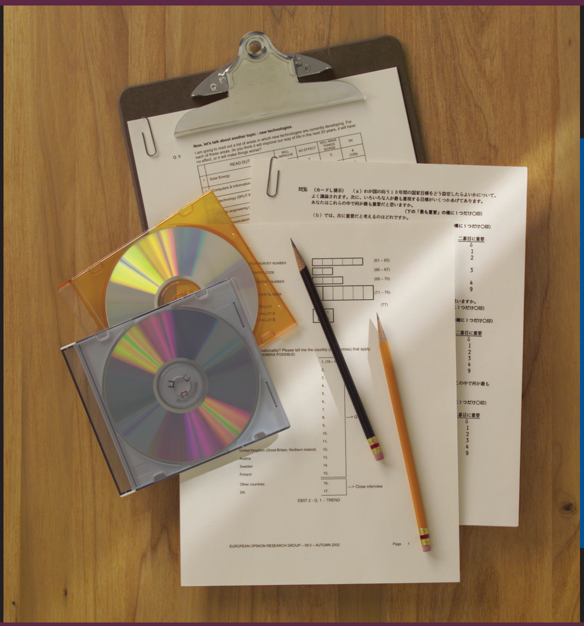
Guide to Social Science Data Preparation and Archiving

Best Practice Throughout the Data Life Cycle



ICPSR
INTER-UNIVERSITY CONSORTIUM FOR
POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH

ICPSR
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

First Edition, 1997. Second Edition, 2000. Third Edition, 2005.

Acknowledgments

First and Second Editions. The *Guide to Social Science Data Preparation and Archiving* was printed in 1997 and 2000 with support from The Robert Wood Johnson Foundation. ICPSR thanks Dr. Richard T. Campbell, former Resident Scientist with the National Archive of Computerized Data on Aging (NACDA), for his intellectual guidance and extensive work on this manual. The first and second editions also drew heavily from two other volumes: Carolyn Geda's *Data Preparation Manual* (1980) and *Depositing Data With the National Institute of Justice* by Christopher S. Dunn and Kaye Marz (2000).

Third Edition. The *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* retains many of the key concepts from the previous two volumes and brings it up to date by discussing new tools and standards, and by emphasizing the life cycle approach to research data. ICPSR thanks the following individuals for their contributions to the current edition: Erik Austin, Corey Colyer, Darrell Donakowski, Russel Hathaway, Cynthia Hoxey, Peter Joftis, Kaye Marz, Shawn Marie Pelak, Amy Pienta, Ruth Shamraj, and Mary Vardigan. The Third Edition is the product of a collaboration among ICPSR and several of its projects, which also provided financial support:

- Child Care and Early Education *Research Connections*, sponsored by the United States Department of Health and Human Services, Administration for Children and Families, Child Care Bureau
- Data Resources Program, sponsored by the United States Department of Justice, National Institute of Justice
- Data Sharing for Demographic Research Project, sponsored by the National Institute of Child Health and Human Development, Demographic and Behavioral Sciences
- Human Subject Protection and Disclosure Risk Analysis Project, sponsored by the National Institute of Child Health and Human Development, Demographic and Behavioral Sciences
- National Archive of Criminal Justice Data, sponsored by the United States Department of Justice, Bureau of Justice Statistics

About ICPSR

Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR) is an organization of member institutions working together to acquire and preserve social science data, to provide open and equitable access to these data, and to promote effective data use. ICPSR facilitates research and instruction in the social sciences and related areas by archiving and disseminating data and by conducting related instructional programs.

The ICPSR Data Archive is unparalleled in its depth and breadth; its data holdings encompass a range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, health and medical care, education, child care research, law, and substance abuse. ICPSR also hosts several sponsored projects focusing on specific disciplines or topics. Social scientists in all fields are encouraged to archive their data at ICPSR.

Why Should I Archive Data?

ICPSR data advance scientific knowledge by making it possible for researchers around the world to conduct secondary analyses. ICPSR supports the social sciences by sharing data and methods with the research community, allowing for replication, verification, and extension of original findings. Archiving data with ICPSR ensures the long-term safekeeping of data, protecting it from obsolescence, loss, deterioration, or irreversible damage. Another advantage of archiving data with ICPSR is that our trained staff is available to provide user support. In addition, researchers often fulfill granting agency obligations by archiving data with ICPSR.

How Do I Deposit Data?

ICPSR works closely with researchers who submit their data collections for use by the social science research community. This publication provides useful information on the deposit process. For more information about the preparation of data for deposit and other general inquiries, please see www.icpsr.umich.edu/deposit or send e-mail to deposit@icpsr.umich.edu.

What Kinds Of Specialized Services Does ICPSR Offer?

ICPSR offers a number of specialized services designed to meet the needs of researchers collecting social science data:

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH

> In addition to quantitative data, ICPSR accepts qualitative research data (including transcripts, audiotapes, and videotapes) for preservation and dissemination. Please contact ICPSR for information specific to depositing qualitative data.

> ICPSR archives and disseminates data that require special handling and restrictions in order to protect human subjects. Restricted-use datasets require a special application process (e.g., data protection plan, IRB approval) and delivery of data on removable media (CD-ROM).

> ICPSR maintains a secure data enclave to store and protect data with the highest confidentiality standards. Data stored in the data enclave may be analyzed in our on-site, supervised computing facility with prior approval.

> Arrangements can be made to deposit data that cannot be disseminated immediately. This occurs when release would pose a confidentiality risk or when other dissemination plans have been made. See ICPSR's policy regarding preservation with delayed dissemination on our Web site.

> Data may be made available to the research community through online data analysis.

> ICPSR can create specialized Web pages to enhance data dissemination. For complex studies, ICPSR creates a user guide to assist new users in working with the data. This can be made available through the Web site along with other online features such as FAQs and electronic mailing lists.

> Training in documentation preparation and the Data Documentation Initiative (DDI) can be provided by ICPSR staff. Learn how to create DDI-compliant documentation, including an XML codebook.

> ICPSR staff can train other scholars in how to analyze your data through the Summer Program in Quantitative Methods.

Contact ICPSR for more information about these services or for other ways to customize your data products for dissemination.

Table of Contents

Table of Contents — *continued*

Introduction

Importance of Data Sharing and Archiving

Archives that preserve and disseminate social and behavioral data perform a critical service to the scholarly community and to society at large, ensuring that these culturally significant materials are accessible in perpetuity. The success of the archiving endeavor, however, ultimately depends on researchers' willingness to deposit their data for others to use.

In recent years, several national scientific organizations have issued statements and policies underscoring the need for prompt archiving of data, and some funding agencies have begun to require that the data they fund be deposited in a public archive. In its 2003 Data Sharing Policy, The National Institutes of Health (NIH) "reaffirms its support for the concept of data sharing" and goes on to state that "data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health." The National Science Foundation policy on data sharing states that "investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."

> "Data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health."
>
> *NIH Data Sharing Policy and Implementation Guidance*, March 5, 2003.

These statements from leading agencies supporting research demonstrate that the data sharing ethic is integral to maximizing the impact of research dollars. There are several other reasons to share data. Fienberg (1994) argues that data sharing, and hence archiving:

> Reinforces open scientific inquiry. When data are widely available, the self-correcting features of science work most effectively.

> Encourages diversity of analysis and opinions. Researchers having access to the same data can challenge each other's analyses and conclusions.

> Promotes new research and allows for the testing of new or alternative methods. Examples of data being used in ways that the original investigators had not envisioned are numerous.

> Improves methods of data collection and measurement through the scrutiny of others. Making data publicly available allows the scientific community to reach consensus on methods.

> Reduces costs by avoiding duplicate data collection efforts. Some standard datasets, such as the General Social Survey and the National Election Studies, have produced literally thousands of papers that could not have been produced if the authors had to collect their own data. Archiving makes known to the field what data have been collected so that additional resources are not spent to gather essentially the same information.

- Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

Frequently, outsiders bring fresh ideas and new analytic perspectives to a project. A good case can be made that a researcher may actually enhance the productivity (and certainly the visibility) of a project by early archiving.

Using the *Guide*

The *Guide to Social Science Data Preparation and Archiving* is aimed at those engaged in the cycle of research, from applying for a research grant, through the data collection phase, and ultimately to preparation of the data for deposit in a public archive. The *Guide* is a compilation of best practice gleaned from the experience of many archivists and investigators. The reader should note that the *Guide* does not attempt to address policies and procedures specific to certain archives, as they vary. Most public social science archives welcome investigators to contact them at any point in the research process to discuss their plans with respect to the design and preparation of public-use datasets.

Many investigators are more than willing to make their data available to others, but are unsure of how to go about preparing data for outside use, particularly in terms of complete documentation. The *Guide* is intended to help researchers document their datasets and prepare them for archival deposit. The *Guide* is written with the assumption that the reader is familiar with basic concepts of computerized data files, such as variables, labels, codes, and so forth.

Another assumption that informs this document is that the vast majority of readers will be familiar with statistical software packages like SAS, SPSS, and Stata, which are used in social science research. We refer to these three packages throughout the manual, although most observations hold true for other statistical programs. And while the *Guide* focuses on survey datasets, researchers routinely deposit many other kinds of data in public archives, ranging from social indicators at the country level to qualitative data. In most cases, the basic issues are similar. Also, we are mindful that the computing environment changes rapidly and that new techniques for managing, documenting, and analyzing data are continually being developed.

