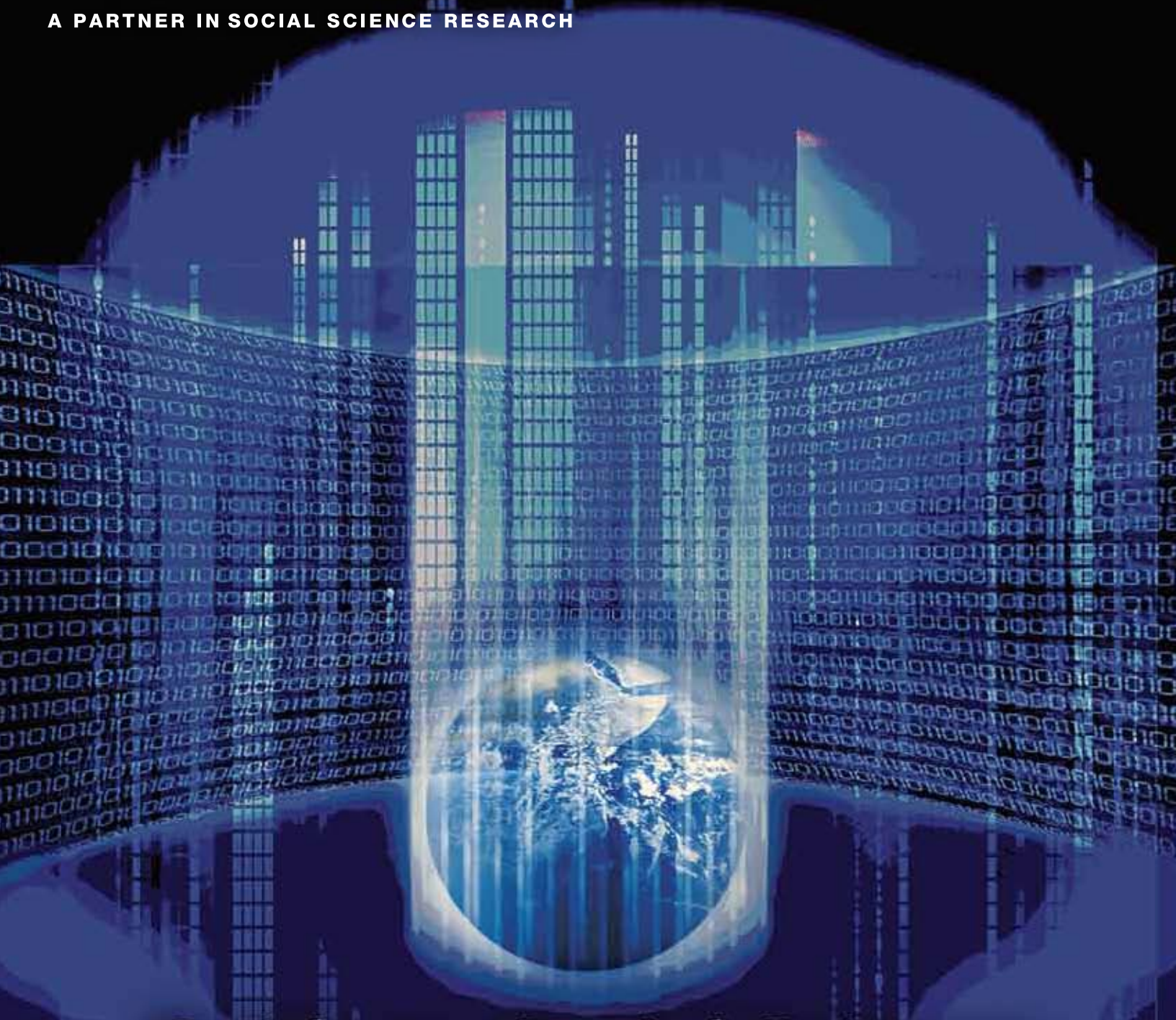
ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

A PARTNER IN SOCIAL SCIENCE RESEARCH



Guide to Social Science Data Preparation and Archiving

Best Practice Throughout the Data Life Cycle • 5th edition

Acknowledgements

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

First edition, 1997. Second edition, 2000. Third edition, 2005. Fourth edition, 2009. Fifth edition, 2012.

ACKNOWLEDGEMENTS

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

Third edition. This edition discussed new tools and standards, and emphasized the life cycle approach to research data. The third edition was the product of a collaboration among ICPSR and several of its projects, which also provided financial support. ICPSR thanks the following individuals for their contributions to the third edition: Erik Austin, Corey Colyer, Darrell Donakowski, Russel Hathaway, Cynthia Hoxey, Peter Joftis, Kaye Marz, Shawn Marie Pelak, Amy Pienta, Ruth Shamraj, and Mary Vardigan.

Fourth edition. This edition contained updated information related to digital preservation, informed consent, copyright, and qualitative and geospatial data. It was the product of a collaboration among ICPSR and several of its projects. ICPSR thanks the following individuals for their contributions to the fourth edition: George Alter, Peter Granda, Russel Hathaway, Cedrick Heraux, Peter Joftis, Felicia LeClere, Jared Lyle, Kaye Marz, Nancy McGovern, Elizabeth Moss, JoAnne McFarland O'Rourke, Beth Panozzo, Amy Pienta, Lisa Quist, Ruth Shamraj, Mike Shove, and Mary Vardigan.

Fifth edition. The latest edition contains new information on data management plans, video files, respondent disclosure protection, and virtual data enclaves. ICPSR thanks the following individuals for their contributions: George Alter, Peter Granda, Russel Hathaway, Jared Lyle, Kaye Marz, Nancy McGovern, Elizabeth Moss, JoAnne McFarland, Amy Pienta, Michael Shove, Lisa Quist, Wendi Fornoff, Sue Hodge, Robbin Gonzalez, Dan Meisler, Jenna Tyson, and Mary Vardigan.

SUGGESTED CITATION

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.). Ann Arbor, MI.

ISBN 978-0-89138-800-5

Table of Contents

About ICPSR

Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR) provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community. The ICPSR data archive is unparalleled in its depth and breadth; its data holdings encompass a range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, health and medical care, education, child care research, law, and substance abuse. ICPSR also hosts several sponsored projects focusing on specific disciplines or topics. Social scientists in all fields are encouraged to archive their data at ICPSR.

Why Should I Archive Data?

