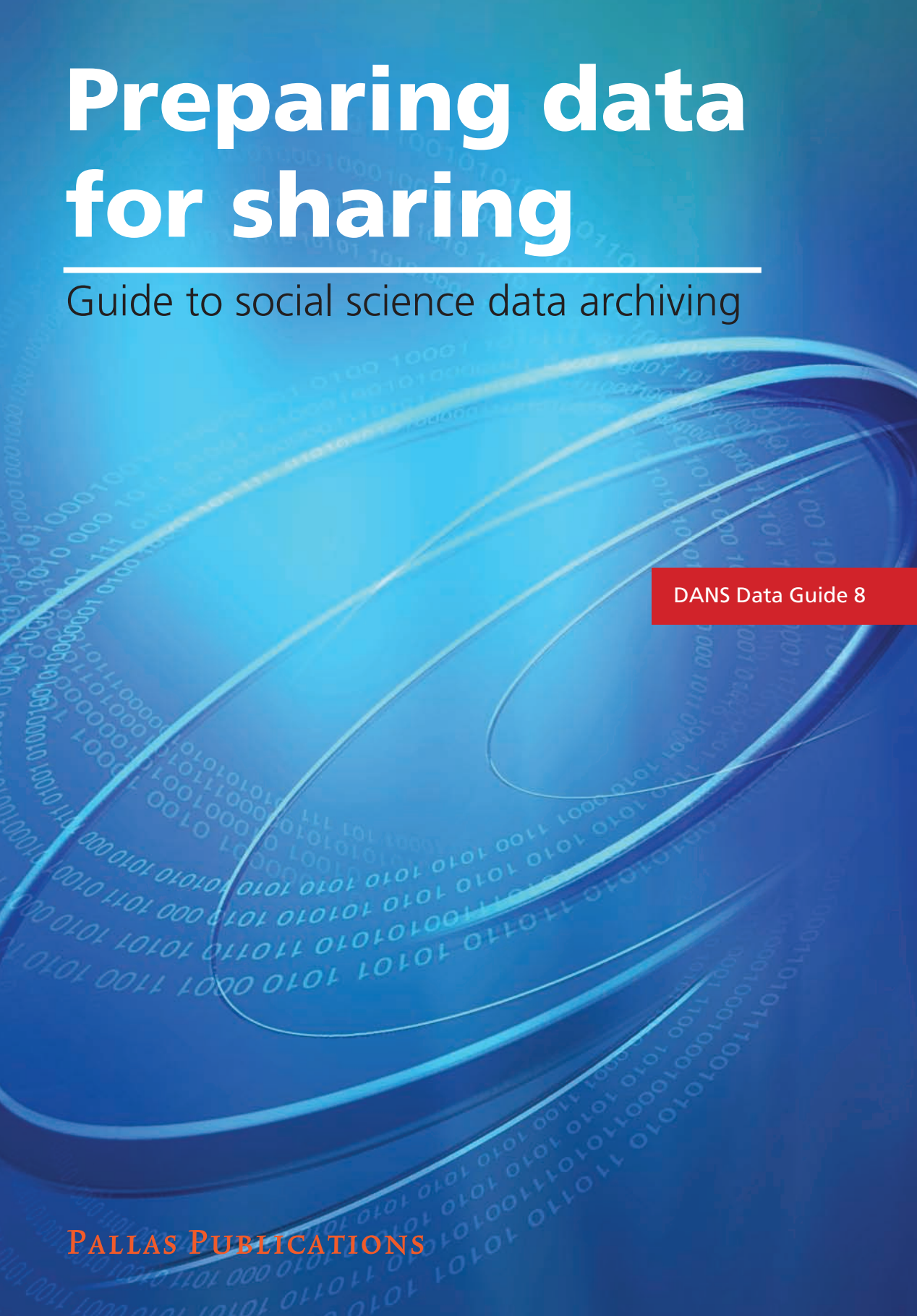
Preparing data for sharing

Guide to social science data archiving

DANS Data Guide 8



Pallas Publications

Preparing data for sharing

Preparing data for sharing

Guide to social science data archiving

DANS Data Guide 8

Pallas Publications

2010 Inter-university Consortium for Political and Social Research ICPSR

© Some rights reserved.

Usage and distribution of this work is defined in the Creative Commons Attribution-Non-Commercial-Share Alike 3.0 Netherlands License. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0/deed.nl.

ICPSR University of Michigan

Institute for Social Research
P.O. Box 1248
Ann Arbor
MI 48106 USA
www.icpsr.umich.edu

Holder Copyright License Agreement

Data Archiving and Networked Services – DANS
Royal Netherlands Academy of Arts and Sciences &
Netherlands Organisation for Scientific Research
P.O. Box 93067
2509 AB The Hague
The Netherlands
www.dans.knaw.

ISBN 978 90 8555 039 6
e-ISBN 978 90 4851 380 2
NUR 740 / 995

Typesetting and design: Ellen Bouma, Alkmaar
Photo cover: istockphoto.com

Pallas Publications is an imprint of Amsterdam University Press

Preface

This publication 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 the DANS data archive, or any other data repository. It is a compilation of best practice gleaned from the experience of data archivists and investigators.

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. This publication focuses in large part, but not entirely, on quantitative data and 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 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. Additionally, it addresses a variety of dataset formats: quantitative (survey, administrative), qualitative, and geospatial. The publication is intended to help researchers document their datasets and prepare them for archival deposit, as well as to think more broadly about the types of digital content that should be deposited in an archive.

This publication is an adaption of the 4th edition of the Guide to social science data preparation and archiving of 2009, published by the Inter-university Consortium for Political and Social Research – ICPSR at the University of Michigan in the United States. We like to thank Mary Vardigan and Ruth Shamraj, staff members of ICPSR, for their very generous permission to rewrite the original ICPSR publication for the social science research community in the Netherlands. As so far, in the Netherlands there is no any other publication available about preparing social science data for archiving with the ultimate goal to share the data with other researchers. To emphasize the focus of this edition for the Netherlands we gave it the title Preparing data for sharing, guide to social science data archiving.

We sincerely hope that this publication will encourage researchers and research institutions to contact DANS at any point in the research process to discuss their plans with respect to the preparation of public-use datasets.

Dr. Peter K. Doorn Director DANS
The Hague, September 2010

Contents

1. Introduction

1.1 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 and documentation for others to use.

Maintaining scientific standards is an endeavor involving the entire scientific community, teachers, students, authors, reviewers, funding agencies, journal and book editors, and so on. Data sharing plays an essential role in this process, allowing scientists to test and replicate each others' findings. 'The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author' (King, 1995). There are many benefits to data sharing that go beyond replication. Fienberg (1994) argues that data sharing:

- 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. 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.

Early archiving may enable a researcher to enhance the impact, and certainly the visibility of a project.

