, AD-23.

References

Pennsylvania State University, Management Services Standards and Procedures Manual #3, Chapter 5, Section 1. (on-line manual)

Pennsylvania State University Administrative Policy, AD-23, p. 1.

[1] Paper Presented at IASSIST 91, 15 May 1991, Edmonton, Alberta, Canada.