ICPSR data advance scientific knowledge by making it possible for researchers around the world to conduct secondary analyses. ICPSR supports the social sciences by sharing data and methods with the research community, allowing for replication, verification, and extension of original findings, and making sure this is possible over time. Archiving data with ICPSR ensures the long-term preservation of data, protecting it from obsolescence, loss, or irreversible damage. Another advantage of archiving data with ICPSR is that our trained staff are available to provide user support. In addition, research funding agencies — including the National Institutes of Health (NIH), the National Science Foundation (NSF), and the National Endowment for the Humanities (NEH) — are increasingly recommending or requiring that all funded projects contain plans for sharing and managing data. Archiving with ICPSR, which continually adheres to prevailing standards and practice for long-term preservation, can meet those requirements.

How Do I Deposit Data?

ICPSR works closely with researchers who submit their data collections for use by the social science research community. This publication provides information on how to prepare data for deposit and how researchers can ensure access to their data by others in the future. For more information about the preparation of data for deposit and other general inquiries, please see our online deposit form or send an email to deposit@icpsr.umich.edu.

What Kinds of Specialized Services Does ICPSR Offer?

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

- In addition to quantitative data, ICPSR accepts **qualitative research data** (including transcripts and audiovisual media) for preservation and dissemination. ICPSR is committed to digital preservation and encourages researchers to consider depositing their data in emerging formats, such as Web sites, geospatial data, biomedical data, and digital video. Please contact ICPSR for information about depositing any and all types of data content or formats related to your project.
- The ICPSR **electronic deposit form** facilitates the secure upload of files on the Web and enables the depositor to describe the data collection being deposited. The depositor may designate others to access and add information to the form, which may be completed in more than one session if desired.

- ICPSR assigns a **persistent identifier** to each dataset.
- ICPSR standard practice for dataset documentation involves **variable-level DDI XML markup**, which enables precise searching.
- ICPSR curates and disseminates data that require **special handling and restrictions** in order to protect human subjects. Restricted datasets require a special application process (e.g., data protection plan, IRB approval).
- ICPSR maintains a **secure data enclave** to store and protect data with the highest confidentiality standards. Data stored in the data enclave may be analyzed in our onsite, supervised computing facility with prior approval. We also are implementing a **virtual data enclave** allowing remote, secure access to restricted-use data that remain on ICPSR's servers.
- Arrangements can be made to **deposit data with delayed dissemination**. This occurs when release would pose a confidentiality risk or when other dissemination plans have been made. See the ICPSR Web site for our policy regarding preservation with delayed dissemination.
- ICPSR requires that all staff handling data undertake special training to receive **certification in handling sensitive data with confidentiality concerns**.
- In June 2011, ICPSR became one of the first six data repositories to earn the **Data Seal of Approval**, an international standard that demonstrates that archives are taking appropriate measures to ensure the long-term availability and quality of the data they hold.
- Researchers may conduct simple or complex analyses, recode and compute new variables, and subset variables or cases for downloading through our **online data analysis** system, Survey Documentation and Analysis (SDA).
- ICPSR can create **specialized Web pages to enhance data dissemination**. For complex studies, ICPSR creates a user guide to assist new users in working with the data. This can be made available through the Web site along with other online features such as FAQs and electronic mailing lists.
- **Training in documentation preparation and the Data Documentation Initiative (DDI)** can be provided by ICPSR staff. Learn how to create DDI-compliant documentation, including an XML codebook. This can enhance data usability during the project phase and make it easier to deposit data when the project is complete.
- ICPSR staff can train other scholars in how to analyze your data by offering **specialized data analysis workshops** during our Summer Program in Quantitative Methods of Social Research. Webinars may also be conducted.
- ICPSR hosts **training in digital preservation** for managers implementing preservation programs.
- ICPSR has received funding to build the capacity to acquire, archive, and securely stream **video research data**. This requires new infrastructures and processes, which are currently being developed. If you have video research data, please contact Robbin Gonzalez at rpgonzal@umich.edu to discuss the most current deposit and dissemination options.

Contact ICPSR at netmail@icpsr.umich.edu for more information about these services or for other ways to customize your data products for dissemination, or visit www.icpsr.umich.edu.

Importance of Data Sharing and Archiving

Archives and domain repositories 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.

In recent years, several national scientific organizations have issued statements and policies underscoring the need for prompt archiving of data, and some funding agencies have begun to require that the data they fund be deposited in a public archive. The National Institutes of Health (NIH) now requires a data sharing plan for large projects, and in 2011 the National Science Foundation (NSF) began to require a data management plan as part of every grant application. The National Endowment for the Humanities (NEH) has followed suit.

These statements from leading research funding agencies demonstrate that the data sharing ethic is integral to maximizing the impact and benefit of research dollars. Experience has demonstrated that the durability of the data increases and the cost of processing and preserving the data decreases when deposits are timely. Further, archived data result in a greater number of publications and a higher profile for data producers (Pienta, 2010).

Data sharing also allows 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. Some standard datasets, such as the General Social Survey and the National Election Studies, have produced literally thousands of papers that could not have been possible if the authors had to collect their own data. Archiving makes known to the field what data have been collected so that additional resources are not spent to gather essentially the same information.
- Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

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

Planning Ahead for Archiving and Preservation of Data

Data management and sharing plans should be developed in conjunction with an archive to maximize the utility of the data and to ensure the availability of the data in the future. We recommend that researchers consult as early as possible with the data archive in which they plan to deposit data; this will facilitate preservation and dissemination of the research data.

