A PROPOSAL FOR THE CONTENTS OF A PROJECT ARCHIVE
Ronald W. Helms and James D. Hosking, University of North Carolina

Prologue: The Funeral Dirge

The process of archiving data and other materials from a project is often like the activities that surround the death of a person. Just as a dying person may lose the function of one organ system after another, a dying project is often disabled to greater and greater extent by the loss of key personnel. When a person reaches the deathbed it is too late to buy life insurance to provide a sizeable estate; when a project is in its last stages it is often too late to arrange for appropriate archiving of the projects; valuable data, or "estate." Just as a person's survivors are required to make funeral arrangements and settle the estate, a project's survivors--those who haven't left the institution--are often left with the responsibility for the "lost remains."

This morbid analogy is actually useful. Like people, every project is certain to "die." Whether the project leaves an "estate" in the form of archived data of potential scientific value is dependent upon the extent of "estate planning" during the vital stages of a project. We take a more cheerful view than the somewhat morbid author (unknown) who wrote, "Life is just one long path to the grave." In contrast, we feel that prudent project managers will allocate appropriate resources to "estate planning" during the vital stages of a project. When the end comes, the survivors will have already completed many of the "final arrangements" and can hold an Irish wake, cheerfully acknowledging the project's accomplishments and secure in the knowledge that the "estate", or archive is in good condition.

1. Introduction: Objectives of Archiving

There are several levels of archiving which one can consider. Two principal ones are report/paper level and project level archiving.

The motivation for record/paper level archiving can be summarized succinctly:

- For a reasonable period of time after its appearance the authors of a paper or report should be able to reproduce any statistical computations presented, regardless of the format--statistics, charts, graphs, etc.

A paper/report level archive is created for the purpose of saving those materials needed to reproduce the computations.

Paper/report level archives are discussed in another report; the topic is raised to distinguish such archives from the topic discussed here, project-level archiving.

A project-level archive stores materials from an entire project rather than just one report or paper.

The objectives of a project-level archive are: to store the project's scientifically important data and related materials in such a manner that other scientists can readily learn and understand the project's scientific environment, objectives, and methods and can readily access the data.

An archive consists of the datasets, documents, and other materials needed to satisfy these objectives.

A project archive can be organized in three major components:

- Project documentation
- Dataset documentation
- Datasets

These three components are described separately in the following sections of the report, with greatest emphasis on the first two.

*The authors' experience and interest lie in the area of projects involved in scientific research. Persons interested in archiving materials from other types of projects should not only change "scientific" to a more suitable adjective, but should also judge the extent to which the principles expounded here are appropriate in the field of application.

2. Project Documentation

To fulfill the archive's objectives the archive must contain sufficient information for "another scientist"--one not associated with or knowledgeable about the project--to readily learn and understand: the project's science; its data collection instruments and methods; its history; and other factors influencing the value and interpretability of the data. The project...
The project documentation can be organized into the following sections:

- Scientific documentation
- Data Collection Instruments and Methods
- Limitations of the Study
- Other Factors Affecting the Value of the Data

These sections are described in detail in the following text.

2.1. Scientific Documentation

For an outside scientist to understand the "science" of a project, he must learn scientific background, objectives, methods, and other pertinent scientific information.

Scientific Background. The objective of a research project is to answer specified "scientific questions." However, to understand the full meaning of these questions it is important to understand the context in which they are addressed, the scientific background of the project. The scientific background section describes work that has gone before, work that has led to the point at which the project is initiated to answer, questions which sprang from preceding work.

The scientific background section is usually not written by data management personnel. This material is usually available from other documents. For grant-supported projects, the grant proposal will contain an appropriate discussion of the scientific background, possibly under the heading "Review of the Literature." Other sources include annual reports, final reports, and "Literature Review" sections of published papers and reports.

Typically the archivist will have to assemble materials from the various sources and edit the collection into a coherent description. The result should be reviewed by scientific personnel.

Scientific Objectives. An outsider must understand the project's scientific objectives in order to fully understand what the project was about and to understand how data might be appropriately used.

Most scientific projects have clearly specified scientific objectives. Statements of the objectives may be found in the sources noted above: grant proposals, annual and final reports, published scientific papers and reports.

Again, the appropriate materials can be assembled and edited into a coherent summary of scientific objectives, which should be reviewed by scientific personnel.

Experimental Design and Scientific Methods. A thorough understanding of the experimental design and scientific methods used to generate the data. For example, a study "lung capacity" in children could use any of several measurement methods; different methods produce different types of characterizations of a person's "lung capacity." "Lung capacity" is not a well defined term, but each of the methods produces well defined characterizations of lung capacity. Each method has its scientific advantages and disadvantages, its characteristic accuracy and precision, and so on. To understand the data one must understand the underlying scientific methods.

