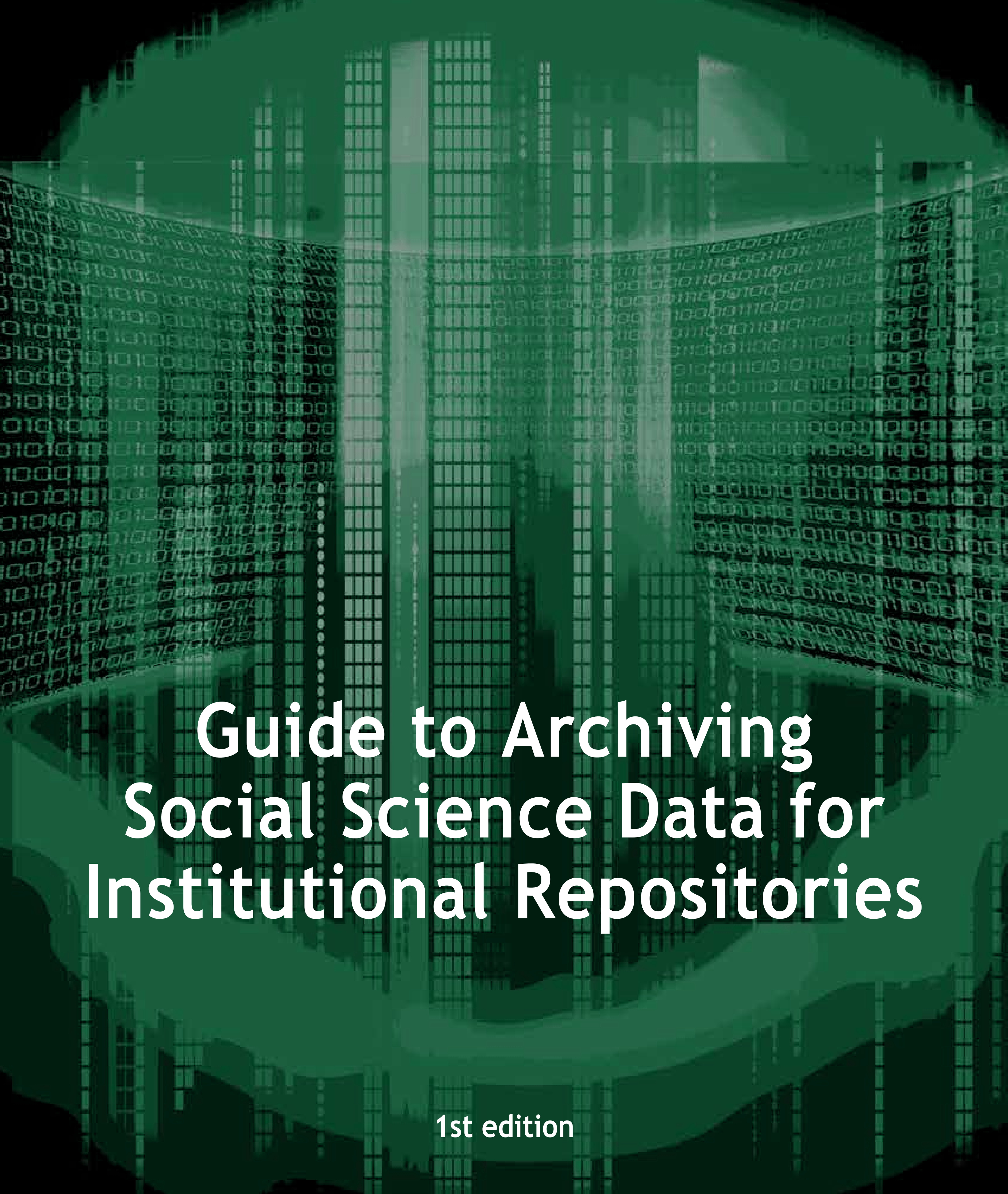
ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH



Guide to Archiving Social Science Data for Institutional Repositories

1st edition

Copyright © 2012 by the Inter-university Consortium for Political and Social Research (ICPSR)

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

First edition, 2012.