Dissemination-ready versus archive-ready data formats

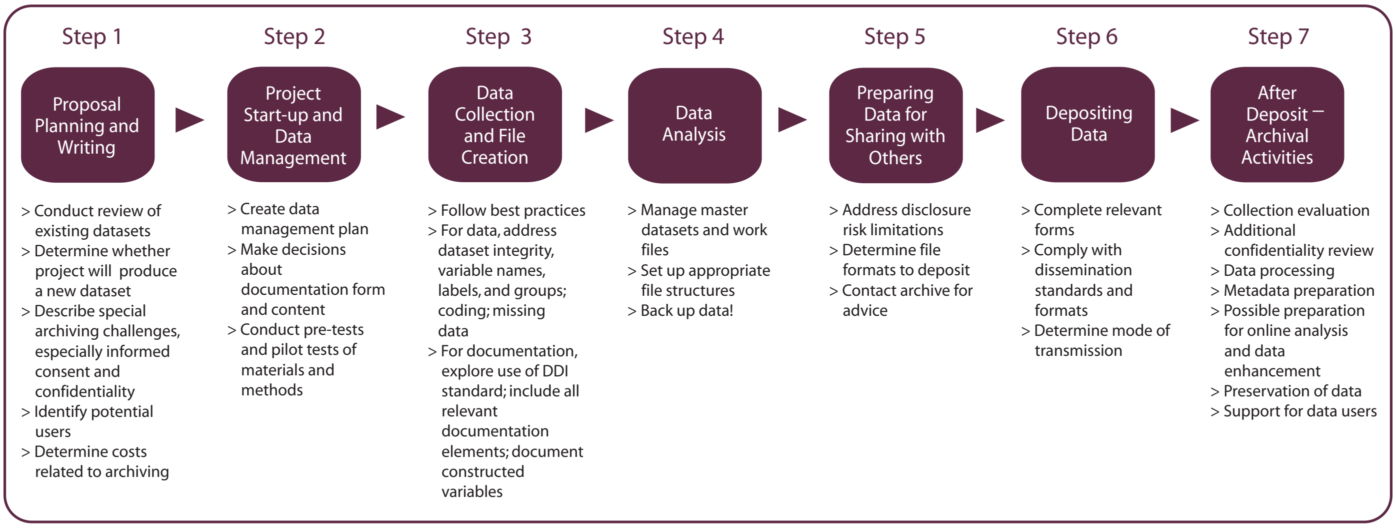
Note that there is a distinction between depositing data with dissemination to others in mind and depositing data in a format for preservation. The *Guide* primarily focuses on preparing data in such a way that it is suitable for dissemination and for sharing with others. Most archives have procedures to ensure that data are preserved for the long term, and those procedures may involve creating alternate versions of the data for preservation purposes.

Planning Ahead for Archiving

In recent years, the importance of archiving data has received more attention in scholarly literature. The social science research community has come to recognize that data producers must create preservation plans early on in the research data lifecycle. According to Jacobs and Humphrey (2004), “Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project’s life cycle and the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method.” Ideally, then, the researcher should plan for eventual archiving and dissemination of project data before the data even come into existence.

The data life cycle

We offer here a schematic diagram illustrating key considerations germane to archiving at each step in the data creation process. The actual process may not be as linear as the diagram suggests, but it is important to develop a plan to address the archival considerations that come into play across all stages of the data life cycle.



Step 1 — Proposal Planning and Writing
- > Conduct review of existing datasets
- > Determine whether project will produce a new dataset
- > Describe special archiving challenges, especially informed consent and confidentiality
- > Identify potential users
- > Determine costs related to archiving

Step 2 — Project Start-up and Data Management
- > Create data management plan
- > Make decisions about documentation form and content
- > Conduct pre-tests and pilot tests of materials and methods

Step 3 — Data Collection and File Creation
- > Follow best practices
- > For data, address dataset integrity, variable names, labels, and groups; coding; missing data
- > For documentation, explore use of DDI standard; include all relevant documentation elements; document constructed variables

Step 4 — Data Analysis
- > Manage master datasets and work files
- > Set up appropriate file structures
- > Back up data!

Step 5 — Preparing Data for Sharing with Others
- > Address disclosure risk limitations
- > Determine file formats to deposit
- > Contact archive for advice

Step 6 — Depositing Data
- > Complete relevant forms
- > Comply with dissemination standards and formats
- > Determine mode of transmission

Step 7 — After Deposit – Archival Activities
- > Collection evaluation
- > Additional confidentiality review
- > Data processing
- > Metadata preparation
- > Possible preparation for online analysis and data enhancement
- > Preservation of data
- > Support for data users

Archiving an existing collection

Steps 1–5 above are covered in Chapters 1–5 respectively. We recognize that not everyone is able to follow the ideal process outlined above. Some projects may have already collected data and may be at the point of depositing a collection into a public archive. While properly documented collections generally do not pose serious difficulties for archiving, we do recommend that data depositors read Chapters 3 (Data Collection and File Creation Phase) and 5 (Preparing Data for Sharing With Others) to make sure that their collections meet minimum archival standards. Steps 6–7 are not addressed in this volume, however they are included to shed light on typical archival procedures. Step 6 represents the activities that a depositor undertakes in conjunction with a specific archive, and step 7 outlines the main activities undertaken by the archive after data are deposited. We recommend that depositors closely follow the procedures of the archive with which they plan to deposit data.

Chapter 1 Proposal Planning and Writing Phase

In keeping with the principles of responsible archiving, plans for archiving should be fleshed out while the researcher is at the stage of outlining and writing the actual grant application. Funding agencies increasingly require that applications for support include data sharing and dissemination plans. Thinking ahead during this early phase of the project permits the researcher to take into account important issues — particularly issues related to disclosure risk — from the very beginning, which can simplify the process and avert problems later on at the data deposit stage.

In its 2003 Data Sharing Policy, NIH suggests that applicants for funding specify in their proposals the following:

- Schedule for data sharing
- Format of final dataset
- Documentation to be provided
- Analytical tools to be provided, if any
- Need for data sharing agreement
- Mode of data sharing

Important Steps to Follow

In addition, we offer the following steps as guidance during the proposal planning and writing stage:

Conduct a review of existing datasets

If the proposed research is to involve data collection or data acquisition, a thorough review of existing data on the topic should be conducted so that the applicant can state why currently available datasets are inadequate for the proposed research. In addition to the usual literature search, it is recommended that the data catalogs of the major archives be reviewed.

Determine whether a new dataset will be produced and whether the data should be archived

Note that when writing the grant proposal, it is useful to think of "data" in the widest sense when looking ahead to depositing materials. Data can exist in a number of formats and a number of types: numeric data files, interview transcripts, and other qualitative materials such as diaries and field notes also qualify as data and may need special handling.

Some projects do not involve original data collection and/or may involve combining data from one or more secondary sources. Derived datasets created by collating materials from existing sources and presenting these in a new way could constitute a new dataset. A project may also be considered to be producing a new dataset if it combines (a) primary and secondary data; (b) secondary data with newly created variables; or (c) secondary data based on data that is not yet publicly available. If the project meets any of the above conditions and would be useful to other researchers in supporting new research, reproducing original findings, or testing new hypotheses, then archiving the dataset should be considered. See Chapter 5, Archiving Data From Secondary Sources for a more in-depth discussion.

Describe any special challenges that might arise when archiving the data

If one envisions any difficulties in making the data available for secondary research, these difficulties should be outlined in the grant application. Difficulties associated with depositing or archiving materials are usually centered around one of the following issues:

- Confidentiality
- Consent
- Copyright

Any problems applicants foresee regarding these or any other issues relevant to the archiving of data should be explicitly spelled out during proposal preparation. This should be accompanied by evidence that thought has been given to strategies to overcome such problems and that there is a general willingness to archive data. If the investigator considers that the data in question may not be appropriate for archiving, it is worthwhile to consult with archive staff at this stage to determine whether this is indeed the case from the archival perspective.

> Plans for archiving should be fleshed out while the researcher is at the stage of outlining and writing the actual grant application.

Informed consent and confidentiality considerations. It is never too early to be thinking about issues related to informed consent and confidentiality. Protection of individuals' privacy is a core tenet of responsible research practice, and any project must address this topic in a careful and thorough manner.

Informed consent is the term given to the communication process allowing individuals to make an informed choice about participation in a research study. This process is reflected in an informed consent document that contains specific, required information about the research study. The informed consent document serves as the formal authorization by an individual of their agreement to participate in the proposed research. The human subjects involved in a project must participate willingly, having been adequately informed about the research. In preparing the informed consent document, investigators are asked to include a statement describing the extent to which confidentiality of records identifying the subject will be maintained. This has implications for one's ability to share data with the research community.

In their informed consent guidelines, the Behavioral Sciences Institutional Review Board at the University of Michigan suggests including the following text in the informed consent document: "You will not be identified in any reports on this study. Records will be kept confidential to the extent provided by federal, state, and local law. However, the Institutional Review Board, the sponsor of the study (i.e. NIH, FDA, etc.), or university and government officials responsible for monitoring this study may inspect these records" (www.irb.research.umich.edu/IRB_HSBS_Shared/consent.html).

If an investigator is planning to archive and share data with the research community, the informed consent document should not promise that the data will be shared with the research team exclusively. Rather, subjects should be informed that when data are shared or made available to others for secondary analysis the information provided will be kept confidential.

Describe the potential users of the dataset

It is helpful to specify in the grant proposal who the likely users (academic or non-academic) of the datasets are. Most potential users will be within the higher education research community, but increasingly policymakers and practitioners are using research data. If the dataset has commercial or other uses, this should also be stated in the application for funding.