The experimental design concerns the selection of subjects or "experimental units" and the method of assigning experimental units to different "treatments." With human subjects, for example, there is a huge scientific difference between (a) studying the first N available subjects who meet the study's admission criteria and (b) selecting a random sample from some well defined population of subjects who meet the admission criteria. There is also a huge difference between a randomized design, in which experimental units are assigned to "treatments" at random, and an "observational study" using non-random assignment.

Thus, the experimental design, as this term is used by statisticians, is an important aspect of the project's scientific methods.

The material describing scientific methods can be assembled and edited from the sources noted above: grant proposals, annual reports, final reports, published papers and reports.

Scientific Limitations. Every scientific study has scientific limitations; some are obvious but perhaps some are subtle. Outsiders, particularly after a period of time, may have difficulty ascertaining the more subtle scientific limitations unless these are documented in detail. Such limitations are usually not noted in a project's documents, perhaps, because authors are more interested in asserting the positive results than in describing limitations.

Limitations arise from many sources. For example, a "sample of convenience" from a human population supports much more limited inferences than a random sample. Some experimental designs, due to confounding or limiting regions of experimentation, are more limited in scope of inferences than others. Some types of scientific measurement are more limiting than others. For example, a Hollingshead socioeconomic scale value is not as economically informative as a subject's taxable income: some types of intelligence scores are valid for only selected ethnic groups, etc.

Data management personnel are typically not qualified to write the scientific limitations section. This material should be written by scientific and statistical personnel early in the study. The resulting document can be extremely useful in the preparation of
If such a document is not available when archiving is begun, the project's highest level of management should be convinced of the value of compiling the document and should assign appropriate resources to get the document prepared. The project's statistician may be a powerful ally in convincing management of the value of the scientific limitations document.

Other Important Scientific Factors. Although the above-described sections will contain much of the important scientific documentation, each project will have other types of scientific documentation important to an in-depth understanding of the data. The archivist is responsible for determining what additional sections are needed and obtaining the necessary text.

2.2. Data Collection Instruments and Methods

The archive must contain complete documentation of all data collection instruments and methods used in compiling the data being archived. The objective is to provide a potential data user with a complete description of how the data were obtained.

This section will contain one section for each data collection instrument/method used. All data forms used and their associated instructions should be described and included, in this section or an appendix, if multiple versions of a data form or related instructions were used, all versions should be described and included. In addition, a clear description of an algorithm should be included which describes how to determine which data records were obtained from each version of the forms.

When machine-readable data are obtained directly from automated instruments; complete documentation of the instruments and their methods of operation should be included unless this information is readily available from public sources such as libraries, in which case only a summary and a reference is needed. Local modifications or additions to "standard" instruments should be described in detail.

The information for this section should be developed throughout the life of the project. It is often difficult or impossible to obtain this information near the end of the project.

2.3. Project History

A summary of all the important events and dates in a project's history is very useful to an outsider trying to understand the project's data. Project personnel operate from an information base which includes most of the important points of project history. The difference between having a full knowledge of this history and virtually no knowledge of it is often not recognizable.
The principal factor to consider for each dataset is its scientific merit (unless there are legal obligations, discussed below). The question, "Does this dataset have sufficient scientific merit to warrant expenditure of the resources necessary to archive it?" calls for a scientific value judgement which must be made by project management.

Some other factors to be considered are discussed below.

Legal obligations. In some situations there are legal obligations to archive data for a period of time. For example, virtually all data collected in support of licensing new prescription drugs must be archived. In addition, more and more research is performed in areas in which the results may have regulatory or commercial implications. Such studies increasingly lead to litigation or regulatory proceedings in which data archives may be important or necessary. (There is additional discussion of this topic in the section "Documentation of Data Processing History, Procedures and Programs," below.)

Confidentiality. An archived dataset which contains confidential data on humans and personal identifiers is a potential source of serious problems. Archived datasets typically cannot be protected as carefully as active datasets. There is rarely a necessity to retain individual identifiers in archived data. An archive which contains several sets of data from the same subjects' pseudoidentifiers, which permit linking records from different datasets, can replace personal identifiers. Some institutions use pseudoidentifiers in this manner for all analysis files, and many master files. The personal identifier-pseudoidentifier linkage is maintained in a special, well-protected dataset, which need not be archived.