1.2 Best practice in archiving

In order for data to be shared and to benefit the scientific enterprise, they must be properly curated and stored. The DANS-developed Data Seal of Approval (DSA) specifies guidelines to ensure that data can be found and shared into the future. These quality guidelines are of interest to researchers and institutions that create digital research files, to organizations that archive research files, and to users of research data. The Data Seal of Approval guidelines can be summarized in the following way.

Digital research data must meet the following quality criteria:

- The research data can be found on the Internet
- The research data are accessible, while taking into account ruling legislation with regard to personal information and intellectual property of the data
- The research data are available in a usable data format
- The research data are reliable
- The research data can be cited

These are important tenets that the social science research community must bear in mind.

For additional information on the Data Seal of Approval, see also chapter 7 and the DSA website: http://www.datasealofapproval.org

1.3 Planning ahead for archiving and preservation of data

Data sharing plans should be developed in conjunction with the data archive to maximize the utility of the data to research and to ensure the availability of the data in the future. We recommend that researchers determine the file formats they will use in consultation with the data archive; this will facilitate preservation and dissemination of the most complete data and documentation files. DANS is committed to maintain social science research data for the long term, for the benefit of future researchers, and to assist data creators in meeting the stipulations of their grantors, e.g. NWO, the Netherlands Organisation for Scientific Research. To ensure the preservation of your data, preservation concerns should be built into the process of proposal writing, data collection, analysis, and eventual deposit.

Ideally, the researcher should plan for eventual archiving and dissemination of project data before the data even come into existence. 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.

2. Proposal planning and writing phase

As indicated earlier, funding institutions increasingly require that applications for support include data sharing and dissemination plans. Plans for deposit and long-term preservation should be fleshed out while the researcher is at the stage of outlining and writing the grant application. Planning 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.

2.1 Important steps to follow

We offer the following steps as guidance during the proposal planning and writing stage:

2.1.1 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 data archives and repositories be reviewed.

2.1.2 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 as one looks 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. Increasingly, social science research data include audio and video formats, geospatial data, biomedical data, and websites, and DANS as well as other data repositories are interested in capturing this broadening array of data.

Some projects may combine data from one or more secondary sources. Derived datasets created by collating materials from existing sources and presenting them 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 are 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. Researchers depositing data from their project that were obtained from another source should identify the original source as well as the researcher's right to distribute the data.

2.1.3 Describe any special challenges that might arise when archiving the data

If one envisions any difficulties in making the data available for secondary analysis, these difficulties should be outlined in the grant application. Difficulties associated with depositing or archiving materials usually center around one of the following issues: informed consent, confidentiality, or copyright.

Difficulties in making the data available for secondary analysis can also arise if the data are deposited with inadequate documentation. Without metadata or a proper description of the data, it could be difficult for others to understand or reuse the data. By including a plan for project documentation and metadata creation in the project proposal, the data creator can limit costs and avoid the challenge of creating metadata after the fact.

Any problems applicants foresee regarding the archiving of data should be explicitly spelled out during proposal preparation to ensure smooth transfer of information to the data archive. Strategies should be developed to overcome these potential problems. 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 the archive shares this perspective.

2.1.4 Craft informed consent language

It is never too early to think about issues related to informed consent agreements 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 agreement that provides required information about the research study. The informed consent agreement serves as the formal agreement by an individual 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

agreement, investigators must include a statement describing the extent to which confidentiality of records identifying the subject will be maintained. This may limit an investigator's discretion to share data with the research community.

2.1.5 Language to avoid in informed consent agreement

If an investigator is planning to archive and share data with the research community, or might consider this in the future, it is best to avoid making inaccurate and overly restrictive promises in the informed consent agreement. For example, language that states that the data will be shared with the research team exclusively or promises that the data will be shared only in aggregate form or in statistical tables will make it more difficult to share the data later. Data producers and archivists have many effective means of modifying datasets to retain the value of human subjects data without disclosing respondent identity. This involves slightly altering the data to blur associations that may present risk while preserving key analytic utility. When obtaining informed consent, investigators should notify respondents that they are obligated as scientists to ensure that information provided will remain confidential when data are shared or made available for secondary analysis.

2.1.6 Model text about confidentiality for informed consent forms

One of the essential components in an informed consent form is a statement about the confidentiality of information collected from subjects. The privacy of subjects must be protected, and consent agreements should affirm that confidentiality will not be compromised by future data analysis. As indicated in the section above, researchers should avoid wording that will preclude archiving and sharing data.

2.1.7 Determine copyright and ownership of research data

Issues surrounding the ownership of research data are often complex. Therefore it is strongly recommend to determine the copyright-holder in the very beginning of a research project.

2.1.8 Describe the potential users of the dataset

The grant proposal should specify who the likely users (academic or nonacademic) 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.

2.1.9 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. Ideally, this should be planned in conjunction with an archive. The various activities typically associated with preparing data are presented below, for which grant applicants should include 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.
- Grant applications should allocate sufficient time and money for the preparation of high- quality documentation. Good documentation is invaluable for both preservation and informed re-use of data.
- Informed consent and confidentiality issues impact costs for archiving. For clarity, informed consent agreement 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 set period of time be dedicated to preparing and collating materials for deposit. This normally comprises the majority of the costs for archiving.

2.2 Consider alternative archiving options

2.2.1 Self-dissemination

Despite the clear advantages that dissemination of data through DANS 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.

2.2.2 Preservation with delayed dissemination

Another possible option, which should be specified in the application for funding, is preservation with delayed dissemination. Under such an agreement the archive and the data producer arrange for archival preservation of the data with dissemination to occur at a later date. With delayed dissemination, the deposit may be completed when it is easiest for the depositor and the archive to manage the data,

as opposed to delaying preservation activities until the time has come to disseminate the data. 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. Archive staff will need to develop expertise with the data and possibly perform processing while knowledgeable project staff are available for assistance.

2.2.3 Data repositories

Many data repositories are found at academic institutions, and have goals of preserving and making available some portion of the academic work of their students, faculty, and staff. Such repositories also provide certain benefits to the researcher, which should be examined and weighed when decisions are made about data preservation and access. DANS recommends researchers who intend to archive their data at those repositories to check if the data repository meets (some of) the guidelines of the Data Seal of Approval.

2.2.4 Restricted-use collections

As previously mentioned, the issue of confidentiality is of paramount importance when conducting research that involves human subjects. Before submitting data to DANS, data depositors are asked to mask or blank information that could be used to identify research participants. 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-access form. DANS 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. As officially agreed with the depositor of the data, DANS provides access to restricted datasets only after consultation and permission of the copyright-holder of the data.