Determine the costs of preparing the data and documentation for archiving

The investigator should outline the plans for and cost of preparing the data and documentation for archiving. The various activities typically associated with preparing data are presented below, for which grant applicants should attach appropriate cost estimates.

- For quantitative data, investigators should allow time to create system-specific files with appropriate variable and value labeling, to supply the syntax for derived variables, etc.
- Investigators need to make sure that they include in their grant applications adequate time and money for the preparation of high-quality documentation. Documentation is invaluable in enabling secondary analysts to understand the data and provide for informed reuse of the material. Good supporting documentation should include (if applicable) the following: questionnaire, coding frame, details of sampling and methodology, interview schedule, and topic guide.
- Consent and confidentiality issues impact costs for archiving. For clarity, consent forms should be drawn up at the start of the project to obtain permission for archiving, with consideration given to whether the data will be available publicly. Confidentiality agreements made with interviewees should not impede data archiving.
- It is strongly recommended that a period of time be costed to prepare and collate materials for deposit. This normally comprises the majority of the costs for archiving.

Consider alternative archiving options

Self-dissemination. Despite the clear advantages that dissemination of data through a public archive can provide, some data collectors may choose to disseminate their own data, especially while funding is available to support this activity. If this is the case, it is recommended that the data producer arrange for eventual archiving of the data after the self-dissemination terminates and specify the schedule for data sharing in the grant application. The archive may want to make a preservation copy during the period of self-dissemination for a number of reasons: (1) to develop expertise with the data; (2) to process the data while knowledgeable staff are available to work with; and (3) for general safekeeping. Generally, an archive will enter into an agreement to make an archival copy for later dissemination only when there is an agreed upon date for eventual dissemination by the archive.

Preservation with delayed dissemination. Another possible approach, which again should be outlined up front in the application for funding, is preservation with delayed dissemination. Under such an agreement the archive and the data producer will arrange for archival preservation of the data with dissemination to occur at a later date. Issues regarding the schedule for eventual dissemination, embargo periods, and human subject protections specific to these studies will be settled prior to deposit as will ground rules on the extent of processing by archival staff while the study remains in the "preservation with delayed dissemination" category. As with self-dissemination, staff will need to develop expertise with the data and possibly perform processing while knowledgeable project staff are available for assistance.

Restricted-use collections. As previously mentioned, the issue of confidentiality is of paramount importance when conducting research that requires human subjects. Before sending data to an archive, data depositors are usually asked to ensure that information that could be used to identify research participants be masked or "blanked." These adjustments, however, may impose limitations on the research uses of such files. In writing a proposal, then, it is useful to think about whether the data will ultimately be made available in a public-use or restricted-use form. In some cases, archives can provide both forms. A restricted-use version of the collection that includes confidential information can be prepared and offered to approved researchers under a set of controlled conditions. The restricted-use dataset approach is an effective way to permit access to confidential and/or sensitive research and has proven acceptable to researchers. To ensure an even greater level of security, some archives provide confidential data through on-site data enclaves, which require that researchers visit the facility to access the data under restricted conditions.

Other considerations

All researchers are encouraged to document the research and fieldwork process and ensure that data are held in an organized manner, and that any consent and confidentiality concerns, which may inhibit subsequent archiving of data, are resolved.

The investigator should outline the plans for and cost of preparing the data and documentation for archiving. This should be planned in conjunction with an archive.

It is also a good idea to review archival guidelines with respect to suitable data formats, types of media, and software for deposit prior to starting fieldwork. Researchers may elect to conduct analyses using a certain software package but have plans to convert the data to other formats later based on the preferences of the archive with which they deposit their data.

Note that if the data to be collected are not typical numeric digital data, costs and plans may vary widely. If the data are, for example, qualitative (e.g., verbatim interviews) rather than quantitative, mixed qualitative and quantitative, or video, audio, or coordinate-based geographic data, special concerns come into play and must be addressed in the grant application.

Chapter 2 Project Start-Up and Data Management Planning Phase

Importance of a Data Management Plan

Once funding is received and the research project has started, the researcher will want to continue to think about and plan for the final form of the collection that will ultimately be deposited in an archive. Planning for the management and archiving of a data collection project at the outset is critical to the project's success. The cost of a project can be significantly reduced if careful planning takes place early in the project.

Documentation as part of the project plan

Documentation should be as much a part of project planning as questionnaire construction or analysis plans. At a minimum, a project plan should involve decisions on the following data and documentation topics:

File structure. What is the data file going to look like and how will it be organized? What is the unit of analysis? Will there be one long data record or several smaller ones?

Naming conventions. How will files and variables be named?

Data integrity. How will data be converted to electronic form, and what checks will be in place to find invalid values, inconsistent responses, incomplete records, etc.?

Preparing dataset documentation. What will the dataset documentation look like and how will it be produced? What information will it contain? (See Chapter 3, Best Practice in Creating Technical Documentation for guidance in using a standards-based approach to documentation production.)

Variable construction. What variables will be constructed following the collection of the original data? How will these be documented?

Project documentation. What steps will be taken to document decisions taken as the project unfolds? How will information be recorded on field procedures, coding decisions, variable construction, and the like?

Integration. To what extent can the various tasks mentioned above be integrated into a single process?

The last point is critical. To the extent that one can use a single computer program or an integrated set of programs to carry out these tasks, they are made simpler, less expensive, and more reliable. A little planning at the beginning of the project regarding which programs will be used for which tasks goes a long way.

Computer-assisted interviewing

Computer-assisted interviewing (CATI/CAPI) is being used increasingly for both telephone and personal interviews. These programs — e.g., Blaise, CASES – typically perform a number of functions simultaneously including direct data entry, integrity checks, and skips and fills. Somewhat similar software can be used to format mail questionnaires and prepare data entry templates.

If at all possible, it is desirable to program the instrument to be fielded according to specifications of the resulting data files. Keeping a focus on the ultimate desired form of the data collection can make dataset preparation that much easier.

Using integrated software

Most large-scale data collection efforts now involve computer-assisted interviewing, but there are still situations in which data entry will be required — e.g., inputting of administrative records, observation data, or open-ended question responses. In the past several years, a number of software tools have become available to make the documentation task easier. For projects requiring data entry directly from mail questionnaires or interview instruments, a variety of programs will not only make data entry a good deal easier but will also carry out data integrity checks as the data are entered and create programming statements to read the data into other programs. A good data-entry program will also recognize automatic skips and fills. For example, suppose that a questionnaire contains a series of items on work experience. If the respondent has never worked, then as soon as that code is keyed, the program skips to the next valid entry, filling in missing data codes in intervening fields as appropriate.

Documentation should be as much a part of project planning as questionnaire construction or analysis plans.

Spreadsheet packages can also be used for data entry. These packages usually can be programmed to perform integrity checks as data are entered. In addition, a variety of database packages such as Microsoft Access, MySQL, and Oracle can be used for both data entry and documentation. Note that when such systems are intended to serve as the format for deposit, it is important to provide full documentation for all of the fields and relationships built into the files.

Other kinds of software can be used to perform many documentation tasks. For example, word processing packages like Microsoft Word and WordPerfect can be used for data entry, maintenance of dataset documentation, and similar tasks, but they are not suitable tools for data integrity checks. Producing an attractive final document using word processing is also quite simple. In fact, if the basic document has been set up in a word processor, retrieving and merging statistical information such as frequencies and descriptive statistics from computer output stored in an external file is a relatively easy task. See Chapter 3 for a discussion of using the Data Documentation Initiative (DDI) metadata specification to produce more robust documentation in eXtensible Markup Language (XML) format. The DDI standard provides a way to produce comprehensive documentation that is consistent in format and thus easy to integrate into larger systems.

Data entry and documentation as part of pretests and pilot studies

Conducting pretests or pilot studies is a good way to uncover potential problems with all aspects of a project. There are two major reasons to include both data entry and documentation as part of this initial phase. First, the cost of a project does include these factors, and the best way to estimate those costs is to pretest them. Secondly, pretest data entry and documentation reveal unanticipated difficulties in record layouts, naming conventions, etc. Actually, the cost of including these steps in the pretest need not be all that high. The most expensive aspect — data entry — should be low, since the pretest covers only a small number of cases. The investigator may not want to prepare a full-scale codebook on the basis of pretest, but it is a good idea at least to prepare a mockup, or to work out the codebook layout for a few variables.

Chapter 3 Data Collection and File Creation Phase

Characteristics of "Good" Data Collections

From the archivist's and the end user's perspective, a "good" dataset is one that is easy to use. Its documentation is clear and easy to understand, the data contain no surprises, and users are able to access the dataset with relatively little start-up time. As noted above, pre-project planning goes a long way to preparing "good" data for there are many details to attend to as a researcher carries a project to a successful conclusion. Much of what follows may be obvious to an experienced researcher, but perhaps not to those undertaking a data collection effort for the first time.

Following best practice in terms of building both the data and documentation components of a collection is critical. Below we describe aspects of best practice that conform to widely accepted norms in the social sciences.

Best Practice in Creating Data Files

Dataset integrity

The path taken from a questionnaire or interview to an actual data record can be a perilous journey with many potential accidents waiting to happen. Inevitably, when the first analyses are performed on newly minted data, anomalies emerge — codes that do not make sense, records that do not match, and 5-year-olds with six children. Such flaws are inevitable. The task is to find them and fix them, but an obvious first step is to prevent them. Although no system is perfect, several steps can be taken in advance to lessen the incidence of errors.