Acknowledgements
The *Guide to Archiving Social Science Data for Institutional Repositories* was created with support from the Institute of Museum and Library Services (IMLS). ICPSR staff Jared Lyle, George Alter, Amy Blau, Dan Meisler, and Jenna Tyson prepared this version, and thank the following individuals for their contributions: Jake Carlson, Jennifer Darragh, Sarah Shreeves, Libbie Stephenson, and Ryan Womack.

Suggested Citation
Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to Archiving Social Science Data for Institutional Repositories* (1st ed.). Ann Arbor, MI.

Guide to Archiving Social Science Data for Institutional Repositories

Overview

Improvements in data processing and storage technology have resulted in an increased production of research data on a variety of social, economic, and political subjects. Many datasets could be profitably reanalyzed, but they are at danger of being lost since they are never properly archived. Institutional repositories (IRs) provide local, personalized services and are playing a role in preserving these data collections, although some repositories do not feel prepared to curate and archive data.[1] This guide provides a detailed overview of what an institutional repository can do to appraise data, prepare them for storage, and ensure that the preserved data are independently understandable.[2]

Scope

This Guide is focused on the kind of material most often collected as part of a quantitative approach to research, conducted in social sciences disciplines. This includes surveys, enumerations, vital statistics and sometimes administrative records, formatted to enable research using statistical analysis software, such as SPSS, STATA, SAS, R, etc. This kind of research data has particular needs with respect to curation and archiving, beyond and different from the kinds of techniques typically used in institutional repository systems. Specific standards and best practices that have been adopted by other practitioners will be important for the IR to carry out in managing these collections. Other forms of 'data' are used in research, however the scope of this Guide is to highlight the particular needs of datasets created using quantitative research methods. The Guide uses terminology specific to this kind of material. An appendix[3] in this document includes a glossary of terms and their definitions.

Data 101: What are data?

What is meant by "data"? According to one definition, data are "a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen."[1] Data may be captured in many different ways: through sensors that read temperatures or other environmental characteristics; through measurements obtained by experimental observation and analysis; through (automated) logging of (digital) transactions; through survey or census results; or through administrative records. Quantitative social science data often take the form of tabular or delimited raw data files, which may be read by statistical software and external documentation (codebooks) that explain the format and structure of the raw data. [1]

[1] For a discussion of the role institutional repositories can play in preserving research data, especially by partnering with domain repositories and researchers, see: Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services*, *23*(1), 35-53. http://dx.doi.org/10.1108/10650750710720757

[2] While this guide focuses on the steps an institutional repository takes to acquire, curate, and preserve data collections, other guides exist that address other aspects of data archiving. JISC's *Policy-making for Research Data in Repositories: A Guide* (http://www.disc-uk.org/docs/guide.pdf) offers a policy decision-making framework for institutions planning or implementing digital services for data management and sharing. ICPSR's *Guide to Social Science Data Preparation and Archiving: Best Practices Throughout the Data Life Cycle (5th ed.)* (http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf) helps data producers who are engaged in the research process prepare data for deposit in an archive; the UK Data Archive's *Managing and Sharing Data: Best Practices for Researchers (3rd ed.)* provides similar guidance. The Digital Curation Centre's *How-to Guides* (http://www.dcc.ac.uk/resources/how-guides) "provide working knowledge of curation topics, aimed at people in research or support posts who are new to curation, but are taking on responsibilities for managing data, whether at local research group level or in an institutional data centre/ repository."

As a social science data archive, the Inter-university Consortium for Political and Social Research (ICPSR) has specialized in data curation and stewardship for a wide range of social science, behavioral science, and related disciplines since 1962. Recommendations in this guide are based on best practices used by ICPSR.

Teamwork

Working with data is a team effort, especially when you are unfamiliar with the data type or content. Effective data curation requires knowledge and documentation of the data collection methods, data formats, and systems on which data were produced and can be read. Fortunately, there are many resources available to your institutional repository. These include data librarians, social science faculty, librarian research subject specialists, and data archivists at social science data archives.[4] As you follow the below steps to review and archive data, it may be useful to build your own support team.

As part of a team approach to data curation, it is helpful to build good relationships with faculty from the very beginning of a project. Informing faculty about best practices in documentation, assisting with metadata, and making faculty aware of the services and support available through the IR will help ensure that data are stored in an easily curated format. Early involvement may also open avenues to support data curation activities at the IR through grant funding processes; researchers will benefit from early consultations that inform their data management plans.