2.3. Other considerations

Data can often include forms of emerging digital content, which can impact the cost of preservation and dissemination. These forms of data should be collected and documented with well-formed metadata, in an approach similar to the management of quantitative data. If the data are qualitative (e.g., verbatim interviews)

rather than quantitative, a mix of qualitative and quantitative, or video, audio, or coordinate-based geographic data, special concerns come into play and must be addressed in the grant application.

3. Project start-up and data management phase

3.1 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, including metadata, which will ultimately be deposited in the DANS data archive or any other data repository. Planning for the management and archiving of a data collection 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.

3.2 Initial questions to consider

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

3.2.1. Data and 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 large data record or several shorter ones?

3.2.2 Naming conventions

How will files and variables be named? What naming conventions will be used to achieve consistency?

3.2.3 Data integrity

With regard to the input of data, will the variable formats be numeric or character? What checks will be used to find invalid values, inconsistent responses, incomplete records, etc.? What checks will be used to manage the data versions?

3.2.4 Preparing dataset documentation

What will the dataset documentation or metadata look like and how will it be produced? How much is necessary for future retrieval and archival processing?

3.2.5 Variable construction

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

3.2.6 Project documentation

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

3.3 Software

To what extent can the various tasks mentioned above be integrated into a single process? Using a single computer program or an integrated set of programs to carry out these tasks simplifies data management, reduces costs, and is more reliable. It is advisable to determine which program or programs will handle data management and documentation tasks at the outset of the project.

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. A number of software tools are 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 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.

Computer-assisted interviewing (CATI/CAPI) is increasingly being used 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. Be aware that not all CAPI-generated variables are needed in the data file that is deposited; variables that are artifacts of the CAPI process do not contribute useful information for analysis. If 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.

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 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.

3.4 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 the initial phase. First, 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. The cost of the most expensive aspect, data entry, may be reduced, since the pretest covers only a small number of cases. The investigator may not want to prepare a comprehensive 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.

4. Data collection and file creation phase

According to the Data Seal of Approval, there is a complementary relationship between the data producer's responsibility for the quality of his/her research data and the capability of the data archive to provide access and preservation for the long term (DANS, 2009). Following best practice in terms of building both the data and documentation components of a collection is critical. This section describes aspects of best practices in creating research data that conform to widely accepted norms for quantitative, GIS, qualitative, and other types of data in the social sciences.

4.1 Quantitative Data

4.1.1 Dataset creation and integrity

Transcribing data from a questionnaire or interview schedule to an actual data record can introduce several types of errors, including typing errors, codes that do not make sense, and records that do not match. For this reason, employing a data collection strategy that captures data directly during the interview process is recommended. Consistency checks can then be integrated into the data-collection process through the use of CATI/CAPI software in order to correct problems during an interview.

However, even if data are being transcribed (either from survey forms or published tables), several steps can be taken in advance to lessen the incidence of errors.

- Separate the coding and data-entry tasks as much as possible. Coding should be performed in such a way that distractions to coding tasks are minimized.
- 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.
- Use a data-entry program that is designed to catch typing errors, i.e., one that is pre- programmed to detect out-of-range values.
- Perform double entry of the data, 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.

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

Despite best efforts, errors will undoubtedly occur regardless of data-collection mode. Here is a list of things to check.

WILD CODES AND OUT-OF-RANGE VALUES

Frequency distributions and data plots will usually reveal this kind of problem, although not every error is as obvious as, for example, a respondent with 99 rather than 9 children. Sometimes frequency distributions will contain apparently valid values but might be incorrect. For example, the columns for a given variable might have been defined incorrectly, and thus the data have been read from the wrong columns. Data plots often instantly reveal outlying observations that merit checking.

CONSISTENCY CHECKS

Checks for consistency require substantive knowledge of the study. Typically, they involve comparisons across variables. Checks can reveal inconsistencies between responses to gate or filter questions and subsequent responses. Other consistency checks involve complex relationships among variables, e.g., unlikely combinations of respondents' and children's ages. At a minimum, researchers should assure that fields that are applicable to a respondent contain valid values, while those that are not applicable contain only missing values. Measures to prevent inconsistencies should be undertaken even before any data are collected. As previously mentioned, implementing a data-collection system that captures data during the interview process and that can correct problems during the interview (such as use of CATI/CAPI software) can eliminate transcription errors that can occur during post-survey data entry. The data-collection instrument should also be tested before data collection begins to ensure that data will be captured correctly, and that any skip patterns are accurately followed. However, these measures do not eliminate the need by the researcher to examine the relationships among variables to ensure consistency.

RECORD MATCHES AND COUNTS

In some studies, each subject or study participant might have more than one record. This occurs most frequently in longitudinal studies in which each subject has one record for each occasion during which s/he is observed even though s/

he was not actually interviewed at a given point in time. In other instances, the number of additional records may actually vary from subject to subject. For example, in a study of families one might have a household record, followed by a varying number of person records. This is sometimes known as a hierarchical file. Here the researcher must make certain, to the extent that software permits, that (a) the header record contains a count of the number of trailer records, (b) consistency checks are made on the counts, and (c) unique identifiers link the header file to the corresponding trailer records.

4.1.2 Variable names

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. When selecting a variable name, choose a name that is consistent in length with the requirements of the software package being used and consider the long-term utility of the variable name to the widest audience of users. Several systems for constructing variable names are as follows:

- *One-up numbers*
  This system numbers variables from 1 through n (the total number of variables). Since most statistical software does not permit variable names starting with a digit, the usual format is V1 (or V0001) ...Vn. This has the advantage of simplicity, but provides no indication of the variable content. Although most software allows extended labels for variables (allowing entry of descriptive information, e.g., V0023 is 'Q6b, Mother's Education'), the one-up system is prone to error.

- *Question numbers*
  Variable names also may correspond to question numbers, e.g., Q1, Q2a, Q2b... Qn. This approach relates variable names directly to the original questionnaire, but, like one-up numbers, such names are not easily remembered. Further, a single question often yields several distinct variables with letters or numbers (e.g., Q12a, Q12a1), which may not exist on the questionnaire.

- *Mnemonic names*
  Short variable names that represent the substantive meaning of variables have some advantages, in that they are recognizable and memorable. They can have drawbacks, however. What might be an 'obvious' abbreviation to the person who created it might not be understood by a new user. Software sometimes

limits the number of characters, so it can be difficult to create immediately recognizable names.

- *Prefix, root, suffix systems*
  A more systematic approach involves constructing variable names containing a root, a prefix, and possibly a suffix. For example, all variables having to do with education might have the root ED. Mother's education might then be MOED, father's education FAED, and so on. Suffixes often indicate the wave of data in longitudinal studies, the form of a question, or other such information. Implementing a prefix, root, suffix system requires prior planning to establish a list of standard two- or three-letter abbreviations.