> As noted previously, use a data-entry program that is designed to catch typing errors, i.e., one that is pre-programmed to catch impossible values.

> Consider double entry in which each record is keyed in and then re-keyed against the original. Several standard packages offer this feature. In the re-entry process, the program catches discrepancies immediately.

> Carefully check the first 5 to 10 percent of the data records created, and then choose random records for quality-control checks throughout the process.

> Separate the coding and data-entry tasks as much as possible. Coding should be performed in such a way that the data-entry person does not have to concentrate on a second task.

> Arrange to have particularly complex tasks, such as occupation coding, carried out by one person or by a team of persons specially trained for the task.

> Let the computer do complex coding and recoding if possible. For example, if a series of variables describing family structure is to be created, write computer code to perform the task rather than create the variables by hand. Not only are the computer codes absolutely accurate if the instructions are accurate, but they can also be easily changed to correct a logical or programming error.

Despite all these good intentions, errors will undoubtedly occur. Here is a list of things to check.

Wild codes and out-of-range values. Frequency distributions will usually reveal this kind of problem, although not every error is as obvious as, say, a respondent with 99 rather than 9 children. Sometimes frequency distributions will contain entirely valid values but just “not look right.” For example, the columns for a given variable might have been defined incorrectly, and so the data might be moved over one place to the right or left. In addition to checking frequency distributions, it is a good idea to generate data plots, which tend to reveal outliers instantly.

Consistency checks. Checks for consistency are far more difficult to perform than checks for wild codes because they require knowledge of the substantive task at hand. Typically, they require comparisons across variables. A common problem is when a “gate” or “filter” item, such as “Did you work last week?,” is inconsistent with items that follow. For example, a respondent says she did not work, but reports earnings. Other consistency checks involve complex relationships among variables, e.g., unlikely combinations of respondents’ and children’s ages. At a minimum, consistency checks should be run to assure that fields that have applicable responses contain valid values, and fields that are not applicable contain only missing values.

> From the archivist's and the end user's perspective, a "good" dataset is one that is easy to use. Its documentation is clear and easy to understand, the data contain no surprises, and users are able to access the dataset with relatively little start-up time.

Record matches and counts. There are some cases in which each “subject” in a study should or might have more than one record. This occurs most frequently in a longitudinal study in which each subject has multiple records, even if the subjects were not actually interviewed at a given time point. In other cases, the number of additional records may actually vary from subject to subject. For example, one might have a household record, followed by a varying number of person records. This is sometimes known as a *hierarchical file*. Here the tasks are ensuring, to the extent that software permits, that (a) the *header record* contains a count of the number of *trailer records*, and (b) consistency checks are made on the counts.

See Chapter 5, File Structure, for more information on best practice in setting up files with different record types.

Variable names

Two basic issues arise regarding variable names: (a) should an investigator construct a set of standard variable names in the first place, and, if yes, (b) what system should be used? There are several methods of constructing a set of variable names.

One-up numbers. In this system, each variable is numbered 1 through *n*, with *n* equal to the total number of variables. Since most computer programs do not handle variable names starting with a digit, the usual format is V1 (or V0001) . . .V*n*. This approach has the advantage of simplicity and the disadvantage of a lack of information. Even though almost any standard program will allow extended labels for variables, so that one can append information that V0023 is really “Q6b, Mother’s Education,” the number system is prone to error.

Question numbers. It is also possible to name variables corresponding to question numbers, e.g., Q1, Q2a, Q2b. . .Q*n*. This approach has the advantage of relating directly to the original questionnaire, but, like one-up numbers, it has the disadvantage of not being easily remembered. Further, it is not uncommon for a single question to yield several distinct variables.

Mnemonic names. As alluded to above, names chosen to represent the meaning of the actual variable have some advantages, principally in that they are recognizable and memorable. However, there are disadvantages. First, what is an "obvious" abbreviation to the person who created it may not be obvious to a new user. Second, with a limited number of characters to work with, it is not all that easy to create names with immediately recognizable referents. Finally, it is difficult to maintain consistency across variables that share common content, e.g., always to use ED for education.

Prefix, root, suffix systems. A more systematic version of the previous point is to think of each variable name as containing a root, possibly a prefix, and possibly a suffix. For example, suppose all variables having to do with education had the root ED. Mother's education would then be MOED, father's education FAED, etc. Suffixes are often used to indicate the wave of data in longitudinal studies, the form of a question, or other such information. Implementing this system requires prior planning to establish a list of standard two- or three-letter abbreviations.

It is important to remember that the variable name is the referent that analysts will use most often when working with the data. At a minimum, it should convey correct information, and ideally it should be unambiguous in terms of content.

Length. For cross-platform and system portability, variable names of eight characters are recommended, although newer versions of some statistical packages allow longer variable names.

Variable labels

Most statistical programs permit the user to include extended labels for each variable. These labels are extremely important. They should provide at least three pieces of information: (1) the item or question number in the original data collection instrument (unless item number serves as the variable name), (2) a clear indication of what the variable contains, and (3) an indication of whether the variable is constructed from other items. If the number of characters available for labels is constraining, it is a good idea to work up a set of standard abbreviations in advance. These should be shown in the dataset documentation.

Variable groups and corresponding variable group lists in the codebook are an effective way of organizing a dataset and are especially recommended if a collection contains a large number of variables.

Variable groups

Variable groups and corresponding variable group lists in the codebook are an effective way of organizing a dataset and are especially recommended if a collection contains a large number of variables. Variable groups enable secondary analysts to get an overview of the data quickly. Further, they are useful when the data are provided through an online analysis system as they offer an existing navigational structure for exploring the dataset.

Codes and coding

Before survey data are analyzed, the verbal interview responses must be represented by numeric codes (Babbie, 1990). This section discusses common coding conventions, which (a) assure that all statistical software packages will be able to handle the data, and (b) promote greater measurement comparability. Note that in computer-assisted interviewing systems, codes are precoded into the instrument and are assigned automatically; this means that most coding decisions are made early on in the process, before the instrument is fielded. Nevertheless, principles provided here should apply to most coding situations.

No attempt is made here to provide standardized coding schemes. However, the U.S. Bureau of the Census occupation and industry codes and the National Bureau of Standards state, county, and metropolitan area codes are standard codes used for coding this type of information. Below are guidelines to keep in mind during the coding process.

> Identification variables. Provide enough space at the beginning of the record to accommodate all identification variables. Identification variables often include a study number and respondent number to represent each study and case uniquely.

> Code categories. Code categories should be mutually exclusive, exhaustive, and precisely defined. Each interview response should fit into one and only one category. Ambiguity will cause coding difficulties and problems with the interpretation of the data.

> Preserving original information. As much detail as possible should be preserved. The recording of original data, such as age and income, is more useful than collapsing or bracketing the information. Original or detailed data permit the secondary analyst to determine other meaningful brackets rather than being restricted to coded versions of the data.

> Closed-ended questions. Questions that are precoded in the questionnaire should retain this coding scheme in the machine-readable data to avoid errors and confusion.

> Open-ended questions. There are two main approaches to developing a coding scheme for open-ended items. An investigator can either use a predetermined coding scheme or review the initial survey responses to construct the scheme based on major categories that emerge. The approach taken depends upon the nature of the research and the researcher's objectives.

- Increasingly, the full verbatim text of open-ended responses is submitted to archives so that users can make their own determinations about how these responses should be coded. Such responses may contain sensitive information and thus will need to be reviewed and possibly treated by archives prior to dissemination.

> Check-coding. It is a good idea to check-code a portion of the cases during the coding process. Finding more than one code for the same interview response will highlight any problems or ambiguities in the coding scheme. Also, check-coding provides an important means of quality control throughout the coding process.

> Series of responses. When more than one column is required to handle a series of responses, such as in the example below, organizing the responses into meaningful major classifications is helpful. Responses within each major category can be given the same first digit. The second order or secondary digits can represent another category or be used to distinguish the specific response within the major category. This type of coding scheme permits analysis of the data in terms of broad groupings as well as individual responses or categories.

- A partial example of the use of this type of scheme, taken from the 1990 Census of Population and Housing Public Use Microdata Samples (PUMS) person record, is presented below.

Coding Scheme for Employment Status of Parents

000 N/A (not own child of householder, and not child in subfamily)

Living with two parents:

Both parents in labor force:

111 Both parents at work 35 or more hours
112 Father only at work 35 or more hours
113 Mother only at work 35 or more hours
114 Neither parent at work 35 or more hours

Father only in labor force:

121 Father at work 35 or more hours
122 Father not at work 35 or more hours

Mother only in labor force:

133 Mother at work 35 or more hours
134 Mother not at work 35 or more hours

Living with one parent:

Living with father:

211 Father at work 35 or more hours
212 Father not at work 35 or more hours
213 Father not in labor force

Living with mother:

221 Mother at work 35 or more hours
222 Mother not at work 35 or more hours
223 Mother not in labor force

Missing data

In the past few years, enormous strides have been made in both missing data imputation and missing data analysis (Little and Schenker, 1995). None of these new methods can be used, however, if the original data file does not handle missing data properly.

In general, it is not a good idea to use blanks as missing data codes. They are acceptable only when an entire case is missing a large number of variables, such as in a follow-up interview in a longitudinal study, or when an entire sequence of variables is missing, such as data on nonexistent children. In any such instance there should be an indicator variable allowing the analyst to determine unambiguously when cases should have blanks in particular areas of the data record.

