Child Care & Early Education
RESEARCH CONNECTIONS

Guide to Archiving Data with *Research Connections*

Considerations throughout the research lifecycle

We are excited that you intend to archive your data with *Research Connection*s and are pleased to offer this Guide to Archiving Data. This document covers data management techniques that can be incorporated into your project immediately. It is intended to be reviewed at the beginning of the research process and consulted along the way in order to incorporate best practices throughout that will facilitate a smooth transition to the *Research Connections* archive at project completion.

Sections of this document were pulled from *ICPSR's Guide to Social Science Data Preparation and Archiving.* The full *Guide* can be found here: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Please feel free to reach out to Johanna Bleckman (bleckman@umich.edu), manager of the *Research Connections* Archive of Datasets, with any questions.

NCCP
National Center for Children in Poverty
Mailman School of Public Health
Columbia University

ICPSR

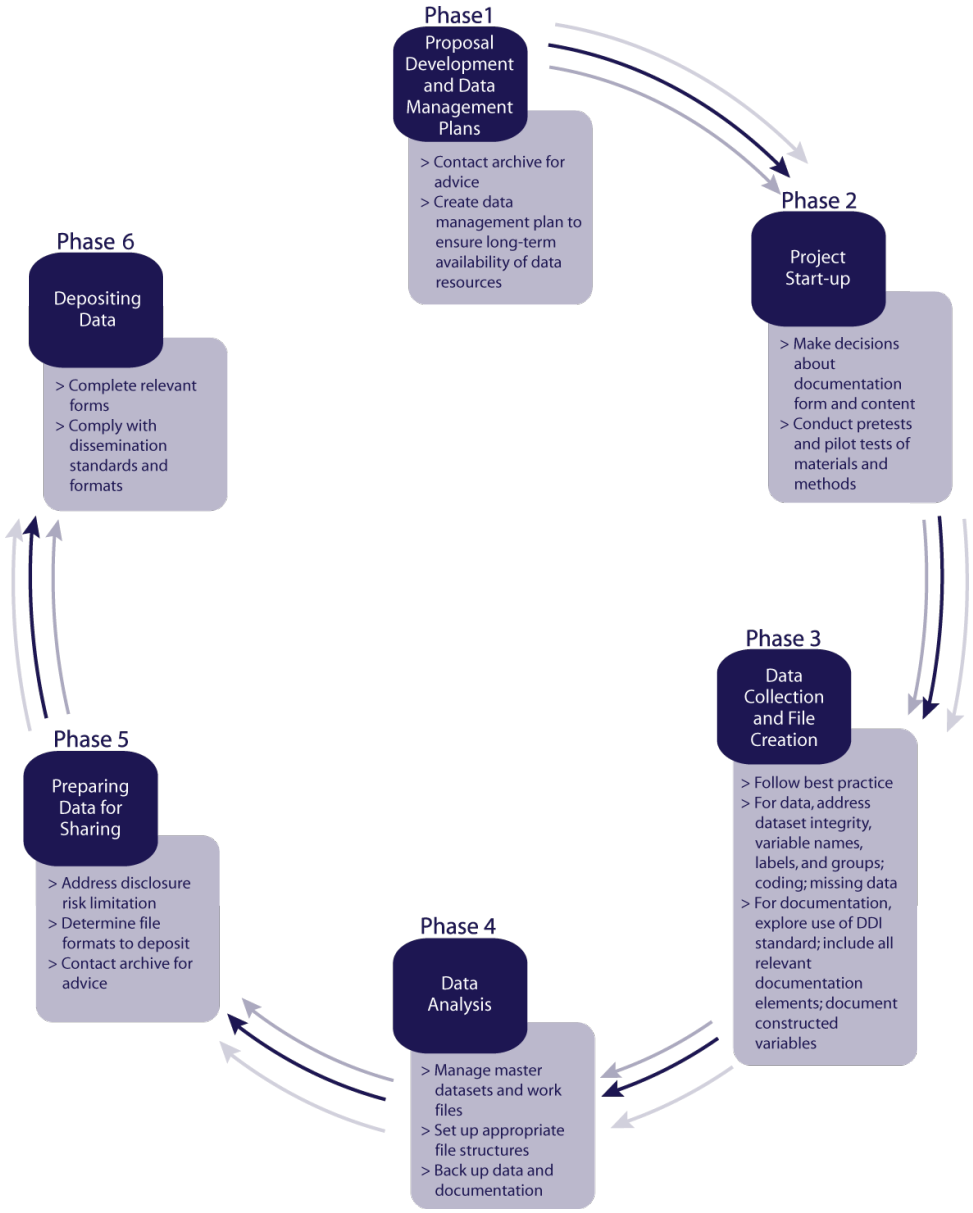
OPRE

Contents

IMPORTANCE OF GOOD DATA MANAGEMENT

This document will discuss and describe best practices for creating research data, both quantitative and qualitative, along with best practices for creating associated metadata[1] and documentation. Once the research project has started, you will want to continue to think about and plan for the final form of the collection, including metadata, which will ultimately be deposited with *Research `Connections*. Planning for the management and archiving of a data collection at the outset is important to the archiving project's success. Poor or uneven data management can yield unreadable, incomplete, or unreliable data collections, which can frustrate and limit future users. Following these guidelines will help ensure your data collection is complete and independently understandable.

