DOCUMENT RESUME

ED 243 904 TM 830 842

TITLE Archiving Methodology. Volume I: Project Officer's Guide.
INSTITUTION Leinwand (C.M.) Associates, Inc., Newton, Mass.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 30 Jul 79
NOTE 50p.; For related documents, see TM 830 843-845.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Archives; *Databases; *Data Collection; *Delivery Systems; Diffusion (Communication); Documentation; *Federal Government; Guidelines; Information Dissemination; Information Utilization; *Models
IDENTIFIERS *Secondary Analysis

ABSTRACT

Recently, to encourage secondary analysis, the federal government has begun to arrange for public policy data to be documented, archived and released to the public. The purpose of this document is to provide government project officers with guidelines for archiving government-sponsored data files. The guidelines represent a model for systematically transferring data from the original data collection contractors to the public domain in a form amenable to secondary analysis. The model has four stages: (1) establishing requirements, policies, and procedures to facilitate data archiving; (2) deciding whether a specific data set will be archived; (3) creating an archived data set; and (4) transferring the data to a consortium which will maintain and disseminate them. (PN)

ARCHIVING METHODOLOGY

VOLUME I: PROJECT OFFICER'S GUIDE

Submitted to

National Institute of Education

by

C.M. Leinwand Associates, Inc.

July 30, 1979

CONTENTS

FIGURES

PROJECT OFFICER'S GUIDE TO DATA ARCHIVING:

I. INTRODUCTION:

Since the Census of 1790, the federal government has been collecting data for public policy purposes. As these data collection activities have expanded, so has the potential for the use of the data themselves. Recently, to encourage secondary analysis, the federal government has begun to arrange for these data to be documented, archived and released to the public. This support has been prompted by the recognition that data collection is an expensive proposition and that it is in the public interest to maximize the use of data acquired with federal funds. In Appendix A, we present a detailed rationale for archiving and releasing data for public policy research.

The purpose of this document is to provide government project officers with guidelines for archiving government-sponsored data files. The guidelines represent a model for systematically transferring data from the original data collection contractors to the public domain in a form amenable to secondary analysis.

The model has four stages: 1) establishing requirements, policies, and procedures to facilitate data archiving; 2) deciding whether a specific data set will be archived; 3) creating an archived data set; and 4) transferring the data to a consortium which will maintain and disseminate them.

Stage one of data archiving takes place entirely at the federal level. It entails establishing policies and procedures for data archiving, including requirements for data archiving in requests for proposals and contracts, and establishing ownership of the data. Stage two occurs when a new data collection project is initiated. The focus of this stage is deciding

whether to archive, what to archive, and how much effort to devote to archiving the data selected. After these decisions have been made, Stage Three creating the archived data set and its associated documentation, begins. Typically, the data are archived through the interaction of two organizations: the organization responsible for the initial data collection and analysis involved in a study, and the data archiving organization responsible for the preparation of the final user-level documentation. The dissemination activities of Stage Four involve storing and maintaining the data in a manner that maximizes their use, publicizing the availability of the data, and providing assistance to interested researchers.

I. STAGE ONE: ESTABLISHING POLICIES

The primary purposes of Stage One are to establish procedures for data archiving and to inform all contractors, present and potential, that they should be prepared to comply with requirements established to facilitate data archiving.

To meet these objectives, all requests for proposals (RFPs) and contracts should specify that machine-readable data generated by government support are to be considered in the public domain. This provision should clearly establish federal ownership of the data and stipulate that primary data files deemed valuable for secondary analysis are to be placed in a repository designated by the federal agency within a reasonable period of time. All data contractors should be prepared to submit data tapes and documentation with their final reports. RFPs and contracts should also state that releasing and disseminating data is the responsibility of the sponsoring federal agency, not that of the primary investigators or other data collection contractors.

Generally, current statutes and judicial interpretation support the public nature of data files collected at the government's expense. But, there are also laws which protect the privacy of individuals and organizations supplying the data. Both the RFP and the contract must make data collectors aware of these laws, before the collection process begins. In addition, it is necessary to develop an agreement which will satisfy archiving requirements and, at the same time, comply with the Freedom of Information Act, the Privacy Act, and other related statutes.

Making such considerations prior to entering into a contract does not necessarily mean that the data a particular project produces will be archived. In fact, since the whole concept of data archiving is relatively new, it

is likely that, in the near future, only a small percentage of the data collected under government contracts will be archived. Deciding which data to archive is the focus of Stage Two.

II. STAGE TWO: DECIDING WHAT TO ARCHIVE

A. BACKGROUND

Since a multitude of research contracts are awarded annually by local, state, and federal agencies, clearly, there is no dearth of studies that could be considered for a data archive. As attractive as this may appear to the archivist and to those interested in maximizing the potential of existing data, the project officer is left with the task of evaluating the worth of the data for archival purposes. Careful evaluation of a study and its data before electing to have them archived is essential to the creation of a useful archive. However, deciding which ones should be archived is not an easy task. In part, this is due to the large number of studies funded (each of which usually generates multiple files) and partly due to the wide range of content encompassed by these studies. Decisions about a study's worth require not only an understanding of the substantive area and its methodological characteristics but also demand that the decision maker have a strong sense of whether or not the study findings will be of interest to others in the field in question. For example, the Division of Policy Research and Analysis of the National Science Foundation (NSF) has funded research in such disparate areas as energy, innovation processes, the socioeconomic effects of science and technology, and public policy related to science and technology. The National Institute of Education (NIE) has sponsored research in such diverse areas as compensatory education, school finance, career education, bilingual education, special education, education for the handicapped, and continuing education. Collectively, these research projects have generated thousands of data files. If either NSF or NIE were to archive all the data

files produced by its programs, it would be necessary to create a separate division for the sole purpose of documenting and archiving these data projects.

It is important to emphasize that not all data are equally valuable in secondary analysis. To determine which data are most valuable, all data bases generated from projects, studies, and awards should be evaluated using specific criteria. These preliminary evaluations will indicate whether a data base is suitable for, and worth, archiving and how much effort should be devoted to archiving it. We have developed a specific set of indicators which can help to ascertain if the time and effort required to archive a particular data set is a wise investment. An excellent method of determining the value of an investment is to consider its "pay off" value. The "pay off" in this case, refers to how importantly the study and its results contribute to science and public policy.