Missing data can arise in a number of ways, and it is important to distinguish among these different instances. There are at least five missing data situations, each of which should have a distinct missing data code.

1. *Refusal/no answer.* The subject explicitly refused to answer the question or did not answer the question when he or she should have.

2. *Don't know.* The subject was unable to answer the question, either because he or she had no opinion or because the required information was not available (e.g., a respondent could not provide family income in dollars for the previous year).

3. *Processing error.* For some reason, there is no answer to the question, although the subject provided one. This can result from interviewer error, incorrect coding, machine failure, or other problems.

4. *Not applicable.* The subject was never asked the question for one reason or another. Sometimes this results from "skip patterns" that occur, for example, when subjects who are not working are not asked questions about job characteristics. Other examples are sets of items asked only of random subsamples and items asked of one member of a household but not another.

5. *No match.* This situation may arise when data are drawn from different sources (for example, a survey questionnaire and an administrative database), and information from one source cannot be located.

If missing data have been imputed in any way, it is highly desirable to indicate this. There are two more or less standard ways of doing so. The first is to include two versions of any imputed variables, one being the original, including missing data codes, and the second being an imputed version, containing complete data. A second approach is to create an "imputation flag," or indicator variable, for each variable subject to imputation, set to "1" if the variable is imputed and "0" otherwise. (Not all missing data need be imputed. In the case of job characteristics, for example, the investigator might want to impute responses for "don't know" and "refuse" cases, but not for cases where the respondent is not working.)

Missing data can arise in a number of ways, and it is important to distinguish among these different instances. There are at least five missing data situations, each of which should have a distinct missing data code.

Selecting missing data codes

Missing data codes should always match the content of the field. If the field is numeric, the codes should be numeric, and if the field is alphabetic, the codes may be numeric or alphabetic. Most researchers use codes for missing data that are above the maximum valid value for the variable (e.g., 97, 98, 99). Occasionally, this presents problems, most typically when the valid values are single-digit values but two digits would be required to accommodate all necessary missing data codes. The same problem sometimes arises if negative numbers are used (e.g., -1 or -9), because of the necessity to accommodate the minus sign. Hence, missing data codes should be standardized such that only one code is used for each missing data type across all variables in the data file, or across the entire collection if the study produced multiple data files.

A note on "not applicable" and skip patterns

Although we have referred to this issue in several places, some reiteration is perhaps in order. Handling skip patterns is a constant source of error in both data management and analysis. On the management side, deciding what to do about codes for respondents who are not asked certain questions is crucial. "Inap." codes, as noted above, should be distinct from other missing data codes. It is not a good idea merely to leave the record blank. Dataset documentation should clearly show for every item exactly who was asked and was not asked the question. At the data cleaning stage, all "filter items" should be checked against items that follow to make sure that no one provides answers to the item who should not, and that those who did not answer the item have the correct kind of missing data code.

Best Practice in Creating Technical Documentation

Preparing high-quality technical documentation, sometimes called a codebook, can be a time-consuming task. In this section, we lay out the contents of ideal technical documentation of a data collection. The cost of preparing such documentation can be significantly reduced by planning ahead. For example, a standard word processor can be used to lay out a page for each variable, and then frequency distributions can be "dropped into" the text by reading computer output directly and using cut-and-paste routines. Or, one can take the computer output itself and turn it into documentation by adding additional material. The task of preparing page layouts can be reduced by creating a standard template and then filling in the blanks for each variable.

Since most standard computer programs will produce frequency distributions that show counts and percents for each value of each numeric variable, it may seem logical to use that information as the basis for documentation, but there are several reasons why this is not recommended. First, the output typically does not show the exact form of the question or item. Second, it does not contain much of the other information referred to above — skip patterns, derivations of constructed variables, etc. On the other hand, as noted above, such output might be an excellent starting point for good documentation. One could simply import the electronic form of the output into a standard word processor and add other required information.

The DDI is a project to establish an international standard for the content, presentation, transport, and preservation of documentation about datasets in the social and behavioral sciences.

Data Documentation Initiative: An emerging standard

We encourage data producers to generate documentation that is "marked up" according to the Data Documentation Initiative (DDI) metadata specification, an emerging international standard for the content, presentation, transport, and preservation of documentation about datasets in the social and behavioral sciences. The DDI specification is written in XML, which permits the markup, or tagging, of technical documentation elements for content — e.g., question text or sampling information.

The DDI approach offers several advantages. First, all information that the analyst needs is available in one document, from which other products, such as setup files, can be produced. Second, the XML file can be viewed with Web browsers and lends itself to Web display and navigation. Third, because the content of each field of the documentation is tagged, the documentation can serve as the foundation for extract and analysis engines and other intelligent agents. Finally, preparing documentation in DDI format at the outset of a project means that the documentation will also be suitable for archival deposit and preservation. For more information on DDI and a list of tools and other XML resources, please consult the DDI Web site at www.icpsr.umich.edu/DDI.

Several XML authoring tools that data producers can employ to create DDI documentation are now available. The user imports the DDI Document Type Definition (DTD) or XML Schema into the software and is then able to enter text for specific DDI elements and attributes. The resulting document is a valid DDI instance or file. There are also DDI-specific tools, such as the Nesstar Publisher, which produce DDI-compliant XML markup automatically.

DDI and question text. Note that most archives in which the data will be deposited will prefer or at least readily accept documentation submitted in DDI format. To be in full compliance, a document should have question text integrated into each variable.

Important documentation elements

It may not be possible for a project to produce documentation that is DDI-conformant. In those situations, using a uniform, structured format with integrated question text is the best alternative, as it will enable the archive to convert the files to XML format easily.

A list of recommended topics to be covered in technical documentation is presented below. Note that many of the high-level elements have counterparts in the Dublin Core Metadata Initiative (DCMI) element set. The DCMI is a standard aimed at making it easier to describe and to find resources using the Internet. For more information on the DCMI, please view their Web site at dublincore.org/.

Principal investigator(s) [Dublin Core Creator]. Principal investigator name(s), include affiliation at time of data collection.

Title [Dublin Core Title]. Official title of the data collection.

Funding sources. Grant number and related acknowledgments.

Data collector/producer. Persons or organizations responsible for data collection, and the date and location of data production.

Project description [Dublin Core Description]. This should describe the project and its intellectual goals, indicate how the data articulate with related datasets, and provide other essential information. Much of the essential information about the project is probably available in various publications resulting from it, and these should be cited. A brief project history, detailing any major difficulties faced or decisions made in the course of the project, would be useful.

Sample and sampling procedures. This section should describe the population being investigated and the methods used to sample it (assuming the entire population is not covered). The discussion of the sampling procedure should indicate whether standard errors based on simple random sampling are appropriate, or if more complex methods are required. If weights are required, they should be described. If available, a copy of the original sampling plan should be included as an appendix. A clear indication of the response rate should be given, indicating what proportion of those sampled actually participated in the study; the retention rate, if applicable, should also be noted. This section should also clearly indicate when the study was conducted.

Weighting. Information on weight variables and how they should be used is critical.

Date, geographic location of data collection, and time period covered [Dublin Core Coverage]. These are key pieces of information that must be supplied.

Data source(s) [Dublin Core Source]. If a dataset was compiled from other resources, it is important to provide a list of the original source files or documents.

Unit(s) of analysis/observation. The unit of analysis describes who or what is being studied.

Variables. For each variable, the following information should be provided:

1. *The exact wording of the question or the exact meaning of the datum.* If the question is drawn from previous surveys or published work, the source should be noted.

2. *The text of the question integrated into the variable text.* If that is not possible, it is useful to have the item or questionnaire number, e.g., Question 3a, so that the archive can make the necessary linkages.

3. *Universe information, i.e., who was actually asked the question.* In other words, if there is a skip pattern such that some items were not asked of all respondents, that information should be shown on the same page as the rest of the information for the item.

4. *Unweighted frequency distributions or summary statistics for the item.* These distributions should show both valid and missing cases.

5. *Missing data codes.* Codes assigned to represent data that are missing. The codes used typically fall outside of the range of valid values.

6. *Imputation and editing information.* If the item has been estimated for any subjects, or if any extensive editing has been done, that information should be made available.

7. *Details on constructed variables created by the project staff.* Often the variables actually used in analysis are constructed on the basis of other variables. Some of these are anticipated at the time of data collection, and others come about in the course of the project. Of course, all of the information listed here should be shown for constructed variables, but it is also important to maintain an "audit trail" for such variables, indicating exactly how they were constructed, what decisions were made about imputations, and the like. If possible, the documentation should show the exact programming statements used to construct the variable.

8. *Exact meaning of codes.* The documentation should show the interpretation of the codes assigned to each variable. For some variables, such as occupational codes, this information might appear in an appendix.

9. *Location in the data file.* For raw data, show the column location and the record number if there is more than one record per case. If the file is in a software-specific system format, location is not important, but the order of the variables in the dataset is. Normally, but not necessarily, the documentation will be in the same order as the file. If not, the position number of the variable needs to be shown.

10. *Variable groupings.* For large datasets, the documentation should categorize variables into conceptual groupings.

Technical information on files. This documentation component includes information on file formats, file linking, and similar information.

Data collection instruments. Include copies of the original data collection forms and instruments if at all possible. Other researchers often want to know the context in which a particular question was asked, and it is helpful to see the survey instrument as a whole. Increasingly, this is difficult because CATI/CAPI programs often do not provide a hardcopy version of the interview, or if they do, it is in a format that is difficult to read.