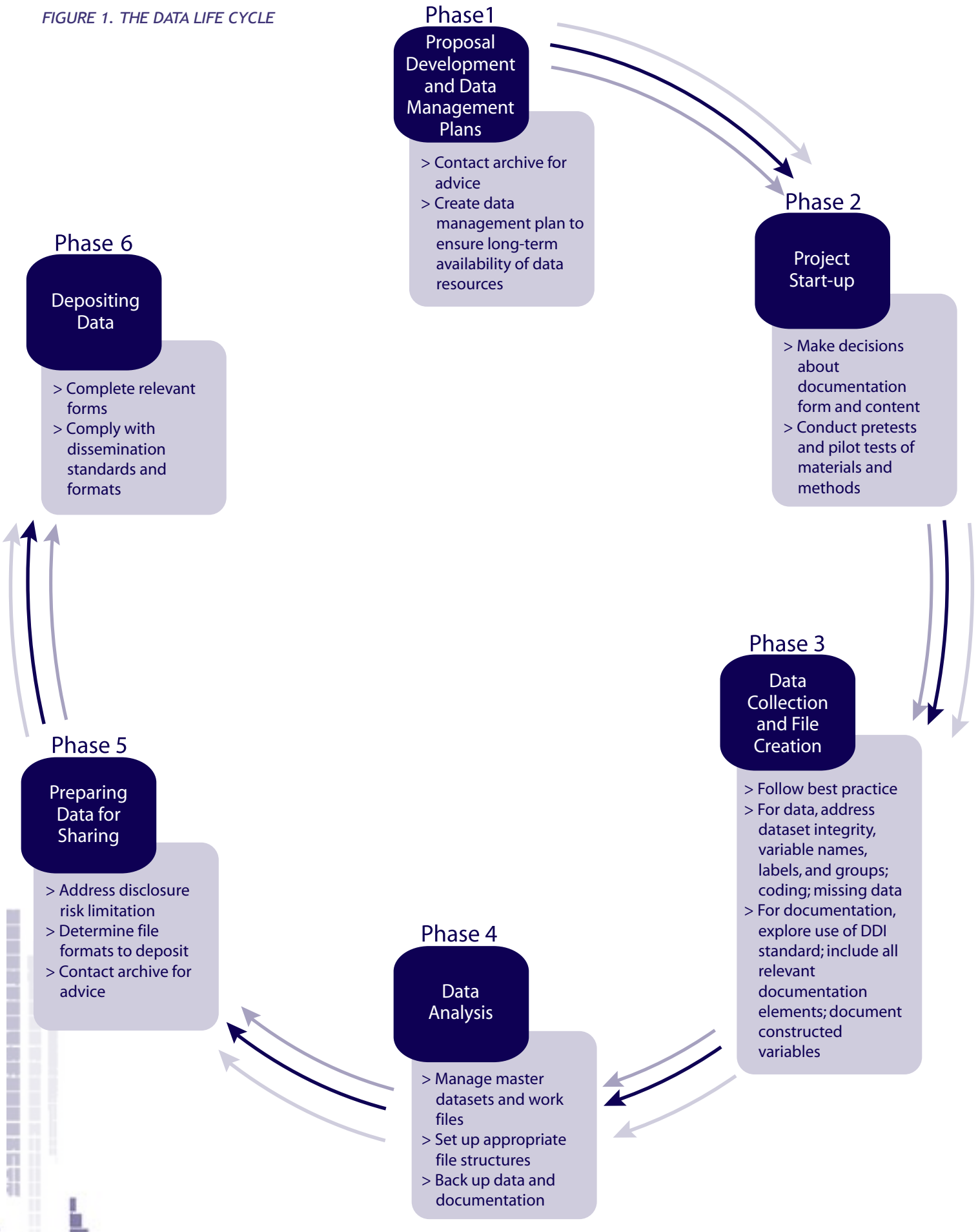
Data archives are committed to maintaining social science research data for the long term, for the benefit of future researchers, and to assist data creators in meeting the stipulations of their grantors. There are several factors to consider when selecting a data archive or domain repository for deposit with a view toward long-term access to your data. These include evidence of an explicit institutional commitment to preservation, and indicators that the preservation program is sustainable and credible and offers preservation and access services that are able to meet your short-term and long-term requirements. Compliance with the OAIS Reference Model is also an important factor to consider when selecting an archive for deposit. For information on digital preservation standards and a glossary of terms, see the Digital Preservation page on the ICPSR Web site.

The Data Life Cycle

Researchers 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."

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.

FIGURE 1. THE DATA LIFE CYCLE



Using the *Guide*

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

For guidance on social science terminology used in the *Guide*, please consult ICPSR's Data Tutorial.

The *Guide* is also available online.

Proposal Development and Data Management Plans

In the earliest stages of proposal development, researchers should consider the growing emphasis — and new requirements, in some cases — on data management plans and data sharing generally. As indicated earlier, funding agencies 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.

In 2010, the National Science Foundation's Social, Behavioral and Economic Sciences Directorate started requiring that all grant applications include a data management plan, which should include the following elements:

- Roles and responsibilities
- Expected data, including types of data to be produced by the research
- Period of data retention
- Data formats and dissemination
- Data storage and preservation of access

Other federal funding agencies such as NIH have long-standing policies recommending similar plans for data management. The information in this section is meant to help researchers meet these requirements, and is taken in large part from the ICPSR Web site on data management plans.

A table listing all recommended elements (including the NSF Mapping category of each), and an example data management plan for depositing data with ICPSR are included at the end of the chapter.

Recommended Elements

DATA DESCRIPTION

Provide a brief description of the information to be gathered, including the nature, scope, and scale of the data to be produced. This will help reviewers understand the data, their relationship to existing data, and possible disclosure risks.

REVIEW OF EXISTING DATASETS

A thorough review of existing data in related journals and data archives will make clear the value of the proposed research and why currently available datasets are inadequate to answer your research questions.

FORMAT

Describe the formats of the data in the submission, distribution, and preservation phases (note that these formats may be the same). Choosing formats preferred for archiving can make processing and release of data faster and more efficient. Platform-independent and non-proprietary formats will ensure that data will be usable over the long term.

Note that when writing the grant proposal, it is useful to think of "data" in the widest sense, including numeric data files, interview transcripts, and other qualitative materials such as diaries and field notes. Increasingly, social science research data include audio and video formats, geospatial data, biomedical data, and Web sites, and many data archives are interested in capturing this broadening array of data.

Archiving and disseminating derived datasets — that is, those resulting from the combination of data from more than one data source, including existing data outside the current research scope — also should be considered. See Phase 6, Depositing Data, for a more in-depth discussion.

METADATA

Describe the metadata to be provided along with the generated data, and discuss the metadata standards used. As metadata are often the only form of communication between the secondary analyst and the data producer, good descriptive metadata are essential for effective data use. Structured or tagged metadata, such as the XML format of the Data Documentation Initiative (DDI), are optimal because of the flexibility they offer in display. XML is also preservation-ready and machine-actionable. For a more detailed discussion on metadata and documentation, please see the "Best Practices in Creating Metadata" section in Phase 3: Data Collection and File Creation.

STORAGE AND BACKUP

Indicate how and where you will store copies of your research files to ensure their safety, as well as how many copies you will keep and how you will synchronize them. The best practice for protecting data is to store multiple copies in multiple locations.

SECURITY

Describe measures you will take to ensure your data are secure. This is an important consideration over the entire life cycle of the data. Raw data may include direct identifiers of study participants and should be well protected during collection and processing. Examples of good security practices include access restrictions such as passwords, encryption, power supply backup, and virus and intruder protections.