B. DESCRIPTION OF EVALUATION CRITERIA

Data archiving costs relatively little in contrast to the expenses incurred in initial data collection and analysis activities. Nevertheless, considerable time and effort are expended to develop and disseminate a data file. To determine whether a data set should be archived, these expenditures are compared with the potential value or "yield" of the file to researchers. Indicators of "yield" are typically subjective, but can be measured.

In this section, we present and describe evaluation criteria used to help measure yield. An evaluation form has been designed to help project officers identify high-yield data sets for archiving. The criteria listed on this form are not hard and fast rules, but rather guidelines to assist project officers in making decisions about what to archive. The form itself appears as Figure 1. A more detailed description of the criteria follows Figure 1.

In summary, there are five criteria for evaluating data yield : scope of data archiving, study design, topical area, public interest, and quality of data. Each criterion is rated on a scale with a low score of "1" and a high score of "5." Although the scale is not weighted, a score of "1" on data quality would cause the data set to be rejected, regardless of scores on other criteria. This is because faulty data creates serious problems in secondary analysis and is a poor basis for public policy research. A total score of 20-25 indicates that the data set is definitely worth archiving, 15-19 points indicate that the data is possibly worth archiving. Data sets scoring less than 15 points should not be archived.

To use the evaluation criteria and form correctly, it is important to understand fully the issues related to each criterion. In addition, these criteria should not be considered fully independent measures. Hopefully,

FIGURE 1

EVALUATION FORM - FOR ARCHIVING DECISIONS

CRITIERIA FOR DATA YIELD	SCORE (lowest)				(highest)
1. Scope of Data	1	2	3	4	5
Guideline: If a data set has a national probability sample, the full five points should be assigned. One point should be assigned to studies employing convenience samples of a narrow scope, such as individual cities, school districts, or families.					
2. Study Design	1	2	3	4	5
Guideline: If a data set is part of a long-term longitudinal sample, the full five points should be assigned. All cross-sectional (one-time) studies should receive a score of one.					
3. Topical Area	1	2	3	4	5
Guideline: If a data set is of broad substantive focus and will be potentially useful in solving either scientific or policy questions, a full five points should be assigned. A score of one should be assigned to studies of very narrow topical interest. A study of stratospheric conditions in Waltham, Massachusetts, from 1971 to 1978 would be considered of low topical interest and would be assigned a score of one. Intermediate scores should be assigned to the degree that a data set can be used to answer current research and policy questions.					
4. Public Interest	1	2	3	4	5
Guideline: If there have been unsolicited requests for a given data set by universities, policy makers, or scientists and informal or formal gauges of public interest are high, the full five points should be assigned. A score of one should be assigned to studies in which no requests have been made and no interest has been shown when public interest has been assessed.					

CRITERIA FOR DATA YIELD

SCORE

lowest highest

5. Quality of Data 1 2 3 4 5

Guideline: Data evaluators must be very cautious when judging this criterion. If data quality is low, this criterion will override other criteria and cause the data set to be rejected for archiving purposes. A score of one means that the data set can not be considered for secondary analysis purposes, regardless of its other merits.

A score of five should be assigned to studies which employ reliability analyses, consistency checks, coding checks, and have a good data collection plan. Intermediate scores should be assigned based on the quality of consistency checks, data formatting, and data architecture.

TOTAL SCORE 1-14 15-19 20-25

The sum of the criteria can be used as an aggregate measure whether data should be archived or not. The scale ranges from a low score of 5 to a high score of 25. A data set with a cumulative score ranging from 20 to 25 should definitely be archived. A score of 15 to 20 should be the basis for seriously considering archiving. A score of below 15 should be the basis for deciding definitely not to archive.

Note: It should be noted that if the data set is scored "1" on data quality, it should definitely not be archived, regardless of its cumulative score.

each will be interrelated in file which appears to have high data-yield potential

1. Scope of Data Set

The scope of the data set is typically the first evaluation criteria considered. It refers to the research population on which the sample is based. Research populations can include all Michigan State University freshman, all urban riots which occurred between 1968 and 1972 in the United States, all court accessions for juvenile delinquents in the State of Washington, or the adult population of Czechoslovkia. The scope is the largest of individuals to which statistical inferences can be drawn from the sample.

Scope is not measured in terms of size, but rather, in terms of representativeness. A sample of 1,500 respondents representative of the nation's employed mothers would be ranked high on scope, while a sample of 10,000 respondents from the city of Moose Jaw, Saskatchewan, would be rated low on scope.

Data sets having a wide scope are often very valuable for archiving purposes. Examples of such data sets are census tract projects, national election studies, general social surveys, and nationwide impact studies of substantive areas such as education, law enforcement, and the utilization of science and technology. Because these studies are based on national populations and not on smaller units, broad statistical inferences can be made. For instance, a narrow scope project, such as a study of women in the Signal Corpsof the U.S. Army in Europe, has a very low scope rating. Statistical generalizations can not be extended beyond this limited group. In contrast, a sample of 1,500 respondents representative of the nation's voting population would have a high scope rating, because its results could be generalized to describe all adult voters in the United States.

As the cost of collecting primary data for nationally representative studies has increased, the value of wide-scope samples has also increased. For example, the Department of Labor and the National Institute of Mental Health spent over $1 million to conduct their 1977 Quality of Employment Survey. Fewer and fewer research organizations will be able to conduct studies of national scope, those which do so will tend to focus on broader issues, problems and needs as costs continue to increase.

2. Study Design

A second criterion of data yield is study design. Two types of survey designs are most often used in social research: cross-sectional and longitudinal. Cross-sectional studies are designed to look at phenomena from one point in time. For example, Title I compensatory education services may be assessed during 1980 or the marital happiness of dual career couples may have been assessed in a 1973 study. In longitudinal studies, phenomena are continually or periodically observed over a length of time. For example, the Census Bureau might look at general trends in fertility rates from 1966 to 1975 or the Bureau of Labor Statistics might look at labor force participation rates for females from 1966 to 1979. Longitudinal studies which use the same respondents and measure the same variables over a period of time are called panel studies. For instance, the National Opinion Research Center might have assessed the same respondents' political attitudes in 1960, 1964, 1968, and 1972.

Longitudinal designs involving panels or repeated observations on other phenomena are valuable because they collect data for the analysis of change. The 1977 Quality of Employment Survey which examined workers' perceptions of labor standards, problems, job satisfaction, job stress, and the meaning of work, re-interviewed a panel of respondents who had also been interviewed in 1973.