Flowchart of the data collection instrument. For complex questionnaires it is sometimes useful to produce a graphic guide to the data, showing which respondents were asked which questions and how various items link to each other. This is particularly useful when no hardcopy questionnaire is available.

Index or table of contents. For large datasets, this is essential. An alphabetized list of variables with associated page numbers to detailed variable information is also extremely helpful.

List of abbreviations and other conventions. Both variable names and variable labels will contain abbreviations. Ideally, these should be standardized.

Interviewer guide. If available, such documents are useful to secondary analysts in order to understand how interviews were administered.

Recode logic. It is important to provide an audit trail of the steps involved in creating recoded variables. This information is sometimes provided in a separate document.

Coding instrument. Rules and definitions used for coding the data are helpful to data analysts.

Chapter 4 Data Analysis Phase

In this chapter, we address important issues to be aware of during the analysis phase when project staff are actively working with data files to investigate their research questions. During this phase, it is essential to maintain the project's data in an organized manner to ensure the accuracy of results and findings.

Master Datasets and Work Files

As analysis proceeds, there will be various changes, additions, and deletions to the dataset. Despite the most rigorous data cleaning, additional errors will undoubtedly be discovered. The need to construct new variables might arise. Staff members might want to subset the data by cases and/or variables. Thus, there is a good chance that before long multiple versions of the dataset will be in use. It is not uncommon for a research group to discover that when it comes time to put out a "final version" of the data for archiving, there are multiple versions that must be merged to include all of the newly created variables. This problem can be obviated to a degree if the research group works on a PC network, where, like a mainframe, there is a single version of the data.

It is a good practice to maintain a master version of the dataset that is stored on a "read only" basis.

It is a good practice to maintain a master version of the dataset that is stored on a "read only" basis. Only one or two staff members should be allowed to change this dataset. Ideally, this dataset should be the basis of all analyses, and other staff members should be discouraged from making copies of it. If a particular user of the data wants to create new variables and save them, a choice should be made between creating a work file for that researcher or adding the new variables to the master dataset. If the latter route is chosen, then all of the standard checks for outliers, inconsistencies, and the like need to be made on the new variables, and full documentation should be prepared. The final dataset reflecting published analyses is the version to archive.

Version numbers

One way to keep track of changes is to maintain explicit versions of a dataset. The first version might be that which comes from the data collection process, the second version that which emerges from data cleaning, the third that which results from composite variable construction, and so forth. With explicit version numbers, which are reflected in dataset names, it becomes easier to match documentation to datasets and to keep track of what was done by whom and when.

Raw data vs. statistical system files

Data may be maintained for analysis purposes in a number of different formats. From the standpoint of data storage, system files take up less space than raw ASCII data and permit the user to perform analytic tasks much more efficiently. System files, which are the proprietary formats of the major statistical programs, are extremely efficient because the statistical package reads in the data values and the various data specifications, labels, missing data codes, etc., only once and then accesses the system file directly afterwards. Because the data are stored on disk directly in the machine's internal representation, the step of translating the ASCII data each time to the internal binary representation of the specific machine is avoided. Many research groups use system files for all their data analysis and data storage after the first reading of the ASCII version. Although this is an efficient way to work, it is important to keep in mind that system files created in older versions of statistical packages may be readable only on the specific systems that created them. Recent versions of some software, however, produce files that are compatible across platforms and systems.

Most packages also create what are called “export/transport files” or “portable files,” as noted earlier. These kinds of files can be read on other machines, preserving all of the variable labeling and identification information. Increasingly, these are the formats that archives prefer to receive.

File Structure

Flat rectangular files. Having collected data, the researcher is faced with the question of what form the computer record should take. For the vast majority of datasets this is a very simple decision; the data are simply strung out in one long record from variable to variable. Typically, an ID number comes first, followed by the set of variables collected on each subject. This is referred to as a “rectangular record,” or sometimes called a “flat file.” The term comes about because each observation has exactly the same amount of information. Again, for the vast majority of studies, the length of the record is irrelevant. Data analysis programs can read very long records containing thousands of “columns” of data. Technically, each character of information consists of one *byte* of data. The maximum length of a record is around 32,000 bytes, depending on the system. This length is so generous that in most studies it is not an issue.

Hierarchical files. Although long records are not a problem for most users, large datasets may be difficult to store, even in this age of generous disk storage space. As a result, it is desirable to reduce the amount of blank space on a record. Blank space typically results when a set of variables is not applicable for the respondent. For example, consider a survey in which the interview elicits detailed information on each of the respondent’s children, with the interview protocol allowing up to 13 children. For most respondents, almost all of this information is “blank” in the sense that no information is collected, although some code to indicate “inapplicable” may appear on the record. Suppose that the average respondent has two children and that for each child 40 bytes of data are collected. On a sample size of 8,000 cases, this means that the file contains something like 3.5 *megabytes* of blanks (8,000 respondents x 11 “missing children” x 40 bytes of data).

All relevant files, particularly datasets under construction, should be backed up frequently — even more often than once a day — to prevent having to re-enter data.

In this case, one might want to consider other ways of storing the data. The first is a “hierarchical record.” In the ASCII file structure, there is a header record containing information on the number of children and a varying number of secondary records, one for each child. From the standpoint of data storage, this is very efficient, but it increases the complexity of the programming task substantially. Most major statistical packages will allow the user to read such data, but some programming is required to produce the rectangular record required for the analysis phase. Analyzing hierarchical files requires sophisticated knowledge of data analysis software. Complex files like these, while they can save lots of disk space, also require a greater level of skill on the part of the user. Unless the file is quite large, the trade-off between the simplicity of a rectangular file and space savings of hierarchical files is rarely worth it.

A second approach to this problem — the preferred approach — is to form separate files for the two kinds of records. One file would be for respondents and the second for children. This approach has the advantage of allowing a user to work with a rectangular respondent record, skipping the child records entirely if they are not of interest. On the other hand, if the children are of interest, then the secondary analyst can write merge routines to match the respondents’ and the children’s data. Therefore, the flexibility of this approach allows separate files to be merged or returned to individual files for analysis, as needed.

Backups

All relevant files, particularly datasets under construction, should be backed up frequently — even more often than once a day — to prevent having to re-enter data. Master datasets should be backed up every time they are changed in any way. Computing environments in most universities and research centers support devices for data backup and storage. Although everyone knows the importance of backing up data, the problem is that few actually follow through. It is also advisable to maintain a backup copy of the data off-site. One fire can destroy years of work.

Chapter 5 Final Project Phase: Preparing Data for Sharing

This chapter addresses the critical final steps researchers should undertake in preparing to archive and/or disseminate their data.

Respondent Confidentiality

For most of this manual, the focus has been on data preparation methods that can serve the research needs of both principal investigators and secondary analysts. In this chapter, however, we highlight one area of divergence, necessitated by the responsibility to protect respondent confidentiality. Special attention must to be paid to this issue by researchers who deposit data with a public archive. Once data are released to the public, it is impossible to monitor use to ensure that other researchers respect respondent confidentiality. Thus, it is common practice in preparing public use datasets to *alter the files* so that information that could imperil the confidentiality of research subjects is removed or masked before the dataset is made public. At the same time, however, care must be used to ensure that the alterations do not unnecessarily reduce the secondary analyst's ability to reproduce or extend the original study findings.

Below we describe some things to consider and suggest steps that can be taken by principal investigators before they submit their data for archiving. But first, a quick review of why this is important.

Two kinds of variables often found in social science datasets present problems that could endanger the confidentiality of research subjects: direct identifiers and indirect identifiers.

Disclosure risk limitation

The principles

Social scientists have a deep and genuine commitment to preserve the anonymity of the subjects whom they study in the course of their research. Most often applied to individuals who consent to be interviewed in surveys, this commitment extends also to groups, organizations, and entities whose information is recorded in administrative and other kinds of records. Archives place a high priority on preserving the confidentiality of respondent data and review all data collections they receive to ensure that confidentiality is protected in the public-use datasets released.

Two major concerns govern policy and practice in this area: professional ethics and applicable regulations. The social sciences broadly defined (as well as a number of professional associations) have promulgated codes of ethics that require social scientists to ensure the confidentiality of data collected for research purposes. (See, for example, the "Ethical Guidelines for Statistical Practice" of the American Statistical Association, which stresses the appropriate treatment of data to protect respondent confidentiality.) Both the rights of respondents and their continued willingness to voluntarily provide answers to scientific inquiries underlie this professional ethic. The ethic applies to all participants in the research enterprise, from data collectors to archivists to secondary analysts who use such data in their research.

Sets of regulations also bind all of us in the research enterprise to measures intended to protect research subjects as well as data obtained from such subjects. These regulations range from federal and local statutes to rules instituted by universities and colleges. Researchers at most universities and in other organizations are subject to such regulations that cover data that they generate or collect.

The practice of protecting confidentiality

Two kinds of variables often found in social science datasets present problems that could endanger the confidentiality of research subjects.

Direct identifiers. These are variables, which may have been collected in the process of survey administration, that point explicitly to particular individuals or units. For instance, in the United States, Social Security numbers uniquely identify individuals who are registered with the Social Security Administration. Any variable that functions as an explicit name can be a direct identifier — for example, license numbers, phone numbers, and mailing addresses. Data depositors should carefully consider the analytic role that such variables fulfill and should remove any identifiers not necessary for analysis.