[1] Metadata are defined as technical documentation at both the variable/value and study levels. Examples include codebooks, user guides, questions text, and project summary information including dates of coverage, PI name, and descriptions of sampling and methodology. For more information, please visit: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf



Planning Ahead for Archiving and Preservation of Data

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

Phases 3-6 are covered in the respective sections of this document. All phases are covered in depth in ICPSR's *Guide to Social Science Data Preparation and Archiving*: http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

INITIAL QUESTIONS TO CONSIDER

At a minimum, a project plan should involve decisions on the following data and documentation topics, many of which are related to the core data management plan. Documentation should be as much a part of project planning as data-related considerations, such as data collection, questionnaire construction, or analysis plans. Further, certain data collection methods (teleform, CASI, etc.) can minimize data entry errors given that they directly record interview data into a database.

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?

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

Data integrity: How will data be input or captured? 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 as the files move through data entry, cleaning, and analysis?

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? What documentation standard will be used?

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

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? Research project Web sites and various Intranet options are increasingly used for capturing this kind of information, and *Research Connections* is prepared to include Web-based information in deposits.

BEST PRACTICE IN CREATING RESEARCH DATA

Following best practice in building both the data and documentation components of a collection is critical. This section describes widely accepted norms for quantitative, qualitative, GIS, and other types of data in the social sciences.

QUANTITATIVE DATA

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 people 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.
+ **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. For example, a respondent indicates that she did not work within the last week, yet the data show that she reported income for that week. 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.

Data File and Variable Creation

Documenting all steps of data file creation will facilitate responsible and informed secondary analysis and help prevent data misuse. Documenting coding decisions, variable creation, patterns of missing data, how and in what method constructed variables were created, scales used for qualitative variables, if applicable, etc.

File Formats

Depositors are encouraged to submit data files as SPSS, SAS, or Stata files. ASCII files are also acceptable as long as they are accompanied with setup files. Datasets in other formats are accepted as well.

Documentation can be submitted as ASCII, Microsoft Word, and DDI XML files, among other formats. *Research Connections* uses question text in various ways, including within online analysis systems when appropriate, and in variable search tools available on the *Research Connections* website. Questions text extraction is easiest from plain text files; we strongly prefer question text to be provided in this way.

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. There are several systems for constructing variable names:

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.

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.

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.

Codes and Coding

Before survey data are analyzed, the interview or questionnaire responses must be represented by numeric codes. Common coding conventions (a) assure that all statistical software packages will be able to handle the data, and (b) 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.

User-coded responses: 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 *Research Connections* prior to dissemination. We recommend that data producers review all files intended for deposit and make any necessary disclosure risk modifications, as the original team is best suited to identify the unique risks of their particular study.

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.

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 in dollars 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.

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 nonexistent 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.

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.

Geographic Identifiers

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 child care center, a residence, etc.). Indirect geographic identifiers include location information such as state, county, census tract, census block, 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 (such as state, county, or census tract, division, or region).

Archiving and dissemination of geospatial data should be discussed with the project officer and *Research Connections* staff prior to deposit.

The Practice of Protecting Confidentiality

Two kinds of variables often found in social science datasets present problems that could endanger the confidentiality of research subjects: direct and indirect 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. For instance, Social Security numbers uniquely identify individuals who are registered with the Social Security Administration. Any variable that functions as an explicit name can be a direct identifier -- for example, 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. For instance, a United States ZIP code field may not be troublesome on its own, but when combined with other attributes like race and annual income, a ZIP code may identify unique individuals (e.g., extremely wealthy or poor) within that ZIP code, which means that answers the respondent thought would be private are no longer private. Some examples of possible indirect identifiers are detailed geography (e.g., state, county, or census tract of residence), organizations to which the respondent belongs, places where the respondent grew up, exact dates of events, and detailed income. Indirect identifiers often are items that are useful for statistical analysis. The data depositor must carefully assess their analytic importance. Do analysts need the ZIP code, for example, or will data aggregated to the county or state levels suffice?

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:** combining values of a single variable or 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.

Data producers can consult with *Research Connections* staff to design public-use datasets that maintain the confidentiality of respondents and are of maximum utility for all users. The staff will also 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.

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 -- 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 *Research Connections* staff following review of the data content, or after consultation between the depositor and archive staff. Maintenance of, and approval of access to, a restricted-use file is managed by *Research Connections* staff in accordance with the terms of access.

Access to restricted-use files is offered to approved researchers under a set of highly controlled conditions. The right to use these files requires acceptance of a restricted data use agreement that spells out the conditions that a researcher must accept before obtaining access. The standard *Research Connections* agreement requires that a researcher provide a detailed summary of the research question and precisely explain why access to the confidential variables is needed. Each user of restricted data must provide a data protection plan outlining steps he or she will take to safeguard the data during the project period. Researchers are given access to the data for a limited time period, at the end of which they must return the original files, or destroy them in good faith. The restricted-use dataset approach effectively permits access to sensitive research information while protecting confidentiality, and has proven acceptable to researchers.