The panel survey was particularly important because it allowed researchers to view social indicators over time.

Like large scale studies, longitudinal studies, especially those involving panels, are expensive; few can be undertaken without government support. Because of their ability to measure social change and the expenses associated with their conduct, studies employing a longitudinal design should be seriously considered for release to the public.

Although cross-sectional studies are less powerful than the longitudinal studies in assessing change, well-executed cross-sectional studies may also be worth archiving, especially when they concern new phenomena and trends. For example, the 1977 Quality of Employment Survey also contained a cross-section of new respondents whose responses were compared with those of the panel. The cross-sectional portion of the study investigated many new topics, among them, employment of the respondent's spouse and the impact of both spouses' working on family life. Given the changes in sex roles, the emergence of the dual career family, and women's increased labor participation, the cross-sectional survey contains data rich in potential for secondary analysis. Another example of a valuable cross-sectional study is the Safe School Study sponsored by the National Institute of Education. This study provides national estimates on an issue that had not previously been examined in depth: the extent of violence in our nation's schools.

3. Topical Area

Studies with a wide scope and longitudinal design are not always of value to secondary analysts. An important additional consideration in archiving is the topical or substantive areas a study covers.

While the criteria of scale and survey design lend themselves to objectivity and quantification, assessing what might be of interest to secondary analysts entails making judgements that are generally more subjective. In the field of sociology, in the 1940s, rural sociology was near the top of specialty areas. Studies of fertilizer dissemination and hog-corn correlations fascinated sociologists and statisticians of the period. But in the urban America of the 1970s, rural sociology is not a popular specialty area. Had machine-readable data on new corn hybrids of the '40s been saved, they might have little relevance for today's predominately urban society.

Efforts designed to collect data to solve major social problems or answer important scientific or policy questions should be reviewed in light of their short- and long-term interest to researchers.

4. Public Interest

Like the topical area criterion, public interest in specific data files may be difficult to gauge. Clearly, a data file for which there is public demand, such as census data, should be seriously considered for archiving, even if it receives low ratings on other criteria. One obvious indicator of public interest is frequent, unsolicited requests for data. If an agency receives many requests for compensatory education, school finance, or school violence data, the potential yield is high.

One way to gauge public interest is to formally or informally survey researcher or university social science data centers which disseminate data.

A more formal approach would be to publicize the potential availability of a data file in professional journals and request inquiries from interested parties. In addition, a number of professional organizations of data users have been formed in recent years which could be queried to provide an indication of interest. Appendix B lists a number of these organizations and persons to contact.

5. Quality of Data

Data quality refers to the care taken to collect, code, and format data. High-quality data is consistent with survey design, correctly coded, and properly formatted. The consistency of actual data with the survey design is a particularly important consideration. In a panel study designed to observe changes in gender roles over time, minimizing such factors as loss of respondents would increase the consistency of, and confidence in, the data. If 60% of the original respondents were not re-interviewed, the survey's representativeness would be lost. Coding and formatting outcomes are also indicators of data quality. For example, if 20% of the responses to an item asking if the respondent's spouse was employed were not coded properly, this would raise serious doubts about the quality of the study. Although a study could be longitudinal, of wide scope and great topical significance, and in public demand, its data yield potential could be undermined due to coding mistakes made during the data collection process. In such a case, the data would be comparatively unimportant for secondary analysis because of their limited reliability.

It is, of course, difficult to make judgements of data quality prior to data collection. One method of assessing data quality prior to data collection is to assess the data collection contractor's experience. In cases where experience is unknown or limited, it is important to emphasize the importance

of data documentation and quality controls to potential data collection contractors. In some cases, low-quality data results from factors beyond the control of the data contractor. In the panel study given as an example earlier, high unemployment, divorce rates, or other external events might have caused attrition of respondents.

6. Summary of Data-Yield Criteria

The most important criteria for evaluating data sets are the scope of data collection, survey design, topical significance, public interest, and data quality. The evaluation sheet which has been provided gives the Project Officer a summary of data yield criteria and can be used for any data set being considered for archiving. Any of the criterion can be weighted, with the possible exception of data quality, depending upon the specific goals of the agency and the specific function the archive will serve.

C. ARCHIVED DATA SETS: CONTENTS AND LEVELS OF SUPPORT

The decision to archive a single data set or collection of data sets demands that a second decision be made: how much effort to devote to archiving. The work involved in archiving a data file or a collection of files may range from the creation of simple documentation to a more complex undertaking, consisting of reformatting data files and writing completely new documents to describe the archive. To present a full picture, we will discuss the components needed to create the best data archive possible, describe the alternatives and options available within each component, and provide the rationale for choosing each alternative. Using this information in conjunction with data-yield scores, project officers can make decisions about the type of archive to be created and the components to be incorporated in it.

The highest level of archiving support would contain all the components described below. The lowest would entail releasing the data as they are received from the data collection contractor and providing copies of any documentation available in the contractor's reports.

Our discussion of the components is organized to reflect the two major tasks involved in data archiving: structuring the files and developing the accompanying documentation. Each of these two areas are discussed in the next section.

1. File Structure

The preparation of data files for archiving and release to the public focuses on two major concerns:

a. Ease of Use

b. Flexibility

There are two primary strategies for dealing with these concerns:

c. Organization/structure

d. Standardization/recoding

a. Concern: Ease of Use

Data files must be prepared in a manner that expedites their use by an analyst involved in secondary research. During the initial data collection and analysis phases, the data files were prepared to fulfill the specific needs of the research project. These needs also dictated the decisions made on file organization or coding. Nonetheless, the files--regardless of how they were organized or coded--had to be analyzed to fulfill the terms of the contract. However, secondary analysts will be working under a significantly different set of constraints, usually to answer a significantly different research question. If the data files are difficult to access or analyze, these analysts may have to redisgn or table their research.

b. Concern: Flexibility

The data archive contractor faces an additional issue, that is, how to maximize data's utility for further analytic purposes. This issue becomes especially important when modifications or recodes to the data are planned as part of the archiving process. The case of recoding of missing values helps illustrate the implications of the issue of flexibility.