Appraising the study materials

The first step in the curation and archiving process ideally begins even before data are received by an institutional repository. It is useful to discuss the data deposit process with the investigator to ensure that both the researcher and the IR have a common understanding of the steps involved, and the roles and responsibilities of each. Collections submitted to a repository should be initially reviewed to see if they fit within the repository's collection development policy, as well as appraised to determine their permanent value. Not all data collections are worth preserving. It is important that the IR develop a policy on the types of data that will be managed and the extent to which they will be curated. There are tools which can be used to develop such a policy, such as *Policy-making for Research Data in Repositories: A Guide*[5]. Understanding what should be included in a full dataset or a codebook will help the IR to appraise the materials being offered for deposit.[6]

Appraisal standards are organized around a repository's priorities. At ICPSR, for instance, data with the following qualities are given greater value for acquisition and preservation:

- Nationally representative
- Theoretically/methodologically unique
- Representing underrepresented research populations
- Widely cited, appearing in top tier journals, or collected by an eminent scholar

[3] See Appendix A: Glossary of Terms

[4] See, for example, "Librarians and Cross-Sector Teamwork", by Joan Lippincott (http://www.cni.org/wp-content/uploads/2011/07/team.pdf)

[5] See: http://www.disc-uk.org/docs/guide.pdf

[6] See Appendix A: Glossary of terms

In addition, data and documentation quality, format, and legal issues should be considered. Many archives also apply tailored appraisal criteria. The Data-PASS partnership of social science data archives specifies its core appraisal questions as:[7]

- How significant are the data for research?
- How significant is the source and context of the data, particularly in regard to scientific progress and society?
- Is the information unique?
- How usable are the data?
- What is the timeframe covered by the information?
- Are the data related to other data in the archives?
- What are the cost considerations for long-term maintenance of the data?
- What is the volume of data?

Another approach to appraisal is to assess the risk of future loss – with riskier data collections given higher priority to acquire and preserve. Are the data from a professor about to retire? Are the data in a software format (e.g., Wordstar) or media type (e.g., 5¼-inch floppy disk) that is becoming obsolete? Are the data part of a grant or project that is losing funding or institutional support?

Understanding the study materials

After appraisal, the next crucial steps in the archiving process are understanding, finding, and selecting study data and documentation. Without the full dataset, future research is limited. Without documentation, context is lost and data cannot be interpreted. The IR collection policy should provide details about file formats and documentation parameters needed for preservation, and be used to assess whether or not a particular study can not only be archived but also reused by others not familiar with the original.

Understanding the research project goals, methodology, and procedures for data collection and data management helps define what to look for and what might be found. For example, knowing that a study used punched cards to enter and then analyze 19th Century Boston voting records provides useful clues about what data to look for (i.e., punched cards) and what documentation might be needed to understand the data (e.g., original census forms, codebooks describing how census responses were classified and numbered on the punched cards).

Ideally, an institutional repository will be able to contact and work directly with the investigators and research associates who initially collected and analyzed the data. The original staff are important resources to gathering old study materials into a cohesive collection for archiving and preservation. Questions you can ask these people that will help you understand the study materials include:

- Is there a grant application, publication, or other document that describes the project as a whole?
- Is there a study description or abstract of the study that describes the theoretical framework that informs the study, research questions addressed by the study, and any specific hypotheses tested?
- If there is no documentation, can the researcher describe what he/she planned to accomplish?

[7] See: http://www.icpsr.umich.edu/files/DATAPASS/pdf/appraisal.pdf

- What sources were used to produce the database? Some data are derived from a single source, like a census, while others are constructed by combining information from a variety of sources.
- What were the steps in data collection? Was information written on paper before it was digitized? Were some variables (like occupations) coded (i.e. classified and numbered)?
- How were the data computerized? What kind of software was used (spreadsheet, database, word processor, statistical package)?
- What changes were made to the data after it was computerized?
- What is the "final" form of the data? Is it in a text file (ASCII) or in a format associated with a specific program (e.g., Excel, SPSS, Access)?
- Were any new computer programs written to generate or analyze the data? Are these programs available for archiving? What kind of system can be used to run them? What kind of output do they generate?