4.1.3 Variable labels

Most statistical programs permit the user to link extended labels for each variable to the variable name. Variable 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 the item number is part of the variable name), (2) a clear indication of the variable's content, and (3) an indication of whether the variable is constructed from other items. If the number of characters available for labels is limited, one should develop a set of standard abbreviations in advance and present it as part of the documentation for the dataset.

4.1.4 Variable groups

Grouping substantively related variables together and presenting such lists in the codebook for a study can effectively organize a dataset and enable secondary analysts to get an overview of a dataset quickly. Groups are especially recommended if a dataset contains a large number of variables. They are especially useful for data made available through an online analysis system as they offer a navigational structure for exploring the dataset.

4.1.5 Codes and coding

Before survey data are analyzed, the interview or questionnaire responses must be represented by numeric codes (Babbie, 1990). Common coding conventions assure that all statistical software packages will be able to handle the data, and promote greater measurement comparability. Computer-assisted interviewing systems assign codes automatically by programming them into the instrument, so that most coding decisions are made before the instrument is fielded. The principles discussed here apply to such situations as well as those in which coding follows data collection.

Guidelines to keep in mind while coding:

- *Identification variables*
  Provide fields at the beginning of each record to accommodate all identification variables. Identification variables often include a unique study number and a respondent number to represent each case.

- *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*
  Code as much detail as possible. Recording original data, such as age and income, is more useful than collapsing or bracketing the information. With original or detailed data secondary analysts can determine other meaningful brackets on their own rather than being restricted to those chosen by others.

- *Closed-ended questions*
  Responses to survey 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*
  For open-ended items, investigators can either use a predetermined coding scheme or review the initial survey responses to construct a coding scheme based on major categories that emerge. Any coding scheme and its derivation should be reported in study documentation. Increasingly, investigators submit the full verbatim text of responses to open-ended questions to archives so that users can code these responses themselves. Because such responses may contain sensitive information, they must be reviewed for disclosure risk and if necessary treated by archives prior to dissemination.

- *Check-coding*
  It is a good idea to verify or check-code some cases during the coding process, that is, repeat the process with an independent coder. For example, if more than one code is assigned to an interview response, this highlights problems or ambiguities in the coding scheme. Such check-coding provides an important means of quality control in the coding process.

- *Series of responses*
  If a series of responses requires more than one field, organizing the responses into meaningful major classifications is helpful. Responses within each major category are assigned the same first digit. Secondary digits can distinguish specific responses within the major categories. Such a coding scheme permits analysis of the data using broad groupings or more detailed categories.

4.1.6 Missing data

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

- *Refusal/No answer*
  The subject explicitly refused to answer a question or did not answer it when he or she should have.

- *Don't know*
  The subject was unable to answer a 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 for the previous year).

- *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.

- *Not applicable*
  The subject was never asked a question for some reason. Sometimes this results from skip patterns following filter questions, for example, subjects who are not working are not asked about job characteristics. Other examples of inapplicability are sets of items asked only of random subsamples and those asked of one member of a household but not another.

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

- *No data available*
  The question should have been asked of the respondent, but for a reason other than those listed above, no answer was given or recorded.

Effective methods for missing data imputation and missing data analysis rely on accurate identification of missing data. For more information on best practice in handling missing data, see Little et al., 2002 and McNight et al., 2007.

4.1.7 Selecting missing data codes

Missing data codes should match the content of the field. If the field is numeric, the codes should be numeric, and if the field is alphanumeric, the codes may be numeric or alphanumeric. Most researchers use codes for missing data that are above the maximum valid value for the variable (e.g., 97, 98, 99). This occasionally presents problems, most typically when the valid values are single-digit values but two digits are required to accommodate all necessary missing data codes. Similar problems sometimes arise if negative numbers are used for missing data (e.g., -1 or -9), because codes must accommodate the minus sign. Missing data codes should be standardized such that the same code is used for each type of missing data for all variables in a data file, or across the entire collection if the study consists of multiple data files.

In general, blanks should not be used as missing data codes unless there is no need to differentiate types of missing data such as 'Don't know,' 'Refused,' etc. Blanks are acceptable when a case is missing a large number of variables (e.g., when a follow-up interview in a longitudinal study was not conducted), or when an entire sequence of variables is missing due to inapplicability, such as data on non-existent children. In such instances, an indicator variable should allow analysts to determine unambiguously when cases should have blanks in particular areas of the data record.

4.1.8 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. 'Not Applicable' or 'Inapplicable' codes, as noted above, should be distinct from other missing data codes. Dataset documentation should clearly show for every item exactly who was or was not asked the question. At the data-cleaning stage, all 'filter items' should be checked against items that follow to make sure that the coded answers do not contradict one another, and that unanswered items have the correct missing data codes.

4.1.9 Imputed data

If missing data have been imputed in any way, this should be indicated. There are two standard ways of doing so. One approach is to include two versions of any imputed variables: the original variable, including missing data codes, and the imputed version that contains complete data. Another 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 to 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 impute for 'Inapplicable' cases where the data are missing because the respondent is not working.)

4.1.10 Geographic identifiers and geospatial data

Some projects collect data containing direct and indirect geographic identifiers that can be geocoded and used with a mapping application. Direct geographic identifiers are actual addresses (e.g., of an incident, a business, a public agency, etc.). Indirect geographic identifiers include location information such as province, municipality, telephone area codes, and place where the respondent grew up.

Investigators are encouraged to add to the dataset-derived variables that aggregate their data to a spatial level that can provide greater subject anonymity. It is desirable for data producers to geocode address data to coordinate data as they can often produce better geocoding rates with their knowledge of the geographic area. When data producers convert addresses to geospatial coordinates, the data can later be aggregated to a higher level that protects respondent anonymity.

In such instances, the original geographic identifiers should be saved to a separate data file that also contains a variable to link to the research data. The file with the direct identifiers should be password-protected and both data files should be submitted to the archive in separate submissions. Investigators are encouraged to contact archive staff for assistance when preparing data for submission that contain detailed geographic information.

When data contain geographic information that pose confidentiality concerns, archive staff can produce a restricted-use version of the data file. The restricted-use version maintains the detailed geographic information and the data can be obtained only through a restricted data use agreement with the archive. In these situations, a publicly available (i.e., downloadable) version of the data may also be distributed that retains the aggregated geographic information but with detailed geographic information masked or removed.

When coordinate-based geographic data are used as units of analysis or variables, the researcher must submit to the archive the relevant geometry files (or information on how to access them) to permit others to recreate or extend the original analysis using the same boundaries. Generally, depositors can submit the geometry (boundary) files in one compressed file containing all of the files that produce the geometry (e.g., single geographic layer visualization, map visualization) for any geographic information system (GIS). Corresponding project files, geospatial metadata, and geocoding rates should also be submitted. Finally, depositors should assure that issues of proprietary visualizations and/or data have been addressed prior to archiving with the understanding that all archived data will be available for distribution.