In NIE's Safe School Study data, missing value codes could have one of six different values, depending on the reason the data were missing. One value indicated "don't know," another indicated "refusal to answer the question," and still another indicated a "legitimate skip." (The respondent should not have answered the question and was routed around it in the survey.) Other values indicated other problems in the data collection. The existence of these six missing values caused difficulties when early versions of the Statistical Package for the Social Sciences (SPSS) were used, since these versions only allowed three independent missing values. Each time an SPSS analysis run was made, the six missing value codes for each variable had to be recoded into three at most. From most analysts' viewpoint, the file would be easier to use if only one or, at most, three missing values existed for each variable. Unfortunately, some analysis might not have been possible unless the six missing value codes were differentiated; therefore, if these variables were recoded on the archive files, a potential analysis opportunity may have been foreclosed.*

c. Strategy: Organization/Structure

In any large-scale data collection project, decisions about structuring and merging data files are based on the specific research issues to be addressed by the analysis. Consequently, each data file is organized in a manner consistent with those needs. For example, a classroom observation study in which an observer completed a data sheet for each ten-minute time period within a school day could be organized in two alternative structures: in the first, the student is the unit of analysis; in the second, the activity is the unit of analysis.

* SPSS versions 7.0 and above allow a range of missing values to be used, making this particular point moot. At the time the Safe School Study was conducted, it was a very real issue.

In the first structure alternative using the student as the basic unit of analysis, one long record would be created for each student observed, and all of the student's activities would be contained in a separate variable. Assuming that there were 25 ten-minute periods in the school day and that the observers marked an activity code for each period, the record would contain 26 variables (a student identifier and 25 period variables). In the second alternative structure, an activity would be the basic unit of analysis. Each ten-minute period for each student would be recorded as a separate record in the file having three variables (the student identifier, a code for the period recorded, and an activity code). Although the first structure is considerably more compact (i.e., it occupies less computer space), the second is more suitable to answer such questions as, "What is the most popular activity?" or "On the average, how many ten-minute periods are spent reading?"

Structuring longitudinal data poses a similar problem. Should the data collected in each longitudinal period be treated as a separate data record? Or, should the data for all periods for each person be merged into one larger record?

The final structure-related issue to be addressed deals with the appearance of data at different levels of analysis on the same data file. This issue is sometimes called the "hierarchical vs. rectangular argument." In a rectangular file, each observation or record on the file is at the same level of analysis and contains data on the same questions. A rectangular file can be envisioned as a piece of graph paper on which each horizontal line represents an observation and each column represents a different question. When the data is viewed, it is rectangular in shape. Although different data items may not appear on some lines because no response was given, potentially, each line could contain data for each column.

In hierarchical files, each observation or record is not necessarily at the same level of analysis, nor does it necessarily represent identical data items. In addition, each record's length can be different, depending on the data it might contain.

The National Crime Survey (NCS) data sets distributed by the Law Enforcement Assistance Administration is a prime example of hierarchical file organization. In the NCS data file, records are located at four different levels of observation.

- Community records contain information about the community in which the survey was taken.
- Household records contain information about the household being interviewed.
- Individual records reference personal questions pertaining to each individual within a household.
- Incident records provide detailed information on each crime that occurred.

Each of these record types is of a different length, and each contains different information. The file is organized so that community records are followed by records of the first household in the community, then records of the first individual within the first household, and, finally, the first individual's incident records. The record for the second individual in the first household in the first community appears next, followed by the second individual's incident records. After the records for all individuals in the first household are completed, records begin for the second household. When all households for that community are reported, another community record begins on tape.

A file in which each data record contains the same type of information--such as a student survey file sorted by student within school--is <u>not</u> a hierarchical file. Since each data record contains the same type of information, this is simply a rectangular file in a particular sort sequence.

For archiving purposes, the major consideration is, How can the files be organized in a manner which preserves maximum flexibility for the secondary analyst? We will address the related issues in turn.

Hierarchical versus Rectangular Structure. Hierarchical and rectangular file structuring each have merits. In choosing one or the other, tradeoffs are necessarily made. For data collected at varying levels, a hierarchical format is the most compact and flexible for data storage and analysis. For instance, the solutions of analytic problems requiring the use of district- and individual-level data are facilitated through a hierarchical data structure. The rectangular format, however, is simpler and easier for an analyst not involved in an original study to use and understand. In addition, the most popular statistical analysis package is the Statistical Package for the Social Sciences (SPSS), which can process only rectangular files. To analyze a hierarchical data set with SPSS, a programmer would have to write a special program to manipulate any hierarchical data within the file. The development of such a program would be costly and time-consuming. The OSIRIS statistical package, also widely used, is similarily unable to handle hierarchical files directly. The Statistical Analysis System (SAS), whose availablility is much more limited, can handle hierarchical files but only in a rather obscure manner. The SAS manual does not address the issue of hierarchical files or give good examples of its use with such files. The alternatives associated with this component are listed below, from the highest level of effort and usefulness to the lowest.

- Structure the files hierarchically; develop and provide programs which would allow analyst to manipulate the data using popular statistical programs.
- Structure the file rectangularly
- Do not restructure; use the files as received from the contractor..

Merging of Files. Frequently, data collected on the same level of analysis but through different instrument or techniques are originally organized as separate files. Within many large projects, multiple files contain the same level of data. For example, district-level data may have been collected with three instruments. It is necessary to decide whether these separate files should be merged into one as part of the archiving process. The answer to this question is based on analysis of two factors: ease of merging and ambiguity of documentation.

Ease of merging refers to the level of difficulty an analyst would encounter in attempting to merge data sets. For example, an initial analysis of school district-level data from a national survey indicated that a consistent district coding scheme was used for all four files. It was assumed that merging these files would be a relatively simple task. However, the student identification numbers in the student data files were changed from year to year to reflect changes in family structure or to identify students who left the school district and later returned. Therefore, it was actually quite difficult to merge the files.

Sometimes, merging may preclude the use of a file for answering a specific research question. For this reason, we recommend that merges be done by secondary analysts working with the archived files, if linkage variables are clear and merging procedures straightforward. In cases where file merging is complicated by complex linkage variables or similar problems, we recommend that merging be performed as part of the archiving process.

The second factor in deciding whether to merge files is ambiguity of documentation. Certain data collection instruments contain the layouts for the resultant data records as part of the instruments themselves. Thus,

an analyst reviewing the data collection instrument can obtain the location of each variable within the data file directly. If the file were merged with others, the location information within the survey form would no longer reflect actual data records. Utilization of the data set would then require another intermediate step to translate the survey question number to an actual location within the file. Other instruments, however, include no locational information and, consequently, possibly confusing factors do not exist. In either case, an analyst would have to use an intermediate codebook to discover an IDS item's location within the data file.