Finding the study materials

The quantity, identifiability, and readability of the digital files produced by a research project vary dramatically. Some projects result in a single data file, while others produce dozens or even hundreds of distinct files. Researchers may also have retained files from various stages of data collection, and there may be duplicates created for backup. Older files might be stored on obsolete external media, such as zip drives, 5 ¼" disks, or even punch cards. In addition to electronic files, many studies have data and documentation in hard-copy format. These might be papers in a filing cabinet or print-outs from an old dot-matrix computer.

Files can be centrally located (e.g., in a researcher's basement) or geographically dispersed. Sometimes the data are discovered on a Web site. Researchers often rely on graduate students or research assistants to handle data management. These people are excellent additional sources for both information about the data collection and copies of the data and documentation.

Data

Data are often stored on computerized or digitized media. However, the below questions also apply for hard-copy data (e.g., census manuscripts). Questions you can ask to help understand and identify the data include:

- How many collections of data does the researcher have? For example, are there logical groupings of data from separate projects, different sources, or types of data collection?
- What type of data are present (e.g., qualitative, quantitative)?
- On what storage media were they stored (e.g., floppy disk, hard drive)?
- In what format were they saved (e.g., Excel, SPSS, proprietary database, text analysis system)?
- Which of these data were used during the final analysis?
- Are the files still readable? For instance, can you still open them if they are stored on a computer? If they are stored on a CD, can the data still be retrieved? If they are in a software dependent format, can you use the data in recent versions of the software? For example, if the data are stored as SPSS 7 system files, can they be used with SPSS 10? Can they be converted to another statistical package, such as STATA? How much work is involved in making the data usable in more current operating systems? For example, a dataset created using a VM system may be difficult to transfer to use in UNIX or Windows.
- Have the data been shared with co-researchers? If so, are there different versions of the data? Are they retrievable?

Documentation

Study documentation is vital for data curation. When possible, work with researchers as their data collection begins to develop good documentation practices and to ensure the resulting material can be well managed. Documentation is often stored in both digital and hard-copy format. Questions you can ask of the researcher or someone with expertise in working with the data, to help understand and identify the documentation include:

- Is there a codebook for the data collection? The codebook describes the data collection, especially how variables and cases are coded, and helps users interpret the data. In older studies, the codebook showed the meaning of every character on each row in a data file. Data stored in statistical packages (SPSS, SAS, Stata), databases (Access, MySQL, Oracle), or spreadsheets (Excel) are organized by variable name, field name, or column.[8]
- Are these data adequately described? For example, if a file includes a row name "LOCATION" and possible numeric responses of 1 to 4, does the file also include a label for the row name, such as "Location of the first recipient", with more labels for responses 1 to 4, such as "North L.A.", "East L.A.", "West L.A.", and "South L.A."?
- If the study included a survey, is there a questionnaire or data collection instrument available? A questionnaire helps the user understand the questions asked during the data collection process.
- Is there other documentation, such as a user guide, that describes how to use the data? Some surveys, for instance, use sampling techniques that are described in draft memos. Publications and memos can also answer questions on: How were the study respondents chosen? What was the response rate? Were certain sub-populations over-sampled? Does the documentation discuss how the sample may be weighted? How should margin of error be handled in analysis?
- Are other resources needed to interpret the data? For example, geographic locations may be coded according to a specific map. Causes of death are often assigned codes from the International Classification of Diseases, which is now in its tenth revision.
- Are there published papers describing the data collection?
- Does the documentation describe how the data were gathered? Was it during face-to-face interviews, mail questionnaire, web survey, telephone survey? Does the documentation show whether or not respondents were asked questions in a random order, or demonstrate the flow of the questions asked?

The big picture

The outcome of these conversations should be an overall picture of the research process that will help you to recognize materials as you find them. You should be able to:

- Identify which data files and documents belong to each collection of data.
- Distinguish between final versions of the data and working files that have been superseded. (Researchers often have multiple versions of data and documentation, which were saved at different stages of their projects.)
- Assess the amount of technical effort required for you to process incoming materials into the archive, including converting the materials to archival formats, or updating software-dependent materials.