4.2 Qualitative Data

With proper and complete documentation, archived qualitative data can provide a rich source of research material to be reanalyzed, reworked, and compared to other data. ESDS Qualidata, a qualitative data archive in the United Kingdom, suggests five possible reuses of qualitative data (2007):

- Comparative research: replication or restudy of original research, comparing with other data sources or providing comparison over time or between social groups or regions, etc.
- Re-analysis: asking new questions of the data and making different interpretations than the original researcher made. Approaching the data in ways that were not originally addressed, such as using data for investigating different themes or topics of study.
- Research design and methodological advancement: designing a new study or developing a methodology or research tool by studying sampling methods, data collection, and fieldwork strategies.
- Description: describing the contemporary and historical attributes, attitudes and behavior of individuals, societies, groups or organizations.
- Teaching and learning: providing unique materials for teaching and learning research methods.

4.2.1 Types of qualitative data

Examples of types of qualitative data that may be archived for secondary analysis include:

- In-depth/unstructured interviews, including video
- Semi-structured interviews

- Structured interview questionnaires containing substantial open comments
- Focus groups
- Unstructured or semi-structured diaries
- Observation field notes/technical fieldwork notes
- Case study notes
- Minutes of meetings
- Press clippings

This is only a partial list and is not meant to be exhaustive. Concerns about what can be submitted for deposit should be discussed with archive staff.

4.2.2 Confidentiality in qualitative data

Ideally, prior to submitting qualitative data to an archive, data depositors should take care to remove information that would allow any of their research subjects to be identified. This process can be made less arduous by creating an anonymization scheme prior to data collection and anonymizing the data as the qualitative files are created for the analysis. The following are examples of modifications that can be made to qualitative data to ensure respondent confidentiality (Marz and Dunn, 2000):

- *Replace actual names with generalized text*
  For example, 'John' can be changed to 'uncle' or 'Mrs. Briggs' to 'teacher.' More than one person with the same relationship to the respondent can be subscripted to represent each unique individual, e.g., friend1, friend2. Demographic information can also be substituted for actual names of individuals, e.g., 'John' can be changed to 'M/W/20' for male, white, 20 years old. Pseudonyms can be used; however, they may not be as informative to future users as other methods of name replacement. Note that actual names may also be store names, names of juvenile facilities, transportation systems, program names, neighborhood names, or other geographic location and their acronyms or well-known and/or often used nicknames.

- *Replace dates*
  Dates referring to specific events, especially birthdates, should be replaced with some general marker for the information, e.g., month, month/year, or mm/dd/yy.

- *Remove unique and/or publicized items*
  If the item cannot be generalized using one of the above options, the entire text may need to be removed and explicitly marked as such, e.g., using either 'description of event removed'.

Since investigators are most familiar with their data, they are asked to use their judgment on whether certain qualitative information in combination with the rest of the text or related quantitative information could allow an individual to be identified.

Data depositors should document any modifications to mask confidential information in the qualitative data. This will ensure that archive staff do not make unnecessary changes to the investigator's modifications when performing their confidentiality review. Such information will also be made available to secondary users of the data to assist them with their use of the data.

4.2.3 Documentation for qualitative data

In order for qualitative data to be used in secondary analysis, it is extremely important that the data are well-documented. Any information that could provide context and clarity to a secondary user should be provided. Specifically, documentation for qualitative data should include:

- Research methods and practices (including the informed consent process) that are fully
- documented
- Blank copy of informed consent form with IRB approval number
- Details on setting of interviews
- Details on selection of interview subjects
- Instructions given to interviewers
- Data collection instruments such as interview questionnaires
- Steps taken to remove direct identifiers in the data (e.g., name, address, etc.)
- Any problems that arose during the selection and/or interview process and how they were
- handled
- Interview roster

The purpose of the interview roster is twofold. First, it provides archive staff a means of checking the completeness and accuracy of the data collection provided for archiving. Second, the interview roster provides a summary listing of available interviews to a secondary user to allow for a more focused review of the data.

Additional information about the DANS requirements for depositing and giving access to specific qualitative data projects can be found in the publication: *Wegwijzer Digitaal Data Deponeren, Interview Data, Getuigen Verhalen* (DANS, 2008).

4.3 Other data types

Social science research is generating new types of data files, such as video and audio. Each data type 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 producer is encouraged to contact the archive to discuss an alternative set of specifications that might be mutually satisfactory. Data archives are developing guidance to assist data depositors in handling these forms of emerging digital content.

4.4 The codebook and the coding instrument

Technical documentation for social science data, often called the codebook, provides communication between the producer of a dataset and the data user, conveying information that is necessary to fully exploit the analytic potential of the data.

A list of important documentation items for an analyst of secondary data is presented below.

Principal investigator(s)
Principal investigator name(s), and affiliation(s) at time of data collection.

Title
Official title of the data collection.

Funding sources
This should include 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
This should describe the project and its intellectual goals and indicate how the data articulate with related datasets. Publications providing essential information about the project should be cited. A brief project history detailing any major difficulties faced or decisions made in the course of the project is useful.

Sample and sampling procedures

This section should describe the target population investigated and the methods used to sample it (assuming the entire population is not studied). 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 were created, 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 presented, indicating the proportion of those sampled who actually participated in the study. For longitudinal studies, the retention rate across studies should also be noted.

Weighting

If weights are required, information on weight variables, how they were constructed, and how they should be used should be presented.

Date and geographic location of data collection, and time period covered [Dublin Core Coverage]

These coverage elements are critical to good documentation.

Data source(s)

If a dataset draws on resources other than surveys, documentation should indicate the original sources or documents from which data were obtained.

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:

- The exact question wording or the exact meaning of the datum. Sources should be cited for questions drawn from previous surveys or published work.
- The text of the question integrated into the variable text. If this 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.
- Universe information, i.e., who was actually asked the question. Documentation should indicate exactly who was asked and was not asked the question. If a filter or skip pattern indicates that data on the variable were not obtained for all respondents, that information should appear together with other documentation for that variable.
- Exact meaning of codes. The documentation should show the interpretation of the codes assigned to each variable. For some variables, such as occupation or industry, this information might appear in an appendix.

- Missing data codes. Codes assigned to represent data that are missing. As discussed above, such codes typically fall outside of the range of valid values. Different types of missing data should have distinct codes.
- Unweighted frequency distribution or summary statistics. These distributions should show both valid and missing cases.
- Imputation and editing information. Documentation should identify data that have been estimated or extensively edited.
- Details on constructed and weight variables. Datasets often include variables constructed using other variables. Documentation should include audit trails for such variables, indicating exactly how they were constructed, what decisions were made about imputations, and the like. Ideally, documentation would include the exact programming statements used to construct such variables. Detailed information on the construction of weights should also be provided.
- Location in the data file. For raw data files, documentation should provide the field or column location and the record number (if there is more than one record per case). If a dataset is in a software-specific system format, location is not important, but the order of the variables is. Ordinarily, the order of variables in the documentation will be the same as in the file; if not, the position of the variable within the file must be indicated.
- Variable groupings. For large datasets, it is useful to categorize variables into conceptual groupings.