File Structure: Level. Level of analysis is a structuring issue which can usually be easily resolved. Decisions about level of analysis have only limited importance in the archiving process, since it is quite easy to transform data records structured at one level to another level. For example, in the classroom observation example discussed above, if the student was chosen as the unit of analysis, a very simple program could be written to transform the file into an activity-level file. The program could be written in the familiar and readily usable SPSS format to further ease this problem. Since no flexibility is lost with either choice, files may, in most cases, be archived at the same level of analysis as they are received.

d. Strategy: Standardization/Recoding

Standardization and recoding are undertaken to resolve problems created by dissimilar treatment of missing values, the use of alphabetic codes, inconsistent coding of linkage variables, the use of similar questions in different instruments, and codes for responses to open-ended questions.

In many studies, data is collected by different contractors, each attempting to undertake independent substudies. Often, one result of this "joint effort"

is that no consistent coding scheme is used to prepare the collected data in machine-readable form. Consequently, differing missing value codes may appear in each of the study's data sets.

This, however, is not the only problem related to standardization and recoding: other potential problems, intentional and unintentional, can arise. For instance, the use of alphabetic codes (the letters "a" through "z") as data values is an all too common and often problematic practice. In addition, linkage variables, such as state codes, are coded inconsistently. Another concern in standardization is with similar questions asked in different ways in different surveys. Although the original coding scheme does not formally account for these similarities, they can be incorporated in the coding scheme used in archiving. Finally, data bases often include items which were coded from open-ended questions. It is necessary to decide whether to collapse some of the infrequently used codes or to delete certain data items completely. We will discuss each of these five types of recoding activities in the next few pages.

Missing Values. Missing values are also rarely standardized access files. This is especially true if the files were prepared by different contractors. We propose a general approach to missing values: to institute a consistent set of missing values throughout archive files. In addition, we recommend the creation of a set of identical missing values for use with all variables within the files of an archive project.

This approach is not quite as simple as it first appears. It suggests many alternatives, each of which presents its own problem. The initial inclination is to use some negative range of values to represent missing values say, for example, -1, -2, -3, -4. This presents a problem when a variable's value can legally be negative, for instance, in certain test scores or, monetary

amounts. One might choose a different set of values--very large or very small numbers which are unlikely to be legal values, for instance, "-999999." This choice presents a problem for variables whose values otherwise occupy only 1 column: the file size is significantly increased. Another alternative is to use all "9's" in each data field as the missing value or create a sequence of mostly "9's" for multiple missing values. Of course, this approach presents the problem of having different sets of missing values, depending on the width of the data field. For example, single-column data fields might have a missing value of "9," two column variables, "99," etc. Obviously, this approach would also cause problems if "9" were a legal value for a single-column data field.

Given the choices and limitations of each alternative, the best initial approach to coding missing values is using a negative number range, since negative values are not legal for most data items. In no case is a blank to be used as a missing value. Its use can lead to ambiguity in analysis because many systems cannot differentiate blanks from zeros. This causes severe problems when zero is a legal value. To make final determinations, it is necessary to review the missing value coding schemes in use within the data sets and the legal values for each of the questions within the data files.

Use of Alphabetic Values. A troublesome but common practice in survey research is to use alphabetic values ("a" through "z") as responses to questions. For example, a survey question has 11 valid responses. Instead of using two columns for this data item, the values "1-9", "A", "B" are used. "A" represents an answer of "10" and "B," "11." This technique is gemerally used to save keypunching time, and it does reduce keypunching costs slightly. However, it impacts most analysis activities adversely, since most statistical packages cannot handle alphabetic responses easily.

Files which contain alphabetic codes should be recoded so that all legal data values in the data files will be numeric. The one possible exception to this nonalphabetic rule may be state identification codes which use the two-character Post Office codes. Determinations on these state codes should be made on a case-by-case basis.

Common Questions across Studies. Sometimes, the independent substudies of a larger substudy may collect similar data using dissimilar questions and methods. One possibility for facilitating the analysis of these data as a group is to create a common coding scheme for the responses to similar questions. We believe that this possibility offers no direct advantage and presents a number of serious disadvantages. In most cases, these similar items were collected as parts of different data collection efforts and, therefore, they are not absolutely equal. Establishing common coding could obscure their important differences and even convince an analyst that they are, in fact, identical precisely because an identical coding scheme appears in the files. Our recommendation for treating similar data items is to defer recoding to the analyst.

Collapsing Codes. It is not usually advisable to collapse or omit a few codes, especially those post-coded for open-ended data items. Collapsing data values permanently obscures some of the file differences in responses. These differences may seem minor or inconsequential, but it is impossible to forecast what analyses might be conducted with the data in the future. It is possible that what seems inconsequential now will become important to someone in the future. The only instance in which we would advise collapsing codes is when the data shows a difference that is inconsequential or not truly representative.

Regarding the omission of data items, we make a similar recommendation: unless the inclusion of data items would mislead and confuse an analyst or actually reflect incorrect or unreliable data, the items should be placed in the archive. We prefer that the analyst make these decisions; (s)he is in a far better position to determine if the data item is valuable and relevant to a particular analysis.

2. ARCHIVE DOCUMENTATION

Developing documentation is the second major task in data archiving. The documentation that will accompany the archive data files is critical: the level and quality of documentation will have a greater influence on the future use of the data than any other factor. It is, therefore, essential that the documentation developed is complete, accurate, and easy to use.

Archive documentation differs substantially from the documentation usually prepared to accompany data files. Since future users of the data archive will not have the luxury of direct contact with the original data collectors and data analysts, the archive documentation is their only source of information. Therefore, this documentation must anticipate and answer questions that may be asked about the data in the future.

Preparing documentation which meets these requirements demands a variety of skills not obviously associated with data archiving, such as an inter-disciplinary team combining the technical skills of programmers and data analysts with professional writers, editors, and graphic designers. Archive documentation must not only be inclusive and accurate; it must be well written, easy to read and comprehend, and contain visual elements which help its readers to focus on what is most important. Documents developed with these goals

in mind are not only more attractive--most people find them more inviting to use.