RESPONSIBILITY

State who will act as the responsible steward for the data throughout the data life cycle. Researchers should describe any atypical circumstances. For example, if there is more than one principal investigator, describe the division of responsibilities between them.

INTELLECTUAL PROPERTY RIGHTS AND DATA OWNERSHIP

Indicate who will hold intellectual property rights to the data and other information created by the project, and whether these rights will be transferred to another organization for data distribution and archiving. If any copyrighted material (i.e., instruments or scales) are used, how will the project obtain permission to use or disseminate it?

Data archives need a clear statement from the data producer of who owns the data before they can be disseminated. However, issues of data ownership can be complex. For example, principal investigators on federally funded projects are responsible for collecting research data and publishing their research findings, but the resulting research data are typically owned by the institution where the principal investigator is employed.

Funding organizations expect researchers to share their data. Public archives can help universities meet those expectations without requiring a transfer of copyright along with research data. A copy of the research data can be shared publicly through an archive while ownership rights remain with the copyright holder. Agreements to publicly archive data typically grant a repository permission to preserve and disseminate the data.

ACCESS AND SHARING

Indicate how you intend to archive and share your data, and why you have chosen that particular option. Mechanisms for archiving and sharing include:

- **Domain repositories,** such as ICPSR (social science)
- **Self-dissemination** through a dedicated Web site created by the research team. Options for eventual dissemination should be arranged through an established archive after the self-dissemination period ends. A schedule of when dissemination will be turned over to a third party should be included. The archive may want to make a preservation copy during the period of self-dissemination for a number of reasons: (1) to develop expertise with the data; (2) to process the data while knowledgeable staff are available; and (3) for general safekeeping.
- **Preservation with delayed dissemination**, in which the data producer arranges with a public data repository for archival presentation with dissemination to occur at a later date, usually within a year. 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.
- **Institutional repositories** at academic institutions, which have the goal of preserving and making available some portion of the academic work of their students, faculty, and staff. Not all such repositories have the capacity to accept and curate data. There are generally two types of institutional repositories: those with a focus on a particular discipline, and those without. Each type provides certain benefits and drawbacks for data producers and users that should be considered when deciding which to use.
- **Restricted-use collections**. In cases in which masking of sensitive data would lessen the analytic power of a dataset, a restricted-use release may be appropriate. Access to restricted-use data can be limited to approved researchers under controlled conditions. Some archives can provide both restricted-use and public-use releases, where the public files have been altered to prevent disclosure of sensitive information about survey participants. See Phase 5, Preparing Data for Sharing, for more on protecting respondent confidentiality.

Other Considerations

Sharing data helps advance science and maximize research investment. Recent research has found that when data are shared through an archive, research productivity is enhanced and the number of publications based on the data is dramatically increased (Pienta, 2010). Experience also has shown that the durability of the data improves and the cost of processing and preservation decreases when data deposits are timely. It is important that data be deposited while the producers are still familiar with the dataset and able to transfer their knowledge fully to the archive.

AUDIENCE

The grant proposal should specify the likely users (academic or nonacademic) of the datasets. 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. This will potentially influence how the data are managed or shared.

SELECTION AND RETENTION PERIOD

Describe how data will be selected for archiving, how long they will be held, and plans for eventual transition or termination of the data collection in the future.

ARCHIVING AND PRESERVATION

Describe how the data will be preserved for the long term. Digital data need to be actively managed over time to ensure they are always available and usable. Digital content requires ongoing preservation action to remain readable, understandable, and meaningful. Depositing data resources with a trusted digital archive can ensure they are curated and handled according to good practices in digital preservation.

ETHICS AND PRIVACY

If applicable, indicate how you will handle informed consent with respect to informing respondents that the personal information they provide will remain confidential when data are shared or made available for secondary analysis. This may mean describing:

- Plans to obtain Institutional Review Board approval
- Any legal constraints on sharing data such as HIPAA
- Methods of managing disclosure risk

INFORMED CONSENT

Generally speaking, informed consent agreements and confidentiality should be considered as early as possible in the research process. Protection of individuals' privacy is a core tenet of responsible research practice, and must be thoroughly addressed.

"Informed consent" refers to the communication process that allows individuals to make informed choices about participating in a research study. An informed consent agreement provides required information about the study and serves as a formal agreement by an individual to willingly participate in the proposed research. A description of how participant confidentiality will be protected must be included in an informed consent agreement.

Language in an informed consent agreement giving the research team exclusive access to the data or promising that the data will only be shared in aggregate form or statistical tables could make archiving and disseminating the data more difficult later. Disclosure protection methods can guard sensitive information while preserving the analytic power of a dataset, rendering such restrictive language in informed consent agreements unnecessary. Two examples of non-restrictive statements on confidentiality are given below:

> Sample 1. We will make our best effort to protect your statements and answers, so that no one will be able to connect them with you. These records will remain confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study. Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public.
>
> Sample 2. The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study.

Optional Elements

BUDGET

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. Some potentially costly activities are listed below:

- For quantitative data, investigators should allocate resources 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.
- 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.
- 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.

DATA ORGANIZATION

Describe how the data will be managed during the project, including information about version control, naming conventions, etc. Indicating how your data may be different than the norm will help other researchers during secondary analysis. For example, if the data is dynamic, version control would be central to how the data will be used and understood by the research community.

QUALITY ASSURANCE

Describe procedures for ensuring data quality during the project.

LEGAL REQUIREMENTS

A listing of all relevant federal or funder requirements for data management and data sharing.