Related publications

Citations to publications based on the data, by the principal investigators or others.

Technical information on files

Information on file formats, file linking, and similar matters.

Data collection instruments

Copies of the original data collection forms and instruments. 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. Providing verbatim data collection instruments is increasingly difficult because computer-assisted data collection modes often do not provide a hardcopy version of the interview, or if they do, it is in a format that is difficult to read. Copyrighted survey questions should be acknowledged with a citation so that users may access and give credit to the original survey and its author.

Flowchart of the data collection instrument
A graphical guide to the data, showing which respondents were asked which questions and how various items link to each other. This is particularly useful for complex questionnaires or when no hard-copy questionnaire is available.

Index or table of contents
An alphabetized list of variables with corresponding page numbers in the codebook that contain detailed information about each variable.

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

Interviewer guide
Details on how interviews were administered, including probes, interviewer specifications, use of visual aids such as hand cards, and the like.

Recode logic
An audit trail of the steps involved in creating recoded variables.

Coding instrument
A document that details the rules and definitions used for coding the data is particularly useful when open-ended responses are coded into quantitative data and the codes are not provided on the original data collection instrument.

5. Data analysis phase

In this chapter, we turn to important issues that should be addressed during the analysis phase when project staff are actively working with data files to investigate their research questions.

5.1 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 prepare 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 avoided to a degree if the research files are stored on a network where a single version of the data is maintained.

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.

5.1.1 Data and documentation versioning

One way to keep track of changes is to maintain explicit versions of a dataset. The first version might come from the data collection process, the second version from data cleaning, the third 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.

The documentation process starts at the beginning of the project and is ongoing, reflecting changes, additions, and deletions to the documentation. Here are a few suggestions to keep track of the various versions of the documentation files that will inevitably develop:

- Establish documentation versions similar to those used for the data. Versions could be established in the following manner: the first version contains results from the data-collection phase, the second version results from the data cleaning phase, and the third version adds any constructed variables, if applicable, to the end of the codebook, with appropriate labels and the formulas used to create them recorded when the variables are created.
- Keep a separate change file that tracks changes to the documentation.
- Denote changes in working documents with special characters (for example, use ??? or ***) that facilitate search, review, and replacement during the creation of the final version of the documentation file.
- Conduct a review of the final files to make sure the data and documentation are harmonized, i.e., that the final version of the documentation accurately corresponds to the final version of the data.
- Store final electronic versions of instruments and reports on a read-only basis.

5.1.2 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 readily. 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, and so on, 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 most software, however, produce files such as export/transport files or portable files that are compatible across platforms and systems. These kinds of files preserve all of the variable labeling and identification information in a format suitable for long-term preservation. Increasingly, these are the formats that archives prefer to receive. However, data producers should consider the implications of software changes

during the project to make certain that stored copies of data remain readable and understandable.

Non-ASCII characters: avoid the use of nonstandard character sets when you create archival quality documentation that will be used by a wide range of people over time. Be sure to remove non-ASCII characters from data and documentation files. Often, these characters are generated by proprietary word processing packages.

5.1.3 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 organized 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 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.

- *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 a 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).

In this case, one might want to consider other ways of storing the data. One option is to create 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.

A second approach to this problem the 'preferred approach' is to form separate files for the two kinds of records: one file for respondents and another file 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.

- *Relational databases*
  A relational database is a collection of data tables that are linked together through defined associations. For example, a database that includes a 'respondents' table and a 'children' table, as in the last example, would use a key variable (Family ID) to associate children with their parents. Relational databases allow a user to perform queries that select rows with specific attributes or combine data from multiple tables to produce customized tables, views, or reports. To preserve relational databases, users should export the database tables as flat rectangular files and preserve the table relationships using, for instance, SQL schema statements. When databases are used as survey instruments or other data input/out mechanisms, the look and feel of the user interface can be preserved by creating a static PDF image of the interface. Promising software is currently under development to normalize relational databases into non-proprietary formats such as XML.

5.1.4 Longitudinal/multi-wave study files

Many multiple data file studies are longitudinal, that is, they contain data collected from the same individuals over multiple points in time, or waves. Longitudinal studies often consist of hierarchical files. For longitudinal data, it is important to make file information as consistent as possible across waves. Data should include clearly specified linking identifiers, such as respondent IDs that are included in data from each wave so that users can link data files across time. In addition, identical variables across waves should have the same variable labels and values to make it easier for users to compare the data across files.

5.1.5 Data 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, in case of an emergency or disaster that could destroy years of work.

6. 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.

6.1 Respondent confidentiality

For most of this publication, the focus has been on data preparation methods that can serve the research needs of both principal investigators and analysts of secondary data. In this chapter, however, we highlight one area of divergence necessitated by the responsibility to protect respondent confidentiality. Researchers must pay special attention to this issue. 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, care must be used to make certain that the alterations do not unnecessarily reduce the researcher's ability to reproduce or extend the original study findings.

Below, we suggest steps that principal investigators can take to protect respondent confidentiality before they submit their data for archiving. But first, a quick review of why this is important.

6.1.1 Disclosure risk limitation, the principles

Social scientists must demonstrate a deep and genuine commitment to preserve the privacy 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.

DANS places 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. DANS adheres to the generally accepted academic codes of conduct for the exchange of knowledge and information in the Netherlands, as laid down in *The Netherlands code of conduct for scientific practice* (Association of Universities in the Netherlands – VSNU, 2005) and the

Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek (Association of Universities in the Netherlands – VSNU, 2005 – only available in Dutch).

6.1.2 The practice of protecting confidentiality

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

- *Direct identifiers* These are variables that point explicitly to particular individuals or units. They may have been collected in the process of survey administration and are usually easily recognized. Any variable that functions as an explicit name can be a direct identifier, for example, a license number, phone number, or mailing address. Data depositors should carefully consider the analytic role that such variables fulfill and should remove any identifiers not necessary for analysis.
- *Indirect identifiers* Data depositors should also carefully consider a second class of problematic variables: indirect identifiers. Such variables make unique cases visible. Some examples of possible indirect identifiers are detailed geography (e.g. province, municipality), organizations to which the respondent belongs, educational institutions from which the respondent graduated (and year of graduation), exact occupations held, places where the respondent grew up, exact dates of events, detailed income, and offices or posts held by the respondent. Indirect identifiers often are items that are useful for statistical analysis. The data depositor must carefully assess their analytic importance.
- *Geographic identifiers* Some projects collect data containing direct and indirect geographic identifiers that can be coordinates used with a mapping application. These data can be classified and displayed with Geographic Information System (GIS) software. Direct geographic identifiers are actual addresses (e.g., of an incident, a business, a public agency, etc.). As described above, the role of these variables should be considered and only included if necessary for analysis. Indirect geographic identifiers include location information such as province, municipality, and place where the respondent grew up.