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.

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
+ Unstructured or semi-structured diaries
+ Observation field notes/technical fieldwork notes
+ Case study notes

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 *Research Connections* staff.

Confidentiality in Qualitative Data

Ideally, prior to submitting qualitative data to *Research Connections*, 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 facility names, 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," or the general indicator "..." 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 *Research Connections* staff do not make unnecessary changes to the investigator's modifications when performing our confidentiality review. Such information will thus also be made available to secondary users of the data to assist them with their use of the data.

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 *Research Connections* 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.

Other Data Types

Early care and education 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 *Research Connections* to discuss an alternative set of specifications that might be mutually satisfactory.

Best Practice in Creating Metadata

Metadata — often called technical documentation or the codebook — are critical to effective data use as they convey information that is necessary to fully exploit the analytic potential of the data. Preparing high-quality metadata can be a time-consuming task, but the cost can be significantly reduced by planning ahead. In this section, we describe the structure and content of optimal metadata for social science data.

XML

ICPSR recommends using XML to create structured documentation compliant with the Data Documentation Initiative (DDI) metadata specification, an international standard for the content and exchange of documentation. XML stands for eXtensible Markup Language and was developed by the W3C, the governing body for all Web standards. Structured, XML-based metadata are ideal for documenting research data because the structure provides machine-actionability and the potential for metadata reuse. XML defines structured rules for tagging text in a way that allows the author to express semantic meaning in the markup. Thus, question text — for example, <question>Do you own your own home?</question> — can be tagged separately from the answer categories. This type of tagging embeds "intelligence" in the metadata and permits flexibility in rendering the information for display on the Web.

Data Documentation Initiative (DDI)

The Data Documentation Initiative (DDI) provides a set of XML rules specifically for describing social, behavioral, and economic data. DDI is designed to encourage the use of a comprehensive set of elements to describe social science datasets, thereby providing the potential data analyst with broader knowledge about a given collection. In addition, DDI supports a life cycle orientation to data that is crucial for thorough understanding of a dataset. DDI enables the documentation of a project from its earliest stages through questionnaire development, data collection, archiving and dissemination, and beyond, with no metadata loss.

DDI Authoring Options

Several XML authoring tools are available to facilitate the creation of DDI metadata. With a generic XML editor, the user imports the DDI rules (i.e., the DDI XML Schema) into the software and is then able to enter text for specific DDI elements and attributes. The resulting document is a valid DDI instance or file. There are also DDI-specific tools, such as Nesstar Publisher and Colectica, which produce DDI-compliant XML markup automatically. For more information on DDI and a list of tools and other XML resources, please consult the DDI Web site at www.ddialliance.org.

Important Metadata Elements

Since most standard computer programs will produce frequency distributions that show counts and percents for each value of numeric variables, it may seem logical to use that information as the basis for documentation, but there are several reasons why this is not recommended. First, the output typically does not show the exact form of the question or item. Second, it does not contain other important information such as skip patterns, derivations of constructed variables, etc.

A list of the most important items to include in is presented below.

+ **Principal investigator name(s), and affiliation(s) at time of data collection**
+ **Title**
+ **Funding sources:** names of funders, including grant numbers and related acknowledgments.
+ **Data collector/producer:** persons or organizations responsible for data collection, and the date and location of data production.
+ **Project description:** a description of the project, its intellectual goals, and 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:** a description of 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 provided, 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.
+ **Data source(s):** if a dataset draws on resources other than surveys, citations to the original sources or documents from which data were obtained.

- **Unit(s) of analysis/observation:** a description of 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, this information might appear in an appendix.
    - Missing data codes: codes assigned to represent data that are missing. 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. Ideally, documentation would include the exact programming statements used to construct such variables. Detailed information on the construction of weights should also be provided.
    - Variable groupings: particularly 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 information.
- **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. Copyrighted survey questions should be acknowledged with a citation so that users may access and give credit to the original survey and its author.

DEPOSITING DATA AND FINDING MORE INFORMATION

Once you have prepared your data and documentation files for deposit, please use the following link to access our Data Deposit Form:

http://www.researchconnections.org/content/childcare/find/contribute.html

Research Connections and ICPSR have information and resources to share on a variety of subtopics related to social science data archiving. For further information, we have provided links below to sections of the ICPSR Guide to Social Science Data Preparation and Archiving. Please don't hesitate to contact us with any questions or feedback.

The full ICPSR Guide to Social Science Data Preparation and Archiving
http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Proposal Development and Data Management Plans
https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter1.html

Data Analysis: File Structure and Versioning
https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter4.html

Disclosure Risk Analysis and Mitigation (Respondent Confidentiality)
https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html
https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html

Depositing Data
https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter6.html

References

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

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

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

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.