In the majority of data collection and analysis projects, data documentation is usually accorded a rather low priority. Generally, data documentation is not a project "deliverable." When it is, no standards exist or are subsequently established for its acceptabil In most cases, this documentation consists of a record layout showing where each data field appears on the data tape; in other cases, only a copy of the data collection instrument with column numbers is provided. Information on collection methodologies, coding techniques, and missing value treatment are usually not reported. Analysts who require this type of information sometimes attempt to piece it together by looking at the data tape or trying to contact someone who has worked with the data. The drawbacks of this limited type of documentation are evident in the following example. The Bureau of Labor Statistics requested that a tape containing information on local area unemployment statistics be reviewed. It was accompanied by a one-page "document" which was supposed to enable researchers to use the tape. (See Figure 3.) The most interesting aspect of this document is that it is presented as user-level documentation for a data file, although

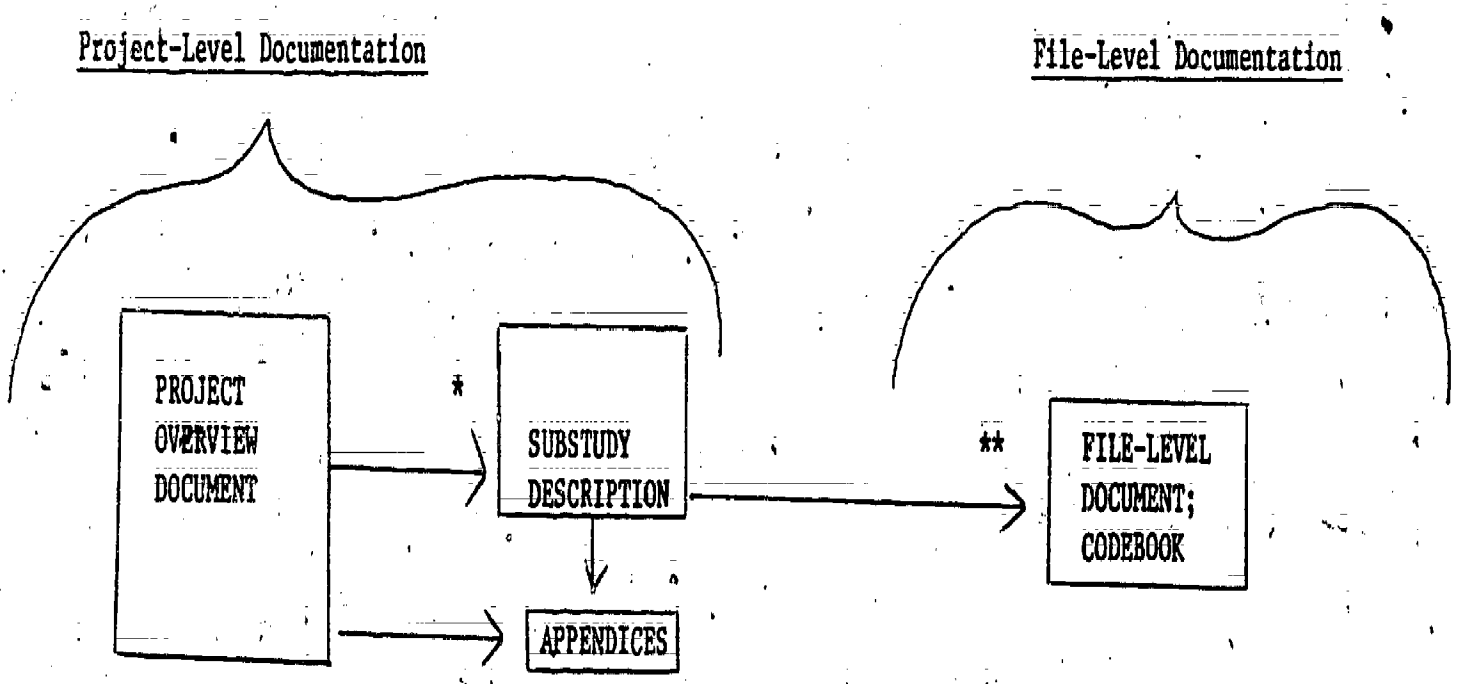
- the record format indicated is incomplete.
- it is not clear whether the state code is an alphabetic or numeric code.
- the data fields in columns 36 through 152 do not indicate what the measure is. Are these values percentages? Is there an implied decimal point in those numbers?

Users would have had to look at a printout of the tape to try to answer these questions. All too often, this type of documentation is considered adequate by its disseminators.

Because a study or project can be conceptualized as a complex whole consisting of multiple parts, the documentation of studies must take into consideration that an archive user needs to understand the study as a whole as well as be familiar with the components of the whole. Documentation should provide a broad overarching description of a research study, and it should also focus in on the components of the study. The first of these descriptions adapts a general perspective toward documentation and the second type of description utilizes a more specific perspective toward documentation. We call the documentation resulting from a general perspective "project-level documentation," and the more specific documentation "file-level documentation."

The next section first presents a figure showing the relationship between the two levels of documentation and secondly, briefly describes the levels. More detail is given in Volumes III and IV of the Archiving Methodology; entitled "Project-Level Documentation Standard" and "File-Level Documentation Standard," respectively. Here, we will discuss the aspects of documentation of most concern to project officers.

RELATIONSHIP BETWEEN PROJECT-LEVEL AND FILE-LEVEL DOCUMENTATION



* A single project is likely to encompass more than one substudy and in such cases, a separate description is developed for each substudy. If there is only one substudy, i.e., one design, then there is no need for a project overview document.

** Substudies usually generate more than one file and each separate file should be documented if its data are judged worthy of archiving.

The project-level document has three major sections: project overview, substudy descriptions, and appendices.

- The project overview summarizes the most important facts about the project and its historical significance.
- The substudy descriptions include sections on substudy purposes, findings, samples, and information on the contents of the data sets in each substudy.
- The appendices include a cross-reference guide, bibliography, and other materials related to the overall project and its substudies.

Project Overview. The first section of the project-level document is an overview of the important facts of the study and is designed to give the reader a thorough grasp of its substance and evolution. These facts are organized thematically:

- Abstract
- Background - historical perspective and significance, issues addressed resulting in the undertaking of the study
- Research topics investigated

The intent of the project overview is to convey clearly and immediately the important elements of the study, how these elements were conceived, and how they articulate with each other. Thus, the overview not only describes the study's historical and theoretical background and the topics it investigated, but also clarifies the overall coherence of these elements, i.e., how they logically flowed together to form the study.