[8] See Appendix A: Glossary of Terms

Taking inventory and selecting materials to archive

Your goal is to find a (1) complete copy of each data collection in its final form (i.e. the form used for analysis), (2) documentation that will allow future researchers to understand and analyze the data without communicating with the person who collected the data originally and (3) be able to convert file formats, storage media, and software dependent files into re-usable content. It may be useful to compile an inventory of all the materials that you have found to keep track of each object as you examine it. Once you have identified the final versions of data and documentation, copy and save them for processing and archiving. Redundant or superseded versions of data and documentation should be returned to the investigator or deleted.

Preparing the study materials for archiving

Using the data and documentation you have selected, you can now prepare the material for archiving and preservation.

Data

For the data, if you have access to the statistical software used to open the data file, you can review and edit information yourself. If you do not have access to or are not able to use statistical software, be sure to include someone with this expertise in your team, such as someone from a local data archive, or someone from a statistical consulting group. Some questions to ask when reviewing data include:

- Are variables and values labeled? For example, if you find a variable with the name "Var100" with no label, can you add a label that would accurately describe the variable and add context for the end user? Likewise, do all categorical variables (e.g., "How satisfied are you with the president's recent speech?") have value labels (e.g., "Very Satisfied"…"Very Unhappy")?
- Are there wild or undocumented codes? For example, if a variable that records the sex of a respondent has documented codes of "0" for female and "1" for male, an undocumented code of "7" would be a wild code. Can you find documentation that would explain this wild code? If not, you can add a note to the codebook explaining that the "7" value is undocumented.
- Are there out-of-range codes? For example, a value of "387" would be out-of-range for a variable that records the age of a respondent. Can you find a reason for the "387" value? If not, you can add a note to the codebook explaining that the "387" value is undocumented.
- Are codes in the data valid (i.e., documented) according to the data collection instrument or the original codebook? If not, you can either correct the mistake in the data code or add a note to the codebook explaining which codes are undocumented.
- Are codes in the data reasonable? For example, if date variables do not contain dates or all instances of a certain variable are defined as system missing, you can add a note to the codebook explaining which codes are unreasonable.
- Do the numbers of respondents defined in the documentation match the data? Is missing data clearly identified?

What is preservation? Why is preservation important?

Preservation is making content available in useable and meaningful formats to current and future users. Materials, especially digital files, are susceptible to deterioration, corruption, and loss, although advance action can do a great deal to prevent loss and insure long-term preservation. For an introduction to digital preservation concepts and strategies, see the *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions online tutorial* (http://www.dpworkshop.org/dpm-eng/eng_index.html).

- Are there variables that may identify human research subjects or otherwise pose a confidentiality concern?

Confidential data

Data that may identify human research subjects pose a special challenge to institutional repositories. If the IR is located at a college or university, the IR must comply with all policies set forth by the local office or review board which oversees research on human subjects. Data repositories are obligated to protect individuals' privacy by preventing the release data that reveals individual identities and/or identifying characteristics, or that could lead to deduction of that information. Examples of data types that pose problems for confidentiality:

- Multilevel studies, i.e., studies with linked variables between files pose serious identification risks because of the multilevel information contained in these which often makes it easy to identify individual subjects. A multilevel study, for example, might include data about patients, doctors, clinics, and treatments.
- Studies with many precisely dated (public) events or birth dates.
- Studies with geocoded information (especially when other related, public information is available).
- Qualitative (narrative interview) studies – i.e., studies with detailed information, the very richness of which is simultaneously the value of the study and the threat to confidentiality.

Examples of potentially problematic variables:

- *Direct identifiers* – Public identifiers (of respondents) or data that directly reveal the identity of respondents that may have been obtained in the process of data collection such as personal names, addresses (including ZIP codes), city codes, telephone numbers, social security numbers, driver license numbers, patient numbers, certification numbers, and other individually unique numbers and codes.
- *Indirect identifiers* – Data that reveal the identity of respondents when they are used in combination with other data. They contain variables such as detailed geography (i.e., state, county, or census tract of residence), exact date of birth, places with a population less than 100,000, organizations to which the respondent belongs, educational institution from which the respondent graduated (and year of graduation), exact occupations held, place where the respondent grew up, exact dates of events, detailed income, and offices or posts held by the respondent. Sometimes, indirect identifiers can also become direct identifiers depending on the features of the research design. It can be more challenging to identify indirect identifiers. Careful attention must therefore be paid to interactions among the context of the study, the nature of the sample, and the characteristics of respondents to prevent ordinarily unrevealing information from becoming the pointer to an individual.