Summary Table: Elements of a Data Management Plan

RECOMMENDED ELEMENTS

Element	Description	NSF Mapping
Data description	A description of the information to be gathered; the nature and scale of the data that will be generated or collected.	Expected Data
Existing data	A survey of existing data relevant to the project and a discussion of whether and how these data will be integrated.	Expected Data
Format	Formats in which the data will be generated, maintained, and made available, including a justification for the procedural and archival appropriateness of those formats.	Data Format and Dissemination
Metadata	A description of the metadata to be provided along with the generated data, and a discussion of the metadata standards used.	Data Format and Dissemination
Storage and backup	Storage methods and backup procedures for the data, including the physical and cyber resources and facilities that will be used for the effective preservation and storage of the research data.	Data Storage and Preservation of Access
Security	A description of technical and procedural protections for information, including confidential information, and how permissions, restrictions, and embargoes will be enforced.	Data Format and Dissemination
Responsibility	Names of the individuals responsible for data management in the research project.	Roles and Responsibilities
Intellectual property rights	Entities or persons who will hold the intellectual property rights to the data, and how IP will be protected if necessary. Any copyright constraints (e.g., copyrighted data collection instruments) should be noted.	Data Format and Dissemination
Access and sharing	A description of how data will be shared, including access procedures, embargo periods, technical mechanisms for dissemination and whether access will be open or granted only to specific user groups. A timeframe for data sharing and publishing should also be provided.	Data Storage and Preservation of Access
Audience	The potential secondary users of the data.	Data Format and Dissemination
Selection and retention periods	A description of how data will be selected for archiving, how long the data will be held, and plans for eventual transition or termination of the data collection in the future.	Data Format and Dissemination
Archiving and preservation	The procedures in place or envisioned for long-term archiving and preservation of the data, including succession plans for the data should the expected archiving entity cease to exist.	Data Storage and Preservation of Access
Ethics and privacy	A discussion of how informed consent will be handled and how privacy will be protected, including any exceptional arrangements that might be needed to protect participant confidentiality, and other ethical issues that may arise.	Data Format and Dissemination

Summary Table: Elements of a Data Management Plan

OPTIONAL ELEMENTS

Element	Description
Budget	The costs of preparing data and documentation for archiving and how these costs will be paid. Requests for funding may be included.
Data organization	How the data will be managed during the project, with information about version control, naming conventions, etc.
Quality Assurance	Procedures for ensuring data quality during the project.
Legal requirements	A listing of all relevant federal or funder requirements for data management and data sharing.

"NSF Mapping" refers to guidelines from the agency available at www.nsf.gov/bfa/dias/policy/dmp.jsp.

An Example Data Management Plan for Depositing Data with ICPSR

This sample plan is provided to assist grant applicants in creating the required data management plans. Researchers should feel free to edit and customize this text before submission. Note that a letter of commitment from ICPSR confirming that it will archive the data should accompany the plan. Please contact ICPSR Director of Acquisitions Amy Pienta, apienta@umich.edu, to request such a letter.

Data Description — [Provide a brief description of the information to be gathered — the nature, scope, and scale of the data that will be generated or collected.] These data, which will be submitted to ICPSR, fit within the scope of the ICPSR Collection Development Policy. A letter of support describing ICPSR's commitment to the data as they have been described is provided.

Designated Archive — The research data from this project will be deposited with the digital repository of the Inter-university Consortium for Political and Social Research (ICPSR) to ensure that the research community has long-term access to the data. The integrated data management plan proposed leverages capabilities of ICPSR and its trained archival staff.

Access and Sharing — ICPSR will make the research data from this project available to the broader social science research community. *Public-use data files*: These files, in which direct and indirect identifiers have been removed to minimize disclosure risk, may be accessed directly through the ICPSR Web site. After agreeing to Terms of Use, users with an authorized IP address from a member institution may download the data, and non-members may purchase the files. *Restricted-use data files*: These files are distributed in those cases when removing potentially identifying information would significantly impair the analytic potential of the data. Users (and their institutions) must apply for access to these files, create data security plans, and agree to other access controls. *Timeliness*: The research data from this project will be supplied to ICPSR before the end of the project so that any issues surrounding the usability of the data can be resolved. Delayed dissemination may be possible. The Delayed Dissemination Policy allows for data to be deposited but not disseminated for an agreed-upon period of time (typically one year).

Metadata — Substantive metadata will be provided in compliance with the most relevant standard for the social, behavioral, and economic sciences — the Data Documentation Initiative (DDI). This XML standard provides for the tagging of content, which facilitates preservation and enables flexibility in display. These types of metadata will be produced and archived:

- *Study-Level Metadata Record*. A summary DDI-based record will be created for inclusion in the searchable ICPSR online catalog. This record will be indexed with terms from the ICPSR Thesaurus to enhance data discovery.
- *Data Citation with Digital Object Identifier (DOI)*. A standard citation will be provided to facilitate attribution. The DOI provides permanent identification for the data and ensures that they will always be found at the URL specified.
- *Variable-Level Documentation*. ICPSR will tag variable-level information in DDI format for inclusion in ICPSR's Social Science Variables Database (SSVD), which allows users to identify relevant variables and studies of interest.
- *Technical Documentation*. The variable-level files described above will serve as the foundation for the technical documentation or codebook that ICPSR will prepare and deliver.
- *Related Publications*. Resources permitting, ICPSR will periodically search for publications based on the data and provide two-way linkages between data and publications.

Intellectual Property Rights — Principal investigators and their institutions hold the copyright for the research data they generate. By depositing with ICPSR, investigators do not transfer copyright, but instead grant permission for ICPSR to redisseminate the data and to transform the data as necessary to protect respondent confidentiality, improve usefulness, and facilitate preservation.

Ethics and Privacy — *Informed consent*: For this project, informed consent statements, if applicable, will not include language that would prohibit the data from being shared with the research community. *Disclosure risk management*: The research project will remove any direct identifiers in the data before deposit with ICPSR. Once deposited, the data will undergo procedures to protect the confidentiality of individuals whose personal information may be part of archived data. These include: (1) rigorous review to assess disclosure risk, (2) modifying data if necessary to protect confidentiality, (3) limiting access to datasets in which risk of disclosure remains high, and (4) consultation with data producers to manage disclosure risk. ICPSR will assign a qualified data manager certified in disclosure risk management to act as steward for the data while they are being processed. The data will be processed and managed in a secure non-networked environment using virtual desktop technology.

Format — *Submission*: The data and documentation will be submitted to ICPSR in recommended formats. *Access*: ICPSR will make the quantitative data files available in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, and Stata. Documentation will be provided as PDF. *Preservation*: Data will be stored in accordance with prevailing standards and practice. Currently, ICPSR stores quantitative data as ASCII along with setup files for the statistical software packages, and documentation is preserved using XML and PDF/A.

Archiving and Preservation — ICPSR is a data archive with a 50-year track record of preserving and making data available over several generational shifts in technology. ICPSR will accept responsibility for long-term preservation of the research data upon receipt of a signed deposit form. This responsibility includes a commitment to manage successive iterations of the data if new waves or versions are deposited. ICPSR will ensure that the research data are migrated to new formats, platforms, and storage media as required by good practice in the digital preservation community. Good practice for digital preservation requires that an organization address succession planning for digital assets. ICPSR has a commitment to designate a successor in the unlikely event that such a need arises.