The length of the project overview varies with the complexity and scope of the study. It is recommended that descriptions of complex studies requiring a lengthy overview employ subtitles for organization and emphasis. (See example below.)

Substudy Descriptions. After the project as a whole has been described, a brief overview of the relationship between the substudies comprising the project is presented. The goal of this section is to inform the reader of how these substudies come together within the archive. Before this goal can be achieved, the archivist must first organize the project into substudies. Each substudy is then detailed. Its description consists of the following components:

- o title,
- o background and purpose,
- o study design,
- o sample,
- o statistical analysis,
- o major findings,
- o file descriptions.

The file description is a key section in the substudy description. It gives a brief outline of each data file within a substudy and informs readers of the type, scope, and scale of data in each file. Each file description tells about

- o the type of data in the file, alerting the reader to unexpected data and highlighting important or unusual contents (e.g., "This file is the only known source of nationally-weighted data on violence in schools broken down by location within school");
- o the data collection instrument used to create the file;
- o the number of data items per subject;
- o the number of subjects.

Appendices. A key feature of the appendices is the Cross-Reference Guide. This guide or index enables a researcher to identify information collected through a number of different activities. A researcher interested in analyzing reading instruction practices, for instance, could utilize this guide to identify which questions focussed on this issue and in which data files they are located.

Because of the large number of data items which make up a study, the creation of this guide is both complex and time-consuming. It can be approached in two ways. The simpler approach is to develop a Key Word in Context (KWIC) list which indexes each word within each question. (See Figure 4.) Although this is the most inexpensive method of creating an index, its utility is limited. The terminology utilized in each set of data may differ; certain conceptual ideas may not be directly described in any individual data item description. A conceptual index is more complex and also more valuable. In a conceptual index, key conceptual ideas within the study are identified and then used to index each data item. (See Figure 5.)

Also included in the appendices are bibliographies for the overall project and for the individual substudies. The appendices can also contain excerpts of related studies, reports, and other materials relevant to the study.

File-Level Documentation. Documents comprising the file-level documentation provide detailed descriptions of each data file contained in the archive. These file-level volumes are individual documents pertaining to each data file. They consist of two parts: a narrative description of the file and a computer-generated codebook.

The narrative file descriptions contain information on the goal of the specific data file, the unit of observation, and the data's scope and scale. The codebooks describe each data field, its location, missing values, and coding scheme, and provide specific notes on the field. For high-yield data set, we recommend that a Software system be used to facilitate the creation of codebooks to allow detailed, machine-readable codebooks to be created efficiently.

A number of factors suggest that a machine-readable format be used for codebooks. Many researchers have indicated that it is frustrating to attempt to read "fifth-generation Xerox" codebook copies. By creating codebooks as computer files and including them as part of the archive tapes, each researcher is able to obtain as many first-generation copies of codebooks as required. In addition, because the codebook may be processed by computer, a researcher can use a computer

text editor to reformat the codebook file into a specification file for a particular statistical system, such as SPSS. Thus, machine-readable codebooks facilitate the use of the archived data files through standard statistical analysis systems.

D. DATA ARCHIVING ADVISORY COMMITTEE

Deciding which data sets to archives and what level of archiving effort a data set merits are not easy tasks. This is due to the fact that the creation of an archive demands input from a variety of sources. Moreover the types of users of a data archive must be anticipated at the earliest stages of archive development: their needs, interests, and professional expertise. The best environment in which to make archive decisions is one which represents the multitude of disciplines that converge in the entire data archiving process from its development to utilization.

A committee is the most natural mode in which to obtain the needed input for these decisions. Four possible strategies for creating a Data Archiving Advisory Committee (DAAC) are presented below.

1. Standing In-House Committee

The standing in-house Data Archiving Advisory Committee is a permanent advisory panel within a federal agency which makes archiving decisions. This panel establishes general policies for data archiving, reviews all awards for data collection, and makes initial and final judgements about whether the data is worth archiving. The advantage of a permanent DAAC within the agency is that it assures continuity in experience. A second advantage is that the same people who decide that a given data file is worth archiving also have the authority to oversee the archiving effort and to release the data to public.

An in-house DAAC also has disadvantages. Since a relatively small number of federal agencies currently archive their data and more archiving is done by academic and private agencies, inhouse committees may be currently unequipped to make archiving decisions. A second related disadvantage is that an inhouse DAAC may not assure sufficient input from researchers who will actually be the users of the data. A final disadvantage is other demands on staff time may preclude full participation, especially in agencies that are understaffed.

2. Ad Hoc Inhouse Committee

A second type of inhouse advisory panel would one convened for a specific archive or topical area and draw its members on the basis of their specific expertise. For example, in the case of the National Science Foundation, inhouse representatives of each division would review all projects within specified topical areas. For example, inhouse staff concerned with domestic environmental policy would review all contracts awarded in that area to determine which are appropriate for archiving. All projects related to stratospheric conditions would be reviewed by inhouse experts in that area. These committees would be temporary; they would be established and later disbanded, according to the allocation of research dollars. The advantage of this model is that it will assure expertise will in decision making. But what is gained from the expertise of the temporary committees may be somewhat diminished as a result of the lack of continuity between committees. In addition, subject area specialist may not necessarily be versed in the technical aspects of archiving, such as file architecture, data structuring, and data documentation.

3. Standing Extramural Committee

The third type of panel is a standing committee of outside experts in the topical area, data base management, and public policy. Consisting of a permanent contigent of archiving experts and specialists in the specific research area, this review board could both make archiving decisions in a continuous and expert fashion. Like an inhouse standing committee, the standing extramural committee has the advantages of continuity and accumulation of experience. However, it lacks the authority to oversee data collection and the archiving process and to release data. In addition, its access to contractors and project officers is more limited.

4. Ad Hoc Extramural Committee

The last type of data archiving advisory committees is the ad hoc extramural

committee, composed of distinguished academics and other data policy experts. This type of committee shares advantages and disadvantages of the inhouse ad hoc committee. Because it is formed on the basis of topical expertise, the committee is in a good position to determine what is worth archiving. Because the committee is temporary, it lacks continuity of data archiving experience and expertise. Moreover, the ad hoc extramural committee does not have the opportunity for access to project officers that inhouse temporary panel does. Nevertheless, the extramural panel may have a better understanding of the needs of interested publics.