- Treating 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 in a special manner when preparing a public-use dataset. Commonly used types of treatment are as follows:

- Removal: eliminating the variable from the dataset entirely
- Top-coding: restricting the upper range of a variable
- Collapsing and/or combining variables: merging data recorded in two or more variables into
- 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 analytic utility and covariate structure of the dataset while protecting subject confidentiality. Swapping is a service that archives may offer to limit disclosure risk. For more in-depth discussion of this technique, see O'Rourke, 2003 and 2006.
- 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.

Data producers can consult with DANS or any other data repository to design public-use datasets that maintain the confidentiality of respondents and are of maximum utility for all users. 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. If the investigator anticipates that significant work will need to be performed before deposit to anonymize the data, this should be noted and funds set aside for this purpose at the beginning of the project.

6.1.3 Restricted-use data collections

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, such as direct or indirect identifiers, will be removed or adjusted through the treatment procedures discussed above. These treatments, 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 and analytic potential of a dataset.

Creating a restricted dataset provides a viable alternative to removing sensitive variables. In such instances, a public-use dataset that has these variables removed is released, while the dataset preserving the original variables is kept as a restricted-use dataset. The restricted-use dataset is released only to approved clients/users who have agreed in writing to abide by rules assuring that respondent

confidentiality is maintained. 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.

7. Data publishing

DANS offers two possibilities for archiving and sharing research data. Data- and documentation files can be published via the online self-archiving system DANS EASY at http://easy.dans.knaw.nl/ or by making use of the advanced data management program, 'DANS EASY Online analysis tool,' at http://nesstar.dans.knaw.nl/webview/. In both systems, it is possible to give free and universal access to the data as well as restricted access to authorized individuals. Although open access to data is preferred, DANS likes to emphasize that it is also possible to deposit data under restricted access conditions. This means that access to restricted data is only possible after consultation and permission of the copyright-holder / original researcher(s) of the data. If a permission request has been denied by the copyright-holder of the data, access to the datafiles is not possible. In general it is recommended to contact DANS about the various options to publish data, in order to discuss for each individual project the best way of creating access to a dataset.

7.1 DANS EASY – Electronic Archiving System

The primary objective of depositing data in EASY is availability for secondary analyses. It is therefore important that the files are sufficiently transparent to enable reuse now as well as in the future. To realize this, data producers are asked to adhere to the following guidelines for depositing social science research data.

- *Data files*
  Researchers are requested to deposit all data files for the study in question. When it concerns several data files per survey, we recommend to supply a file list. This file list comprises an overview of the names of the files and a description of the contents.

- *Format*
  DANS prefers the following data formats: SPSS portable files, SAS transport files or STATA export files, but depositing other formats is also possible.

- *Labels*
  The files must have complete and clear variable and value labels.

- *Anonymization*
  The files must no longer contain any variables that might lead to the identification of individual respondents. The following variables, if applicable, therefore need to be removed:
  - Name of the respondent
  - Address data of the respondent
  - Telephone number
  - Social security number

  and the following variables must be recoded:
  - Date of birth (to be recoded to year of birth)
  - Letters of the postcode (recode to 4 digits)

If the exact names of professions are used, we recommend to create an entirely new file in which the professions have been recoded. DANS will then make this new file available, but can also upon special request and after permission from the depositor of the data, make the file with the exact names of the professions accessible. This way, the original data remain available for specific analysis purposes.

- *Documentation files*
  Researchers are kindly requested to submit the following documentation, preferably as a PDF file, in ASCII format or as a MS-Word file:
  - The questionnaire(s) or other survey instruments
  - A description of the variables or a codebook
  - The survey report with the description of:
    - » the sampling
    - » the non-response
    - » the data collection method
    - » the fieldwork report
    - » the weighting variables with an explanation (if applicable)
    - » the construed and/or derived variables with an explanation (if applicable).

DANS encourages researchers to document their research data as completely as possible, but only a few Dublin Core metadata elements are required when data are deposited in DANS EASY.

7.2 Persistent identifiers

DANS EASY automatically generates a persistent identifier in every new metadata record. This guarantees that if the persistent identifier has been included in references to the dataset, the data always can be found on the web. The Internet allows researchers to share, obtain and refer to scientific publications, research information and scientific data easily. References to scientific information on the Internet are often achieved by using URLs. However, after some years, URLs tend to suffer from link rot: readers are confronted with a 'page not found' message when clicking on a link. Persistent identifiers provide a technique and an organizational structure for avoiding this problem. The cause of link rot lays in the fact that URLs are meant to identify a location whereas researchers actually want to specify the resource at that location. This works fine as long as the resource can be maintained at that specified location forever, but in practice that is not feasible. Scientists need a trustworthy way of referring to scientific output on the Internet. A solution to ensure the integrity of scientific referencing is called 'persistent identifiers.' Persistent identifiers allow unique naming of a resource on the Internet, independent from its location. This enables researchers to refer to the resource itself instead of its location. A resolver is used to facilitate translation and forwarding to the registered location. The essential part of this construction is that the locations can be updated when objects move, without having to modify references. Obviously, this requires effort from the organizations that wish to keep their resources accessible. It is the responsibility of DANS to keep research data permanently accessible. Therefore, it aims to assign and maintain persistent identifiers. DANS has chosen the Uniform Resource Name (URN) as the best suitable identifier for this purpose and the W3C entrusted it with an appropriate sub-name space.

7.3 DANS EASY online analysis tool

DANS also offers data producers the option to publish survey data and metadata via the DANS EASY online analysis tool at http://nesstar.dans.knaw.nl/webview/, which is based on the Nesstar publishing tool. It uses an advanced data management program that increases the accessibility of data and documentation. This means that it is possible to disseminate all the variables in a dataset and to visualize data with maps, graphs and tables. This tool is also DDI compliant. The DDI, Data Documentation Initiative, is a metadata standard used for documenting datasets developed in European and North American data archives, libraries and official statistics agencies. With the use of the publishing tool it is easy to create

datasets and generate documentation according to the DDI standard. For additional information about DDI see: http://www.ddialliance.org/

With the DANS EASY online analysis tool, data can be published on a DANS server, as well as on a dedicated server for data from a special project. For a presentation of the system, or instruction and support, please contact DANS.