Storage and Backup — Research has shown that multiple locally and geographically distributed copies of digital files are required to keep information safe. Accordingly, ICPSR will place a master copy of each digital file (i.e., research data files, documentation, and other related files) in ICPSR's Archival Storage, with several copies stored with partner organizations at designated locations and synchronized with the master.

[NOTE: A version of this document in Microsoft Word is available here.]

Importance of Good Data Management

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 an archive. 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.

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 questionnaire construction or analysis plans.

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?

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? For example, archives increasingly use checksums and other techniques for ensuring integrity.

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? (See Phase 3, Data Collection and File Creation, for guidance in using a standards-based approach to documentation production.)

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

Project documentation. What steps will be taken to document decisions 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 archives are prepared to include Web-based information in deposits.

INTEGRATING DATA AND DOCUMENTATION

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.

Computer-assisted interviewing. 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 in an archive; 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.

Using integrated software. Most large-scale data collection efforts now involve computer-assisted interviewing, but there are still situations in which data entry will be required — e.g., inputting of administrative records, observation data, or open-ended question responses. 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.

Spreadsheets and databases. 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. See Phase 3 for a discussion of using the Data Documentation Initiative (DDI) metadata specification to produce documentation in eXtensible Markup Language (XML) format. The DDI standard provides a way to produce comprehensive documentation that is consistent in format and thus easy to integrate into larger systems.

DATA ENTRY AND DOCUMENTATION AS PART OF PRETESTS AND PILOT STUDIES

Conducting pretests or pilot studies is a good way to uncover potential problems with all aspects of a project. There are two major reasons to include both data entry and documentation as part of 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. See the "Important documentation elements" section in Phase 3: Data Collection and File Creation for essential codebook components.

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, GIS, qualitative, 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. 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. 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.

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. 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.

See the "File structure" section in Phase 4: Data Analysis for more information on best practice in setting up files with different record types.

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.

No attempt is made here to provide standardized coding schemes for all variables. However, the U.S. Census Bureau occupation and industry codes and the National Institute of Standards and Technology's state, county, and metropolitan area codes (also known as Federal Information Processing Codes or FIPS) are standard schemes used to code these types of information. 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 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.

Figure 2 presents an example of the use of this type of scheme for coding parental employment status, from the 1990 Census of Population and Housing Public Use Microdata Samples (PUMS) person record. The first digit of the scheme describes the number of parents present in the household; the second indicates the employment status of parents; the third tells whether employed parents work full- or part-time.

FIGURE 2

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

Living with two parents:

Both parents in labor force:

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

Father only in labor force:

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

Mother only in labor force:

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

Living with one parent:

Living with father:

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

Living with mother:

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

MISSING DATA

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.

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

2. *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).

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

4. *Not Applicable.* The subject was never asked 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.

5. *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.

6. *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.

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.)

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 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). 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 (see later section on restricted-use datasets).

Geospatial data. 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. This is encouraged even if the boundary file is easily obtained from the U.S. Census Bureau or from a known third party, and is absolutely necessary if the original spatial analysis used specially created zones. 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.

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
- 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.

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 or events involving the criminal justice system, 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 archive staff do not make unnecessary changes to the investigator's modifications when performing their 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 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.

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, so you should seek an archive that meets your requirements.

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.

DEPOSITING DDI METADATA

ICPSR encourages the deposit of DDI metadata with deposits of research data. There are currently two main versions of the DDI specification — DDI Codebook (Version 2.*) and DDI Lifecycle (Version 3.*). Most archives will prefer or at least readily accept documentation submitted in either of the DDI versions. To be in full compliance, a document should have question text integrated into each variable.

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

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

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

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

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 [Dublin Core ~ 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.

Substantive, temporal, and geographic coverage of the data collection [Dublin Core Coverage]. Descriptions of topics covered, time period, and location.

Data source(s) [Dublin Core ~ Source]. 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:

1. The exact question wording or the exact meaning of the datum. Sources should be cited for questions drawn from previous surveys or published work.

2. 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.

3. 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.

4. 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.

5. 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.

6. Unweighted frequency distribution or summary statistics. These distributions should show both valid and missing cases.

7. Imputation and editing information. Documentation should identify data that have been estimated or extensively edited.

8. 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.

9. 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.

10. 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.

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 hardcopy questionnaire is available.

Index or table of contents. A list of variables either in alphabetic order or organized into variable groups with corresponding page numbers or links to the variables in the technical documentation or codebook.

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

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

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

Data Analysis

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.

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.

DATA AND DOCUMENTATION VERSIONING

One way to keep track of changes is to maintain explicit versions of a dataset. The first version might result 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.

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. For example, a curly non-ASCII apostrophe in the text string 'Respondent's Age' is read in binary with a different ASCII code than a straight ASCII apostrophe in 'Respondent's Age'.

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 should 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.

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.

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. 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.

Preparing Data for Sharing

This chapter addresses the critical final steps researchers should undertake in preparing to archive and/or disseminate their data. We also provide information on ways to access and analyze archived confidential data.

Respondent Confidentiality

Much of this guide has focused on data preparation methods that can serve the research needs of both principal investigators and analysts of secondary data. In this section, 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 submitting their data for archiving. But first, a quick review of why this is important.

THE PRINCIPLES OF DISCLOSURE RISK LIMITATION

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.

Institutions conducting research using human subjects funded by the federal department of Health and Human Services are responsible for compliance with the federal regulation on Protection of Human Subjects (45CFR46). Every such university and research institution must file an "assurance of compliance" with the HHS Office for Human Research Protections that includes "a statement of ethical principles to be followed in protecting human subjects of research." For more information, see www.hhs.gov/ohrp.

College or university Institutional Review Boards (IRBs) approve proposals for research involving human subjects and take actions to ensure that any research is carried out appropriately and without harming research participants.

Archives place a high priority on preserving the confidentiality of respondent data and review all data collections they receive to ensure that confidentiality is protected in the public-use datasets released. Two major concerns govern policy and practice in this area: professional ethics and applicable regulations. The social sciences broadly defined (as well as a number of professional associations) have promulgated codes of ethics that require social scientists to ensure the confidentiality of data collected for research purposes. (See, for example, the American Statistical Association's "Ethical Guidelines for Statistical Practice," [1999] which stresses the appropriate treatment of data to protect respondent confidentiality.) Both the rights of respondents and their continued willingness to voluntarily provide answers to scientific inquiries underlie this professional ethic. The ethic applies to all participants in the research enterprise, from data collectors to archivists to secondary analysts who use such data in their research.