Indirect identifiers. While the direct identifiers are easy to recognize, data depositors should also carefully consider a second class of problematic variables — indirect identifiers that focus attention on unique cases. Put differently, they make the unique cases visible. For instance, a United States ZIP code field may not be troublesome in the univariate case, but when combined with other attributes like race and annual income, a ZIP code may allow unique individuals (extremely wealthy, poor) residents of that ZIP code to become visible. This visibility means that answers the respondent thought would be private are no longer private. Some examples of these indirect identifiers are detailed geography (e.g., state, county, or Census tract of residence), organizations to which the respondent belongs, educational institution from which the respondent graduated (and year of graduation), exact occupations held, place where the respondent grew up, exact dates of events, detailed income, and offices or posts held by the respondent. Indirect identifiers typically include items that are useful for statistical analysis (indeed, that is probably why such information was collected in the first place). The data depositor must carefully weigh the analytic importance of these variables. Do analysts need the ZIP code, or will data aggregated to the county or state levels suffice? As we discuss in the next section, there are several strategies for working with problematic variables.

How to handle indirect identifiers. If, in the judgment of the principal investigator, a variable might act as an indirect identifier (and thus could be used to compromise the confidentiality of a research subject), the investigator should "treat" that variable when preparing a public-use dataset. Commonly used types of treatment are as follows:

> Removal — Eliminating the variable from the dataset entirely.

> Bracketing — Combining the categories of a variable.

> Top-coding — Restricting the upper range of a variable.

> Collapsing and/or combining variables — Merging the concepts embodied in two or more variables by creating a new summary variable.

> Sampling — Rather than providing all of the original data, releasing a random sample of sufficient size to yield reasonable inferences.

> Swapping — Matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases. This retains the covariate structure while retaining the analytic utility. Swapping is a service that archives may offer to limit disclosure risk. (For a more in-depth discussion of this technique, see O'Rourke, 2003.)

> Disturbing — Adding random variation or stochastic error to the variable. This retains the statistical properties between the variable and its covariates, while preventing someone from using the variable as a means for linking records.

An example from a national survey of physicians (containing many details of each doctor's practice patterns, background, and personal characteristics) may help to illustrate some of these categories of treatment of variables to protect confidentiality. Variables identifying the school from which the medical degree was obtained and the year graduated should probably be *removed* entirely, due to the ubiquity of publicly-available rosters of college and university graduates. The state of residence of the physician could be *bracketed* into a new "Region" variable (substituting more general geographic categories such as "East," "South," "Midwest," and "West"). The upper end of the range of the "physician's income" variable could be *top-coded* (e.g., "$150,000 or more") to avoid identifying the most highly-paid individuals. Finally, a series of variables documenting the responding physician's certification in several medical specialties could be *collapsed* to a summary indicator (with new categories such as "Surgery," "Pediatrics," "Internal Medicine," "Two or more specialties," etc.).

Staff at most archives will consult with principal investigators to help them design a public-use dataset that maintains (to the maximum degree possible) the confidentiality of respondents. The staff will additionally perform an independent confidentiality review of datasets submitted to the archive and will work with the investigators to resolve any remaining problems of confidentiality. The goal of this cooperative approach is to ensure that all reasonable steps have been taken to protect the privacy of research respondents whose information is contained in public-use datasets.

Restricted-use data collections

To handle the most confidential data, archives grant access through a secure data enclave environment. A data enclave is a secure data analysis laboratory that allows researchers access to the original data in a controlled setting.

Public-use data collections include content that has been carefully screened to reduce the risk of confidentiality breaches, either directly or through deductive analyses. Some original data items will be removed or adjusted through data-masking procedures. These adjustments, however, frequently impose limitations on the research uses of such files. It is possible that the loss of the confidential data could detract from the significance of the dataset and its analytic potential.

Treating a restricted dataset provides a viable alternative to removing confidentiality variables. In these cases, a public-use dataset that has these variables removed is released, and the original dataset is kept as a restricted-use dataset that preserves the original variables. The restricted-use dataset is released only to approved clients/users who have agreed in writing to abide by the rules governing the use of these restricted datasets. It is important to note, however, that designating data for restricted-use can occur at the request of a data depositor, upon determination by the archive staff following review of the data content, or after consultation between the depositor and the archive. Maintenance of, and approval of access to, a restricted-use file is managed by archive staff in accordance with the terms of access.

Access to the restricted-use files is highly controlled and only offered under a set of controlled conditions to approved researchers. The right to use these files requires acceptance of a Restricted-Use Data Agreement. This agreement will completely spell out the conditions that a researcher must accept before access can be obtained. Most agreements will require a detailed summary of the research question and will require the researcher to precisely explain why access to the confidential variables is required. Researchers are usually also given a time period during which they may use the data. At the end of the prescribed time period, researchers are usually asked to return the original files, or are asked, in good faith, to destroy the files. The restricted-use dataset approach is an effective way to permit access to confidential and/or sensitive research information and has proven acceptable to researchers. In general, the more identifying information there is in the data, the more restrictive are the regulations governing access and location of use.

Data enclaves

To handle the most confidential data, archives grant access through a secure data enclave environment. A data enclave is a secure data analysis laboratory that allows researchers access to the original data in a controlled setting. In such an environment, one might find data such as medical records with identifying information (name and address) included. The secure data enclave has added security features in place to ensure the safekeeping of the most confidential data. Enclaves typically have appropriate physical security measures (no windows, video monitoring, key card entry) to strictly control access to the data. The computing environment of the enclave is not connected to the Internet, but rather has its own network server (connected to a small number of work stations). Researchers who use the enclave are monitored by archive staff who ensure that no unauthorized materials leave the enclave. Further, any analysis that is produced is scrutinized to ensure that it does not include any potential breaches of confidentiality. There is also a set of policies and procedures that govern the use of data in the enclave.

Data with geographic identifiers or GIS-related files

In situations where data contain detailed geographic information, archive staff often opt to produce a restricted-use version of the data file. The restricted version maintains the detailed geographic information, but may be obtained only by special arrangement with the archive.

Quantitative data. Some projects collect data containing direct and indirect geographic identifiers that can be geocoded and used with a mapping application. In those cases, the geographic identifiers should be retained in the data submitted to the archive. Direct geographic identifiers are actual addresses (e.g., of an incident, a victim, an offender, etc.). Indirect geographic identifiers include state, county, Census tract, Census block, telephone area codes, and place where the respondent grew up.

> Geographic information is particularly sensitive and thus must be handled carefully.

When data contain detailed geographic information, archive staff often opt to produce a restricted-use version of the data file. The restricted version maintains the detailed geographic information but may be obtained only by special arrangement with the archive. In these situations, a public and freely available version of the data may also be distributed that retains the aggregated and/or coordinate geographic information with the detailed geographic information masked or removed (see previous section on restricted-use datasets). Investigators should contact archive staff for assistance when preparing data for submission that contain detailed geographic information.

Qualitative data. As is the case with quantitative data submissions, most archives ask that investigators submitting qualitative data review the data prior to submission to determine if the files contain information that would allow any of the subjects to be identified. In most cases, such information *should not be included* in the files sent for archiving and for public release. However, in the case of data with explicit geographic information (for example, court transcripts), depositors should discuss elimination of this information with archive staff prior to data submission. One technique that can be used is to replace geographic names with generalized text. For example, “Layton’s Corners” can be changed to “neighborhood” or to “neighborhood block,” as appropriate. References to more than one location in the text can be subscripted to identify locations — e.g., neighborhood1, neighborhood2. Demographic information can also be substituted for actual names of locations, e.g., “Layton’s Corners” can be changed to “SE/W” for Southeast, predominantly white. Pseudonyms can be used, but they may not be as informative to future users as other methods of name replacement. Note that actual names may also be other geographic locations and their acronyms or well-known and/or often-used nicknames. Investigators are asked to provide archive staff with information on what modifications were undertaken to mask confidential information in the qualitative data. This helps ensure that archive staff do not make unnecessary changes to the investigator’s modifications when performing confidentiality review. Such information will also be made available to secondary users of the data to assist them with effective use of the data.

Mapping software files. With regard to situations in which geographic locations are used as units of analysis or variables, the researcher must supply the relevant geometry files (or information on how to access them) to permit others to recreate or extend the analysis. For example, if a spatial analysis is done at the Census tract level or some special zone, those boundary files should be stored with the data so as to allow for secondary analysis. This is not necessary if the boundary file is easily obtained from the U.S. Census Bureau or from a known third party. However, if variables pertaining to specially-created zones were added, secondary analysts will need these files to perform any spatial data analysis using the same boundaries. A simple solution is to submit the geometry (boundaries) files in one compressed file containing all of the files that make up the geometry for any Geographic Information System (GIS).

Final Data File Preparation

File formats

If a dataset is to be archived, it must be organized in such a way that other people can read it. Ideally, the dataset should be accessible using a standard statistical package, such as SAS, SPSS, or Stata. Essentially, there are three broad choices: (a) provide the data in raw (ASCII) form along with documentation and let users prepare their own programs; (b) store the data in ASCII form, along with setup files to read them into a standard program; or (c) store the data as a "system file" in an analysis package. Each of these alternatives has its advantages and disadvantages.