Variables with confidential information must be either recoded or deleted. As more and more datasets of all kinds become available online, the risk that datasets may be combined to reveal identities through indirect identifiers also increases. It is therefore important to consider how best to preserve privacy and confidentiality when datasets are made public. Data archives, such as ICPSR, have significant experience handling confidential data and can provide additional resources.

Documentation

For the documentation, you can prepare versions to archive together with the data. ASCII documentation files can remain unchanged. If the documentation is in another electronic format, such as Word or Excel, you can convert them to PDF. If the data are in hard-copy format, you can scan and convert them to PDF.

Study description

In addition to converting and scanning original documentation, you can create a study description, which provides an overview of the entire data collection. While your IR may choose to capture information differently, the study description for ICPSR data collections includes the following information (much of the study description is created from the information submitted through the online data deposit form[9]):

I. Study citation

1. Principal investigator(s) and their respective affiliation(s) at the time of data collection (for multiple investigators, give proper name order).

2. A descriptive title of the data collection, including the time period(s) and geographic location(s) that the data cover.

3. Place of production (city/state) of data collection, date of production, and organizational name of data producer.

4. Sponsoring or funding agency (if applicable) and grant number.

5. Person/organization responsible for collecting data.

6. Special collaborator(s) (if applicable).

II. Study description

Provide an abstract of the study. The abstract should describe the theoretical framework that informs the study, research questions addressed by the study, and any specific hypotheses tested. A useful source of information is the major publications produced by the investigator. These may be part of a CV or bibliography about the data. Grant application information, if available, can also provide an overview or abstract of the study.

III. Study methodology and sampling

Provide an in-depth description of the study sampling and methodology, if available, or answer the individual questions below.

1. Unit of analysis (for example, individuals, households, metropolitan area):

[9] See: https://www.icpsr.umich.edu/cgi-bin/ddf2

2. Source(s) of data (for example, personal interviews, telephone interviews, self-enumerated questionnaires, administrative records):

3. Type of data collection (e.g., survey, aggregate, census/enumeration, experimental, event/transaction, clinical, program source code, machine-readable text, administrative records, etc.):

4. Date(s) the data were collected (provide specific dates and ranges, e.g., month/day/year-month/day/year):

5. Time span covered by the data collection (months/days/years -- include specific dates and ranges):

6. Geographic area(s) to which data are relevant (for example, New York City, Singapore, United States, Springfield [Ohio]):

7. Unit of geographic analysis (for example, Census tract, state, precinct):

8. Describe the universe of the study:

9. Is the data collection one of a series? If so, provide a description of the series:

10. Describe the type of sample(s) obtained. If a complete sampling description is available in a printed publication, cite the complete citation for that publication, or provide a separate file with your submission.

11. Provide the response rate for each sample:

12. Describe the established measurement tools used in your study (for example, MMPI, CPI, DSM-IV, CES-D, SF-36, Addiction Severity Index):

13. Indicate any weights used in your data collection and describe them in detail:

IV. Study bibliography

List all publications describing or resulting from the data collection. You may provide a list in a separate document. Include the full title, full names of author(s), place of publication/publisher, journal name/volume/issue, full date, and page numbers, as applicable: you may find publications in the researcher’s CV or web site.

Conclusion

This guide provides a detailed overview of what an institutional repository can do to appraise data, prepare them for storage, and ensure that the preserved data are independently understandable. By leveraging the expertise on campus and best practices from domain repositories and professional organizations, institutional repositories can make data available within and beyond their campus research communities.

Appendix A. Glossary of Terms

Definitions are taken from the ICPSR Glossary of Social Science Terms.[10] All other references are specified.