Potential Useful Life of Data. In some situations the potential useful life of one or more datasets is too short to warrant the expense of archiving in a project archive. Such data may be archived in a report/paper archive for an appropriate duration. This situation can arise, for example, in a study comparing different measurement methodologies. If one of the methods is demonstrated to be clearly inferior data from that measurement method may have little or no scientific interest.

Data compiled for project management purposes may have little or no scientific interest. For example, a study depending upon laboratory measurements may have a very active quality control program to monitor accuracy and precision of the laboratory instruments. While a summary of the results of such a program is an important part of an archive (see a subsequent section addressing the data quality issue) the detailed quality control data may have very little scientific value.

The selection of datasets to be archived is an important task with important consequences for a project's resources. Decisions should be reviewed by scientific personnel and approved at the highest level of project management.

3.2. Selection of an Archival Medium

Selecting the medium for storing archived datasets is perhaps one of the most comfortable decisions for a data management oriented archivist because the alternatives involve familiar considerations.

At this time the primary possibilities for selection are:

- magnetic tape
- magnetic disk
- punched cards

and a near future choice might be

- laser(video/data) disk

The laser disk, when fully marketed, will be a special version of the already familiar laser video disk. The video disks are mass produced by pressing a "blank" disk between two patterned dies. The computer version will apparently involve a laser/write mechanism which will use a strong laser to burn holes ("bits") into a layer of material under a plastic coating. Data can be recorded only once; the burning process is irreversible. These devices are not widely available at the time of this writing.

Some factors to use for selecting a medium are:

- Compactness
- Cost of recording data on medium
- Cost of maintenance
- Stability/long life
- Robustness to environmental extremes
- Will there be hardware and software to process the volume?

C.M. Dollar (1980) compares the various storage media for archival purposes.

The first choice at the time of this writing is magnetic tape, with appropriate provisions for environmentally controlled storage and for tape exercising (unwinding, rewinding) approximately annually.

One of the reasons tape is preferable to disk is the rapid change in disk recording methodology. Only 10 years ago IBM 2314 disks/drives were a standard fixture in many computer centers. Today, many centers do not have the hardware to mount an IBM 2314 disk pack. One can expect similar rapid changes over the next decade. The pack archived today may be so obsolete 10 years from now as to be unusable.
3.3. Selection of a Dataset Format for the Archive

In this context, dataset format refers to the format of the dataset rather than to the format of the records in the dataset. Examples include EBCDIC "flat file" datasets, SAS datasets in SAS databases, SPSS SAVEFILES, etc.

We feel it is important that an archive be stored in one of the self documenting formats as, for example, SAS datasets and SPSS SAVEFILES. Such datasets contain at least two parts, one containing data descriptions: type, length, format, labels, etc., of variables. While this documentation, by itself, is totally insufficient for archive purposes, it is absolutely essential. By choosing a self-documenting format one ensures that these essential descriptions remain with the data values.

The dataset format should not depend upon software which is esoteric and/or may cease to be available or supported, for obvious reasons. Similarly, the software should be relatively stable, with assured long term compatibility. For example, SAS disk datasets may not be in a long-term stable format. Although SAS 79.5 supports reading of SAS 72 disk datasets, rapidly changing disk hardware technology, and the obvious need to support new, more efficient devices, could lead to a lack of "upward compatibility" and a version of SAS that would not read SAS 79.5 datasets. (This discussion is centered on SAS because of the context of the paper; similar remarks apply to dataset formats produced by other software.)

SAS tape datasets have a particularly simple structure, amenable to processing by PL/I or FORTRAN programs. For this reason, and because tape hardware is changing less rapidly than disk hardware, SAS tape datasets, recorded at the highest available density, are a good hardware/software choice for archiving datasets at the time this is written.

Once datasets are selected for archiving, and the medium, format, and software have been chosen, creating archival copies is a simple matter. Documentation of the datasets, the topic of the next section, is not so easy.

4. Dataset Documentation

The dataset documentation component of the archive may conveniently be organized in the following sections:

- Structure and access
- Data Quality
- Statistical Summaries
- Data Processing History, Procedures, Programs
- Hard Copy
- Other

These are the titles of the following sections of text.

4.1. Structure and Access

The structure and access section is organized "top down" into subsections addressing greater and greater levels of detail.

Contents. The contents section is a list of all the archived datasets using convenient mnemonic names and brief (1-3 sentence) descriptions of the contents. The names should be used consistently throughout the documentation and need not be as succinct as SAS dataset names. Each dataset's entry should also define the "observation/record unit", e.g., one record represents measurements made on one subject during one calendar month.