Regulations also bind all participants in the research enterprise to measures intended to protect research subjects as well as data obtained from such subjects. These regulations range from federal and local statutes to rules instituted by universities and colleges.

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, in the United States, Social Security numbers uniquely identify individuals who are registered with the Social Security Administration. Any variable that functions as an explicit name can be a direct identifier — for example, 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, 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 are often 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?

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 state, county, census tract, census block, telephone area codes, 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 — 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. 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.

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

Data producers can consult with a social science data archive 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 contain 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. Funding agencies, principal investigators, or archive staff may designate data as restricted-use. This decision usually involves consultation among the interested parties. Maintenance of, and approval of access to, a restricted-use file is managed by archive staff in accordance with the terms of access.

Access to restricted-use files is offered under a set of highly controlled conditions to approved researchers. 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. Most agreements require 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 usually 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.

However, the advent of virtual data enclaves (see below) may eliminate the need for physical transfer of data files.

DATA ENCLAVES

In general, the more identifying information there is in a dataset, the more restrictive are the regulations governing access and location of use. Archives grant access to the most confidential data — for example, medical records containing identifying information such as respondent name and address — through a data enclave environment, either physical or virtual.

Virtual data enclaves. These data portals allow users to obtain remote access to restricted data that would not otherwise be available for research. This often includes using a restricted access application system, getting set up with secure remote access to the restricted data (including possible on-site inspection), monitoring research behavior during data access, and having analytic results reviewed for disclosure risk before they are permitted to leave the secure environment. Such systems generally prevent users from emailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally.

Physical data enclaves. A physical data enclave is a secure data analysis laboratory that allows access to the original data in a controlled setting. Secure data enclaves have added security features to ensure the safekeeping of the most confidential data. They typically have appropriate physical security measures (no windows, video monitoring, key card entry) to strictly control access. Their computing environments are not connected to the Internet, but rather have their own network server (connected to a small number of work stations). Researchers using the enclave are monitored by archive staff who see to it that no unauthorized materials are removed. Any analyses produced are scrutinized to determine that they do not include any potential breaches of confidentiality. Other policies and procedures also govern the use of restricted data in enclaves.

Secure Survey Documentation and Analysis (SSDA). SSDA is an online data analysis program that performs bi-variate cross-tabulation, comparison of means, correlation, and regression analyses. SSDA is designed to provide a safe, reliable way to distribute restricted-use data publicly, thereby democratizing access to data that was previously unavailable or required special procedures to obtain. SSDA automates several disclosure protections that prevent the use of organization-defined high-risk variables, singularly or in combination, and restrict types of output commonly associated with disclosure risk (e.g., small un-weighted sample sizes). When the organization-defined rules are violated by an attempted analysis, the resulting output is completely or partially suppressed.

In addition to adhering to the specific requirements of a data archive, data creators who intend to deposit their data should be aware of the OAIS Reference Model standard for what to deposit. See ICPSR's Digital Preservation site regarding the Submission Information Package for more information. The SIP includes a deposit form, the original files, and associated study-level and variable-level metadata (e.g., codebooks).

When preparing data for final deposit in a data archive, it is important to consider the factors detailed below. Since the following list is not all-inclusive, data creators should watch for developments relating to the digital preservation of social science data.

File Formats

If a dataset is to be archived, it must be organized in such a way that others can read it. Ideally, the dataset should be accessible using a standard statistical package, such as SAS, SPSS, or Stata. Three common approaches to data file preparation are: (1) provide the data in raw ASCII format, along with setup files to read them into standard statistical programs; (2) provide the data as a system file within a specific analysis program; or (3) provide the data in a portable file produced by statistical program. Each of these alternatives has its advantages and disadvantages.

SOFTWARE-SPECIFIC SYSTEM FILES

System files are compact and efficient, and archives increasingly encourage the deposit of system files and use this format for dissemination. Older system files may not always be cross-platform compatible, however. Newer versions of statistical software packages not only incorporate new data management and analytical features, but may also support new operating systems and hardware. In such cases, previous versions of system files may need to be migrated to newer versions. To prepare system files, consult the user manual for the statistical software of your choice.

PORTABLE SOFTWARE-SPECIFIC FILES

Some archives prefer to receive data in portable or transport file format. One advantage of portable versions of software-specific files is that they can be accessed on any hardware platform. SPSS calls transportable files "portable," and SAS calls them "transport" files, while for Stata data files, no portable equivalent is necessary. However, users should be careful to preserve missing data values. It should be noted that when SAS transport files are generated, missing data are blanked out unless SAS alpha missing codes are used. This can be a problem because the distinctions between different types of missing data (such as legitimate skip vs. refused) become irretrievable, and they may be very important to the secondary analyst. When preparing SAS transport files, it is recommended that the missing data command not be activated but that separate program files be created instead. SPSS maintains the original missing data values when creating portable files. Stata allows the user to assign alpha missing values, and since no separate transportable files are created, alpha missing values are not affected.

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

ASCII DATA PLUS SETUP FILES

For this option, it is necessary to determine which syntax will be used — SAS, SPSS, Stata, or another statistical program. In the case of large datasets, for which users will want to create subsets, the setup files can be edited to meet specific needs. Many archives view ASCII (raw) data files as the most stable format for preserving data. They are software-independent, and hence are apt to remain readable in the future, regardless of changes in particular statistical software packages. Most archives are capable of producing ASCII data and setup files from data files provided in proprietary formats. If a researcher has maintained the dataset in ASCII and read it into a statistical package for analysis, a "raw" ASCII data file may be the most cost-efficient way to archive the data.

Writing an ASCII file can be time-consuming and prone to error, even when a software system has been used to store the data. For example, if SAS has been used to manage and analyze a dataset, the following steps are required: writing SAS statements to export the data in ASCII format, careful checking to make sure the conversion procedure worked properly, and documentation telling users where to find variables in the ASCII data file.

ONLINE ANALYSIS-READY FILES

Online data exploration and analysis packages allow users not only to perform analysis online, but to select only those variables and cases actually required for an analysis in the form of subsets. Increasingly, these systems accept DDI XML as input. Depositing documentation in DDI facilitates online analysis after archival deposit.