Each variant of Data Archiving Advisory Committee has its own unique advantages as can be seen in the Rating of their DAAC Variants below (Table 1). The standing inhouse DAAC is strongest in continuity, in accumulated data archiving experience, and in continual contact with project officers. Its weaknesses lie in its possible lack of expertise in data file architecture and topical expertise. The ad hoc inhouse DAAC has strength in topic choice and contact with project officers but lacks continuity and the accumulation of experience. The standing extramural DAAC is strongest in both continuity and expertise. A possible weakeness is that it may lack contact with project officers. The ad hoc extramural DAAC is strongest in expertise and weakest in the area of contuintiy.

In addition, the tools necessary to develop archive data files and machine-readable codebooks and the skills necessary to describe their effective use are more likely to be found in an organization specializing in archiving than in a research firm whose major purpose is research and evaluation. In combination, the specialized functions of data collection/analysis contractors and archive contractors encourage the optimal use of the resources needed to create an archive data set.

From the project officer's viewpoint, a variety of actions are necessary to perform the archive activities of Stage Three:

A. Data Collection and Analysis Contract Modifications;

B. Archive Contract;

C. Review of Archive Deliverables.

A. DATA COLLECTION AND ANALYSIS CONTRACT MODIFICATIONS

To insure that collected study data can be archived in the future by the archive contractor, the data collection and analysis contract should include the "Guide for Data Collection Contractors" Volume II of this Archiving Methodology This guide presents four versions of a "Cross-Reference Information Form" (CRIF) and recommends the contractor use CRIF to identify key information about the project and its data to the archivist. The collection contract should also specify the format of the final data files according to procedures outlined in the "Data Transfer Guide" presented elsewhere in this report and provide for a small amount of funding for consultation between the archiver and the collector. As the Contractor's Guide states, the additional work requirements imposed on the contractor in archiving are few. In most cases, the contractor simply turns over information and data that is normally produced during the analysis activities.

B. ARCHIVE CONTRACT

The contract for data archiving should be awarded at the same time the

III. STAGE THREE: CREATING THE DATA ARCHIVE

In Stage Two, three decisions are made: 1) whether to archive; 2) what to archive; 3) how much effort to devote to archiving the data selected. In Stage Three, the most important decision to be made is, Who will actually create the archive?

For many reasons, data archiving is best performed by an organization separate from the organization which initially collected and analyzed the data. Research data are usually collected and analyzed for a specific reason, within a limited budget. Therefore, the data collection and analysis activities focus on specific research issues being addressed as part of the study. The documentation created by the analysis contractors reflects this orientation and is, in general, limited to the information needed to complete the work required by contract. In addition, if a project is operating on a limited budget, data documentation is always one of the first areas from which funds are diverted, since it is usually considered a "secondary" product of the project.

While many researchers affirm the value of good data documentation, the task of writing clear user documentation is less interesting to them than data collection and analysis. Often, writing documentation occurs at the close of a project when most researchers are ready to start a new project. Under these conditions, the quality of the documentation is likely to suffer.

The mix of skills needed to develop accurate, easy-to-use archive products is quite different from those usually represented on a research team. Writers, editors, and graphic artists form the core of people who can develop good documentation. Although a background in social science research is essential for writers of documentation, such writers must be able to create background descriptions of projects which are sufficiently clear to inform an audience which knows little or nothing about the project. In this case, the writer is employing a different set of skills than the original social science research and analysis team used.

collection and analyses contract is awarded, or shortly thereafter. It is inadvisable to award the archiving contract after a project has been completed. Initiating archiving activities when the project begins allows direct interaction between the archivists and research staff. An early start insures certain advantages: the information about the data is still fresh in the researcher's minds; and the research staff is accessible, usually working together in one place.

The archive contract will describe the level of effort that will be devoted to archiving. This "level of effort" refers to the project's wide range of alternative activities that can be incorporated into the archiving process. In summary, before the archiving contract can be issued, the following questions must be answered.

- o level of file restructuring - Will hierarchical files be rectangularized? Will common levels of data be merged into a single file?
- o Type of data recoding - Will a consistent set of missing values be used throughout all of the data sets? Will consistent coding schemes be used for various questions?
- o Type of documentation - Will new project and file documents or only one of these be written? Will an index of data items be created?

C. REVIEW OF ARCHIVE DELIVERABLES:

Two kinds of general preliminary checking procedures are appropriate for all archived data: data checks and documentation checks. After the data tapes and appropriate documentation have been delivered to the agency, all data files should be subject to preliminary review.

1. Tape Review for All Data

The tape review entails an initial reading of the tapes to check the following items:

- o readability;

- o correct file and record counts
- o sample record checks.

All contractors should provide a printout of the first ten records with the tape and documentation.

It is also recommended that frequency distributions be run on selected variables within the data files. The selection of variables may be random or in accordance with some other rationale. Each variable chosen should be reviewed for appropriate range. For example, in the case of the variable, "religious preference at age 16," legitimate values may range from 100 to 800. A value of 81 or 1200 would therefore be considered "out of range." This is called a range check and involves comparing actual values with legal ranges. Variables with out-of-range values cast doubt on the construction of the variable or the respondent's understanding of the question.

2. Checking the Documentation

Documentation checks depend on the level of support given the data file. Preliminary documentation checks are appropriate for all data files.

3. Preliminary Documentation Checks

The project officer should make sure that the archivist has supplied a record layout (codebook) which shows where each data field is located on the data tape, a copy of data collection instruments, and a description of how the data files relate to the study.

The archivist will provide project-level documentation, including substudy descriptions, file-level documentation, machine-readable codebooks, and programmer's guides. All of these documents should accompany the data tapes.

The first step in reviewing data documentation is to refer to the project-level documentation. This document describes the entire study, including major research questions, historical background, and individual substudies, as well as other broad information about how individual substudies fit into the project.

FIGURE 6

Comparison of DAAC Models

Criteria for Variant

Variant Name:	Continuity	Accumulated Experience	Project Officer Contact	Topical Expertise	Data File Expertise	Contact With Interested Publics	Time Limitations
Standing Inhouse DAAC	+	+	+	-	-	-	-
Ad Hoc Inhouse DAAC	-	-	+	+	-	?	+
Standing Extramural DAAC	+	+	-	+	+	?	-
Ad Hoc Extramural DAAC	-	+	-	+	-	+	+