ASCII data files and record layouts. This approach has the advantage of being readable on almost any system. However, if the researcher has been using some other system to store the data, writing an ASCII file can be time-consuming and prone to error. For example, if a research group has been using SAS to manage and analyze a dataset, then the following are required: a series of SAS statements to write the data out in ASCII format, some careful checking to make sure the conversion procedure worked properly, and record layout documentation telling users where to find variables in the data file. On the other hand, if the researcher has been maintaining the dataset in ASCII and reading it into a statistical package for analysis, a "raw" ASCII data file may be the most cost-efficient way to move the data to an archive.

ASCII data plus setup files. For this option, it is necessary to determine which syntax will be used — SAS, SPSS, Stata, or another statistical program. It is unfortunate that there is no standard format for naming variables and labeling their values that could be used in *any* statistical package. At this writing such a standard is not widely available. In the case of large datasets, for which users will want to create subsets, the setup files can be edited to meet specific needs.

Software-specific system files. System files are compact and efficient, but one should keep in mind that older system files may not be cross-platform compatible. To prepare system files, consult the user manual for the statistical package of your choice.

Portable software-specific files. Portable versions of software-specific files have the advantage that they can be accessed on any platform. SPSS calls its transportable files "portable," and SAS calls its version transport files (for Stata data files, no portable equivalent is necessary). It should be noted that when SAS transport and Stata data files are generated, missing data are blanked out (this is not true, however, if SAS alpha missing codes are used). This can be a problem because the distinctions between different types of missing data (such as legitimate skip vs. refused) become irretrievable, and they may be very important to the secondary analyst. When preparing this type of file, it is recommended that for SAS and Stata the missing data command not be activated but instead separate program files be created. There is no comparable problem in SPSS portable files because the original missing data values are maintained.

Another problem surfaces with respect to SAS proc formats (value labels), which are not stored in SAS transport data files. SAS proc formats can be provided using program files or stored in SAS catalog files, which are operating-system-specific. Again, the best approach seems to be to provide user-defined SAS proc formats and formats in separate program files.

In many ways, portable files are the most attractive alternative because the user has to do very little work to access the data. On the other hand, at least for the initial reading of the dataset, the user has to access the entire dataset to save a work file. For small datasets this is not a problem, but for large ones it can be problematic, particularly if the user is working on a PC with limited disk capacity.

Choosing among these various file formats is a difficult task. The advantage of having raw data plus setup files to read the data is that the user can customize the syntax (switching from SAS to SPSS, or vice versa, is not that difficult), and read in only those variables of interest. At present, many archives view ASCII (raw) data files as the most "stable" format to archive. Since they are software-independent, they may have a better chance of being "read" in the future, regardless of what happens to individual statistical packages.

Online analysis-ready files. Recently, a number of online data exploration and analysis packages have been developed. These programs have the advantage of allowing users not only to perform analysis online, but also to select only those variables and cases actually required for an analysis in the form of subsets. Increasingly, these systems accept DDI XML as input. Thus, creating documentation in DDI facilitates online analysis after archival deposit.

Alternative data types

Social science research is generating many new types of data files, e.g., qualitative, coordinate-based geographic data, video/audio, and each requires special handling in terms of documentation and disclosure risk analysis. If providing data in any of these special formats is unusually difficult, the data depositor is encouraged to contact the archive to discuss an alternative set of specifications that might be mutually satisfactory.

Archiving data from secondary sources

For projects that do not involve original data collection and/or may involve combining data from one or more secondary sources, the decision regarding whether or what to archive is less clear. This decision should be made in conjunction with an archive, as archives can differ in their acquisition policies. Here are some guidelines to consider:

Primary data not publicly available. If the primary data used for secondary analysis are not already publicly available, researchers are encouraged to submit the data for archiving, with the permission of the original data producer.

Combination of primary and secondary data. If the researcher collects some primary data and also appends secondary data to it for analyses, the guiding questions on whether to submit the whole dataset or just the primary dataset to the archive are: (a) how easily the secondary data can be linked to the primary data, and (b) whether the secondary data are publicly available.

Straightforward links/publicly-available secondary data. If the linkage is straightforward and the secondary data are publicly available, then users can easily obtain the secondary data themselves and link them to the primary data submitted by the researcher. In this case, the project report(s) should clearly identify the source of the secondary data including version and/or date so that other users know which data to obtain. Information about the variable (or combination of variables) that constitutes the unique identifier used to link the data also should be provided.

Census data as secondary data. When the primary data are linked to Census data, the linked Census data should be archived as well, even though the link between the data files can be quite straightforward and the Census data are publicly available. This is mostly due to the fact that the original Census files are large in size and contain a large number of variables. Determining which Census variables to use and at what level to extract the data for the subsets can be time-consuming. Archiving the linked Census data makes it unnecessary for other users to repeat these subsetting steps.

Links that are not straightforward. If the linkage is not straightforward, then the researcher is providing a useful service by archiving the linked data. Examples that fit this type of situation include: (a) the link needs to be made using a judgment call of a combination of nonunique variables, for example, age, sex, and race of an individual and date of incident; (b) an understanding of local geographic factors is needed to link correctly, for example, neighborhoods or block levels, especially over a range of years when boundaries shift. Here, the redundancy of having data stored twice at the archive is outweighed by the usefulness of providing others with the data already linked.

Nonpublicly-available secondary data. If the secondary data are not publicly available, the investigator is encouraged to archive the secondary data as well, especially if these data are needed to replicate the original findings. Sometimes this is not an option, as in instances when the secondary data were obtained informally from individuals who did not want their data publicly available through these means or when the data were obtained from other government sources and the researcher does not have approval to publicly archive the data. But, if the data are not yet publicly available and can be archived, researchers are encouraged to do so.

Variables derived from secondary data. Sometimes, after the data are linked, the researcher may compute new variables based on the linked data (new categories are created, rates are produced, scales are developed). All useful derived variables should be archived also, especially if they are cited in publications. Depending on the factors mentioned above, the derived variables may be archived with the primary data alone or with the linked data.

Secondary data used solely. If the project only involves secondary analysis of data already publicly available and the product of the project is the analysis alone, then data may not need to be submitted.

However, sometimes during secondary analysis of data that are already publicly available, variables are created that would be useful to add to the archive. In this type of situation, the original data and derived variables may be submitted or the derived variables may be submitted alone, as long as the linkage to the original data is clear or well-documented. Determining the usefulness of submitting derived variables for archiving in these instances is at the discretion of the depositor.

Overall, the decision on whether to archive the secondary data and/or the derived variables needs to be made on a case-by-case basis. If the decision on whether to submit data for a project is unclear, the principal investigator should contact archive staff for assistance.

References

American Statistical Association. *Ethical Guidelines for Statistical Practice*. Prepared by the Committee on Professional Ethics, Approved by the Board of Directors, August 7, 1999. http://www.amstat.org/profession/index.cfm?fuseaction=ethicalstatistics

Babbie, Earl. *Survey Research Methods*. 2nd ed. Belmont, CA: Wadsworth, 1990, 209-211.

Blank, Grant, and Karsten Boye Rasmussen. "The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." *Social Science Computer Review* 22 (2004) : 307-318. Retrieved November 15, 2004, from http://ssc.sagepub.com/cgi/content/abstract/22/3/307

Bourque, Linda B., and Virginia A. Clark. *Data Processing: The Survey Example*. Newbury Park, CA: SAGE Publications, Inc., 1992.

Fienberg, Stephen E. "Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions." In *Annual Review of Public Health*, 15 (1994). Palo Alto, CA: Annual Reviews, Inc.

Geda, Carolyn L. *Data Preparation Manual*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, January 1980.

Inter-university Consortium for Political and Social Research (ICPSR). Human Subject Protection and Disclosure Risk Analysis. http://www.icpsr.umich.edu/HSP/

Inter-university Consortium for Political and Social Research (ICPSR). http://www.icpsr.umich.edu

Jacobs, James A., and Charles Humphrey. "Preserving research data." *Communications of the ACM*. 47, 9 (2004): 27–29.

Little, Roderick J.A., and Nathaniel Schenker. "Missing Data." In Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York, NY: Plenum Press, 1995.

Marz, Kaye, and Christopher S. Dunn. *Depositing Data With the Data Resources Program of the National Institute of Justice: A Handbook*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, August 2000.

National Institutes of Health ((NIH), Office of Extramural Research. "NIH Data Sharing Policy and Implementation Guidance" (Updated: March 5, 2003). http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

O'Rourke, JoAnne McFarland. "Disclosure Analysis at ICPSR." *ICPSR Bulletin*, Vol. XXIV, No. 1 (Fall 2003), 3–9. http://www.icpsr.umich.edu/org/publications/bulletin/fall03.pdf

SAS 9.1 Language Reference: Concepts. SAS Publishing. April 2004. ISBN: 1-59047-198-9. http://www.sas.com

SAS 9.1 Language Reference: Dictionary, Volumes 1, 2, and 3. SAS Publishing. April 2004. ISBN: 1-59047-199-7. http://www.sas.com

SPSS 12.0 Command Syntax Reference. SPSS Inc. 2003. http://www.spss.com

Stata Base Reference Manual (4 volumes). Stata Press. 2003. ISBN: 1-881228-69-X Volumes 1–4. http://www.stata.com

University of Michigan, Institutional Review Board (IRB), Behavioral Sciences. http://www.irb.research.umich.edu/IRB_HSBS_Shared/consent.html

Zelenock, Tom, and Kaye Marz. “Archiving Social Science Data: A Collaborative Process.” *ICPSR Bulletin*, Vol. XVII, No. 4 (May 1997): 1–4. http://www.icpsr.umich.edu/org/publications/bulletin/may97.pdf