OTHER FILE FORMATS

Video files. Video file formats are changing rapidly. These changes bring improvements to quality and flexibility while reducing the size of the compressed file. It is essential to deposit the source files for video data, along with the compressed files, to ensure the long-term playability of video files. An archive is able to migrate video data to better formats over time with the source file, which maintains all the original captured information. The compressed file, in comparison, already has much of the original captured information removed. An archive is much more limited in its ability to ensure playability over time without access to the source file, because the technologies of the future may require file information not kept in a compressed file.

Technical information, such as the video file type, compression format, and video source should be included, as well as a written summary of what the video contains. Ideally, the summary should have time codes indicating where significant events occur in the video. A text file with time-coded subtitles for hearing-impaired viewers would also be valuable to include.

Geospatial data files. 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. This is encouraged even if the boundary file is easily obtained from the U.S. Census Bureau or from a known third party, and is absolutely necessary if the original spatial analysis used specially created zones. 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.

Archiving Files from Analysis of Existing or Secondary Data

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

- **Existing data not publicly available.** If the existing data used for analysis are not already publicly available, researchers are encouraged to submit the data for archiving, with the permission of the original data producer.
- **Combination of primary and existing data.** If the researcher collects some primary data and also appends existing data to it for analyses, the guiding questions on whether to submit the whole dataset or just the primary dataset to the archive are: (a) how easily the existing data can be linked to the primary data, and (b) whether the existing data are publicly available.

LINKED DATA

Linked census data. When the primary data are linked to census data, the linked census data should be archived also, even though the link between the data files is straightforward and the census data are publicly available. Since the original census files are large in size and contain a large number of variables, determining which census variables to use and at what level to extract the data for the subsets can be time-consuming. Archiving the linked census data makes it unnecessary for other users to repeat these subsetting steps.

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

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

Derived variables. Often, after the data are linked, the researcher may compute new variables based on the linked data (e.g., new categories are created, rates are produced, or scales are developed). All useful derived variables should be archived also, especially if they are used in analyses included in publications. The derived variables may be deposited in a data file that includes the primary data collected for the project, the existing data from another source, or the derived variables may be deposited with the primary data alone. The code or setup file used to link the files and create the derived variables should also be provided.

Programming code. If the project involves only analysis of data already publicly available and the product of the project is the analysis alone, then data may not need to be submitted for archiving. However, researchers are encouraged to deposit their programming code that created new variables or scales, especially if the derived variables are not deposited within a data file and are cited in publications.

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. http://www.amstat.org/about/ethicalguidelines.cfm (accessed Nov. 15, 2011).

Babbie, Earl (1990). *Survey Research Methods*. 2nd ed. Belmont, CA: Wadsworth [pp. 209-211].

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. http://ssc.sagepub.com/cgi/content/abstract/22/3/307 (accessed Nov. 15, 2011).

Council on Government Relations (2006). "Access to and Retention of Research Data Rights and Responsibilities." http://www.cogr.edu/viewDoc.cfm?DocID=151536 (accessed Nov. 15, 2011).

Data Archiving and Networked Service (DANS) (2008). "Data Seal of Approval, An Overview." http://datasealofapproval.org/sites/default/files/DSA_informationfolder_new_2011.pdf (accessed Nov. 15, 2011).

ESDS Qualidata (2007, September). "Reusing Qualitative Data." http://www.esds.ac.uk/qualidata/support/reuse.asp (accessed Nov. 15, 2011).

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. http://www.annualreviews.org/doi/abs/10.1146/annurev.pu.15.050194.000245 (accessed Dec. 14, 2011).

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. http://hdl.handle.net/2027.42/41214 (accessed Nov. 15, 2011).

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: Space data and information transfer systems — Open archival information system – Reference model." International Organization for Standardization, Geneva, Switzerland. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683 (accessed Dec. 14, 2011)

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 & Politics*, 28(3): 443-499.

King, Gary. 2006. "Publication, Publication." *PS: Political Science & Politics*, 39(1): 119-25.

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." http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed Nov. 15, 2011).

National Institutes of Health, Office of Extramural Research (2003, March 5). "NIH Data Sharing Policy and Implementation Guidance" http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm (accessed Nov. 15, 2011).

National Science Board, National Science Foundation (2005, September). *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_10.pdf (accessed Nov. 15, 2011).

National Science Foundation (1989). "NSF Important Notice 106." http://www.nsf.gov/pubs/stis1996/iin106/iin106.txt (accessed Nov. 15, 2011).

National Science Foundation (2010). Directorate for Social, Behavioral and Economic Sciences, "Data Management for NSF SBE Directorate Proposals and Awards." http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf (accessed Nov. 15, 2011).

National Science Foundation (2011). Directorate for Social, Behavioral and Economic Sciences, "Data Archiving Policy." http://www.nsf.gov/sbe/ses/common/archive.jsp (accessed Nov. 15, 2011).

O'Rourke, JoAnne McFarland (2003, Fall). "Disclosure Analysis at ICPSR." *ICPSR Bulletin*, Vol. 34(1):3–9. http://www.icpsr.umich.edu/files/ICPSR/org/publications/bulletin/2003-Q3.pdf (accessed Nov. 15, 2011).

O'Rourke, JoAnne McFarland, Stephen Roehrig, Steven G. Heeringa, Beth Glover Reed, William C. Birdsall, Margaret Overcashier, 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). http://www.jstor.org/pss/10.1525/jer.2006.1.3.63 (accessed Nov. 15, 2011).

Pienta, Amy, George Alter, and Jared Lyle (2010). "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data". Presented at the Organization, Economics and Policy of Scientific Research workshop, Torino, Italy, April 2010. http://hdl.handle.net/2027.42/78307 (accessed Nov. 15, 2011).

United States Department of Health and Human Services. Office for Human Research Protections (2010, Aug. 30). "What are the key features of the Federalwide Assurance?" http://answers.hhs.gov/ohrp/questions/7145 (accessed Nov. 15, 2011).

United States Department of Health and Human Services. Office for Human Research Protections (2009, July 14). Human Subjects Research, Title 45 Code of Federal Regulations Part 46 http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html (accessed Nov. 15, 2011).

University of Michigan, Human Research Protection Program. Office of the Vice President for Research. "Operations Manual." http://www.hrpp.umich.edu/om/ (accessed Nov. 15, 2011).

Zelenock, Tom, and Kaye Marz (1997). "Archiving Social Science Data: A Collaborative Process." *ICPSR Bulletin*, 17(4): 1–4. http://www.icpsr.umich.edu/files/ICPSR/org/publications/bulletin/1997-05.pdf (accessed Nov. 15, 2011)