Relationships. This section describes relationships, if any, among the various dataset. An archive may contain some datasets which are subsets of others, for example. Or there may be a hierarchical relationship between two datasets. For example, one may be a flat file ("standard data matrix") containing static data (name, SSN, date of birth) with one record per subject while another contains longitudinal medical measurements with little or no static data, and with one record per subject per month.

Another typical relationship is complementary subsets. For example, one dataset may contain data from males, another data from females. Or, data from different years may be stored in different datasets.

Relationships are often best described by a formal schema or informally by charts and tables, both augmented by text understandable by an outsider.

Access Information. This section describes to a potential user the hardware and software information needed to access (read) the datasets, including hardware type (e.g. IBM 3330 disk), volume serial numbers (software and hardware serial number), DCB information such as RECFM (record format using a specific operating system code), blocksize, logical record length, etc. It is useful to completely specify JCL (or similar) information needed to access the data, including the sorts of information normally retrieved from the dataset label (recording density, technique; label serial number, etc.). In subsequent access, the dataset label may be separated from the dataset (disk volumes) or the software may process labels differently.

Record Level Information (Within Dataset). One section is included for each record type within each dataset, specifying the actual record format, if relevant. Names of variables fields used in these sections should be consistent with documentation in other sections (project documentation of data
collection instruments; statistical summaries described below).

For SAS datasets, PROC CONTENTS printout provides a useful table to which text, understandable to an outsider, is added, especially if all variables have been provided with formats and labels.

This section should specify key variables if the dataset uses explicit keys. The dataset structure (organization) should be described if it is not a flat file ("standard data matrix). Any relationships among records within a dataset should be described as, for example, multiple records for the same subject or experimental unit.

Field/Variable Level Information. One section of field/variable specifications is provided for each record type in each dataset. Variables may be considered in two types: original data variables and "derived" ("transformed," "computed") variables.

Original data variables need only be listed together with a reference to the documentation of the source data instrument in the project documentation section.

Each derived variable should be described in detail. The description should list all the original data variables and derived variables used in the computation of the derived variable. The algorithm for computing the derived variable should be described in sufficient detail to permit re-programming.

Some variables may require data processing as, for example, SAS variables which require multiple DATA and/or PROC steps to compute. The documentation of such variables may be organized into two parts, one describing the data processing steps and one describing the specific computations producing each variable. Data processing description should be a combination of text and flow charts; the actual code used to perform the processing may be appended.

Other relevant information. One should carefully consider what other types of access information are needed for an outsider to readily access/process the data and write additional subsections providing this information. For example, are passwords needed to access the data?

4.2. Data Quality

During the course of a project, the managers and analysts usually develop both objective information and subjective opinions about the quality of various components of the data. Information on data quality is extremely useful to potential users.

The objectives of the Data Quality section are to:

• Describe procedures used to assure and/or assess data quality
• Provide potential users with objective evaluations of data quality
• Document extreme and questionable data values
• Document known measurement or processing problems

This section may be organized into several subsections, described in the following text.

Procedures to Assess and Control Data Quality. The amount of resources invested in assessing and controlling data quality varies widely from project to project. This section should contain descriptions of the programs employed in the project being archived. Some types of programs often used are:

• Data validation studies
• Inter-observer variability studies
• Data processing quality control program
• Post hoc studies of data quality

The purpose of this subsection is to describe the programs, not the results. These programs, and their results, are usually carefully described in reports written specifically for that purpose. Such reports should be summarized in this subsection, referenced, and included in the archive.

Objective Evaluations of Data Quality. These subsections contain statistical summaries of data quality, as estimated by the programs described in the previous subsections. All available information should be summarized; the summaries should be organized by (a) type of program generating data quality estimates and (b) data type. If the complete texts of separate data quality reports are contained in the archive, the summaries contained in this subsection can be at a very general level and the reader can be referred to the reports for details.

Extreme and Questionable Data Values. Every project encounters extreme and questionable data values, which can usually be placed in three categories:

• known errors
• suspicious data values (correctness unknown)
• extreme/questionable values known to be correct.

A careful compilation of extreme/suspicious/questionable data values should be maintained throughout the life of the project. (Values which are found to be in error and are subsequently corrected are deleted from the list.) This list of "questionable, etc." values should be provided for potential future users who, lacking this information when performing their own edits of the data, might not detect known (to project personnel) errors and who might delete extreme values known (to project personnel) to be correct. Advance
plannining is needed to capture this information for archival purposes at the time the information is generated, usually as a part of standard operations and analyses.