7.4 Data guides

DANS EASY offers online help (in English) during the process of data depositing and for additional information the following data publications (in Dutch) are available:

Wegwijzer Digitaal Deponeren, Sociale Wetenschappen – DANS Data Guide 5 (DANS, 2009)

Wegwijzer Digitaal Deponeren, Historische Wetenschappen – DANS Data Guide 4 (DANS, 2009)

For additional information about depositing qualitative data from specific projects see:

Wegwijzer Digitaal Data Deponeren, Interview Data, Getuigen Verhalen (DANS, 2008)

7.5 Data Seal of Approval – DSA

The Data Seal of Approval was established by a number of institutions in the Netherlands as well as abroad committed to the long-term archiving of research data. By assigning the seal, the DSA group seeks to guarantee the durability of the data concerned, but also to promote the goal of durable archiving in general. The guidelines pertain to data producers, data consumers and data repositories.

DANS adheres to the following Data Seal of Approval guidelines for data repositories:

- The data repository has an explicit mission in the area of digital archiving and promulgates it
- The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects
- The data repository applies documented processes and procedures for managing data storage
- The data repository has a plan for long-term preservation of its digital assets
- Archiving takes place according to explicit workflows across the data life cycle

- The data repository assumes responsibility from the data producers for access and availability of the digital objects
- The data repository enables the users to utilize the research data and refer to them
- The data repository ensures the integrity of the digital objects and the metadata
- The data repository ensures the authenticity of the digital objects and the metadata
- The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Additionally the DSA Guidelines enable data consumers to assess data repositories in a reliable manner.

More information about the Data Seal of Approval see: www.datasealofapproval.org.

References

American Statistical Association. (1999, August 7). *Ethical Guidelines for Statistical Practice*. Prepared by the Committee on Professional Ethics, Approved by the Board of Directors.

Association of Universities in the Netherlands – VSNU (2005). *Gedragscode voor gebruik van persoonsgegevens in wetenschappelijk onderzoek*.

Association of Universities in the Netherlands – VSNU (2005). *Netherlands code for conduct of scientific research*.

Babbie, Earl. (1990). *Survey Research Methods* (2nd ed.). Belmont, CA: Wadsworth

Blank, Grant, and Karsten Boye Rasmussen. (2004). 'The Data Documentation Initiative: The Value and Significance of a Worldwide Standard.' *Social Science Computer Review 22*: 307-318

Data Archiving and Networked Services – DANS (2009). *Data Seal of Approval*.

ESDS Qualidata. (2007, September). *Reusing Qualitative Data*.

Fienberg, Stephen E. (1994). 'Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions.' *Annual Review of Public Health 15*. Palo Alto, CA: Annual Reviews, Inc. Accessed May 22, 2009.

Fishbein, Estelle. (1996). *Policy Considerations: Access to and Retention of Research Data*. Council on Government Relations.

Green, Ann G., and Myron P. Gutmann. (2007). 'Building Partnerships Among Social Science Researchers, Institution-based Repositories, and Domain Specific Data Archives.' *OCLC Systems and Services: International Digital Library Perspectives 23*: 35-53.

Groves, Robert M., F.J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and R. Tourangeau. (2004). *Survey Methodology*. New York: Wiley.

International Organization for Standardization. (2003, February 24). *ISO 14721:2003: Reference Model for an Open Archival Information System*. Geneva, Switzerland: International Organization for Standardization.

Jacobs, James A., and Charles Humphrey. (2004). 'Preserving Research Data.' *Communications of the ACM 47(9)*: 27-29.

King, Gary. (1995). 'Replication, Replication' *PS: Political Science and Politics, 28(3)*: 443-499.

King, Gary. (2006). 'Publication, Publication.' *PS: Political Science & Politics, 39(1)*: 119-125

Little, Roderick, and Donald Rubin. (2002). *Statistical Analysis with Missing Data* (2nd ed.) Hoboken, NJ: Wiley.

Marz, Kaye, and Christopher S. Dunn. (2000). *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.

McKnight, Patrick E., Katherine M. McKnight, Souraya Sidani, and Aurelio Jose Figuerdo. (2007). *Missing Data: A Gentle Introduction*. New York: The Guilford Press.

National Institutes of Health. (2003, February 26). *Final NIH Statement on Sharing Research Data*. Retrieved May 22, 2009.

National Institutes of Health, Office of Extramural Research. (2003, March 5). *NIH Data Sharing Policy and Implementation Guidance*.

National Science Board, National Science Foundation. (2005, September). *Current Policies on Data Sharing and Archiving. Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* (pp. 57-71).

National Science Foundation USA. (1989). *NSF Important Notice 106*.

O'Rourke, JoAnne McFarland. (2003, Fall). 'Disclosure Analysis at ICPSR.' *ICPSR Bulletin 34(1)*: 3.

O'Rourke, JoAnne McFarland, Stephen Roehrig, Steven G. Heeringa, Beth Glover Reed, William C. Birdsall, Margaret Overcashier, and Kelly Zidar. (2006, September). 'Solving Problems of Disclosure Risk While Retaining Key Analytic Uses of Publicly Released Microdata.' *Journal of Empirical Research on Human Research Ethics, 1(3)*.

United States Department of Health and Human Services. Office for Human Research Protections. (2006, October 25). *Basic Assurance Requirement*.

United States Department of Health and Human Services. Office for Human Research Protections. (2005, June 6). *Step-by-Step Instructions for Filing a Federalwide Assurance for Institutions Within the United States*.

University of Michigan, Institutional Review Boards (IRB), *Health Science and Behavioral Sciences (2007)*. Office of the Vice President for Research 'Informed Consent.'

Wegwijzer Digitaal Data Deponeren, Interview Data, Getuigen Verhalen (DANS, 2008).

Wegwijzer Digitaal Data Deponeren, Sociale Wetenschappen (DANS, 2009).

Wegwijzer Digitaal Data Deponeren, Historische Wetenschappen (DANS, 2009.)

Zelenock, Tom, and Kaye Marz. (1997). 'Archiving Social Science Data: A Collaborative Process.' *ICPSR Bulletin, 17(4):* 1-4.

Contributors to the fourth edition

George Alter
Peter Doorn (edition for the Netherlands)
Peter Granda
Russel Hathaway
Cedrick Heraux
Peter Joftis
Felicia LeClere
Jared Lyle
Kaye Marz
Nancy McGovern
Dan Meisler (edition for the Netherlands)
Elizabeth Moss
JoAnne McFarland O'Rourke
Beth Panozzo
Amy Pienta
Lisa Quist
Jetske van der Schaaf (edition for the Netherlands)
Ruth Shamraj (ICPSR edition, and edition for the Netherlands)
Mike Shove
Mary Vardigan (ICPSR edition, and edition for the Netherlands)