Measurement and Processing Problems. Most long term studies encounter operational problems of various sorts which affect the quality of data. Instruments may be found to have seasonal trends, or biases. Political events during a sample survey may seem to affect peoples' responses to relatively non-political questions. Certain instruments may be found to have precision than others. This subsection provides a place to summarize objective information of this sort. Note that this section may simply highlight or interpret events recorded in the project history, described earlier.

4.3. Statistical Summaries of Archived Data

The objective of this section is to provide a potential user with a descriptive statistical study of each type of data from the project, that is, to describe the distribution of each important variable. Such a study is usually generated as the first step of a statistical analysis, so that if archive planning is done in advance it is only necessary to "capture" the results at the proper time. The statistics included in a descriptive study are those generated by PROC UNIVARIATE for continuous variables or PROCs FREQ and CHART for discrete variables.

In addition to providing the potential user with the capability of assessing the relevance of the data to his/her particular objectives, a descriptive study also provides a series of "check" statistics. One of the first tasks of a future user should be to repeat the computation of descriptive statistics for variables of greatest interest. These statistics can be compared with the statistics in the archive to determine that the appropriate datasets are being used, that the data are the same as those which were archived, etc.

4.4. Documentation of Data Processing History, Procedures, and Programs

The first issue to consider for this section is whether the section should be created. This is a very serious issue which must be considered in view of the great expense, the difficulty of developing the material, and the possibly limited utility of the material.

It may be important to understand one of our assumptions about the research data management process. Assumption: The datasets produced by a project are data products, produced by skilled craftsmen/women who are justified in taking pride in the quality of their work.

If the assumption is correct there may be no need to include documentation of data processing history, procedures, and programs in the archive. One does not ask a master clockmaker to write down all the details of how he built the clock. If the clockmaker has a good reputation, and the clock runs well and keeps good time then we accept the product at face value. The same attitude can be applied to data processing.

In spite of the cost, etc., of developing this documentation for the archive, it may be necessary or desirable. Some of the issues one should consider in making the decision are:

- Is an understanding of data processing history, procedures, and programs important to a full understanding of the data? (For example, in a longitudinal study of lung function/capacity in children, the measurement process changed from manual measurements of an instrument-produced graph to computer-automated measurement. Although this is documented in both the scientific documentation and documentation of variables, the change in data processing (from paper forms to computer generated tape records) may be sufficiently important to justify documentation here).

- Are there legal or regulatory requirements to document processing? -- Pharmaceutical firms are required to retain and document "everything" pertaining to licensing of prescription drugs.

-- If the study's results have important implications for commercial products or governmental actions, an "audit" may be ordered and one may be required to explain discrepancies between original data values and data values in analysis files.

If documentation of processing, procedures, and programs is to be included in the archive, one can obtain guidance for this task from any of several books describing documentation of programs, systems, etc. One should also note that such documentation must be planned in advance and produced during processing phases of the project.

4.5. Hard Copy of Datasets

In some situations it is appropriate to include a hard copy of the data, such as a printout listing, microfilm, etc., in an archive. Some of the issues to consider in deciding whether to include hard copy are:

- Will the usefulness of the hardcopy justify the cost of obtaining and storing it?

- Is there a legal requirement to store a hard copy?
Assuming the datasets remain in machine-readable condition, how difficult would it be to produce hard copy if needed?

Should hard copy be considered as the ultimate backup? It would be possible, though expensive, to key data from a hard copy.

Some institutions keep a hard copy and argue that this copy is analogous to a chemist’s lab notebook, which is both an ultimate data source and archive, with legal and ethical implications.

Although one should give this issue serious consideration, in many instances archiving a hard copy is not justified.

5. Conclusions

The suggestions and proposals contained in this paper are not, alas, the authors’ standard operating procedure … yet.

Although we have suffered through the demise of several projects and have a variety of experiences with archiving we have yet to see any project which was archived to these “exacting requirements.”

On the other hand, we have seen scientifically valuable data wasted—at great cost to taxpayers and the advancement of science—due to inadequate archiving.

We believe these guidelines, produced in response to a very kind invitation by Sally Carson of the Rand Corporation, will be very helpful to us in present and future projects. In particular, we believe the paper will help us convince project managers of the need to allocate resources throughout the life of the project to assure that the project will leave an appropriate “estate”.

6. Reference


7. Credits

Photograph from Encyclopedia of World Art, Volume IV, p. 589.

Research for this paper was supported, in part, by Public Health Service Grant Number 5 D04 AM 01816, “Sampling and Managing Health Information,” and National Heart, Lung and Blood Institute Grant Number HL019171, “Pulmonary SCOR: Lung Injury by Childhood Infection.”