ASCII

A character-encoding scheme used by many computers. (ASCII stands for American Standard Code for Information Interchange.)

code

In most numeric data files, answers to questions are recorded with numbers rather than text, and often even numeric answers are recorded with numbers other than the actual response. The numbers used in the data file are called "codes." For example, when a respondent identifies herself as a member of a particular religion, a code of "1" might be used for Catholic, a "2" for Jewish, etc. Likewise, a person's age of 18 might be coded as a 2 indicating "18 or over." The codes that are used and their correspondence to the actual responses are listed in a **codebook**. See also **value**.

codebook

Generically, any information on the structure, contents, and layout of a data file. Typically, a codebook includes: column locations and widths for each variable; definitions of different record types; response codes for each variable; codes used to indicate nonresponse and missing data; exact questions and skip patterns used in a survey; and other indications of the content of each variable. Many codebooks also include frequencies of response. Codebooks vary widely in quality and amount of information included.[11] Here is an example of a codebook at ICPSR.

data[12]

For social science, data is generally numeric files originating from social research methodologies or administrative records, from which statistics are produced.

file

A collection of any form of data that is stored, usually on a computer disk or tape.

qualitative data[13]

Information that is difficult to measure, count, or express in numerical terms.

quantitative data[14]

Information that can be expressed in numerical terms, counted, or compared on a scale.

study

All the information collected at a single time or for a single purpose or by a single principal investigator. A study consists of one or more files. *Examples:* General Social Survey; Gallup Polls; 1990 Census of Population and Housing STF 1A.

text file

In computer usage, any file written in pure character format. Sometimes called a "plain text file."

undocumented code

(See **wild code**.)

value

[10] See: http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary

[11] An example codebook from ICPSR is here: http://www.icpsr.umich.edu/icpsrweb/ICPSR/help/cb9721.jsp

[12] See: http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary

[13] See: http://www.ojp.usdoj.gov/BJA/evaluation/glossary/glossary_q.htm

[14] See: http://www.ojp.usdoj.gov/BJA/evaluation/glossary/glossary_q.htm

In most numeric data files, answers to questions are recorded with numbers rather than text, and often even numeric answers are recorded with numbers other than the actual response. The numbers used in the data file are called "codes." Thus, for instance, when a respondent identifies herself as a member of a particular religion, a code of "1" might be used for Catholic, a "2" for Jewish, etc. Likewise, a person's age of 18 might be coded as a 2 indicating "18 or over." The codes that are used and their correspondence to the actual responses are listed in a **codebook**. See also **code**.

value label

The textual description of a numeric value or code. For instance, the value label for a code of “1” in a question about religion might be “Catholic,” and “Jewish” for a code of “2”.

variable

In social science research, for each unit of analysis, each item of data (e.g., age of person, income of family, consumer price index) is called a variable.

wild code

In survey research, "wild" codes are codes that are not authorized for a particular question. For instance, if a question that records the sex of the respondent has documented codes of "1" for female and "2" for male and "9" for "missing data," a code of "3" would be a "wild" code, sometimes called an "undocumented code."

Appendix B. Resources

Data Curation Profiles
Data Curation Profiles are structured interviews with researchers about their data practices. The Toolkit provides background information on Data Curation Profiles, an interviewer's manual, an interview worksheet, and a template. A directory of example Data Curation Profiles is freely available.
Toolkit: http://www4.lib.purdue.edu/dcp/
Profiles: http://docs.lib.purdue.edu/dcp/

Data Documentation Initiative (DDI)
The DDI is a metadata specification for social and behavioral sciences. The Website offers information and resources for creating DDI metadata, including tools that generate XML.
http://www.ddialliance.org/

Digital Curation Centre's How-to Guides
The *How-to Guides* cover a range of practical topics about managing and curating data, including appraisal and selection of research data, data citation, data management plans, data management services, and licensing research data.
http://www.dcc.ac.uk/resources/how-guides

Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions
This online tutorial presents central issues in digital preservation and provides advice on how to address them, especially in library/archives/repository settings.
http://www.dpworkshop.org/dpm-eng/eng_index.html

ICPSR's Guide to Social Science Data Preparation and Archiving: Best Practices Throughout the Data Life Cycle (5th ed.)
Provides information relevant to data curation at every stage of the data lifecycle.
http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

ICPSR's Guidelines for Effective Data Management Plans
Goes through the recommended topics that should be addressed in data management plans, explains the importance of discussing each section, and gives examples of language that may be used.
http://www.icpsr.umich.edu/files/datamanagement/DataManagementPlans-All.pdf

IHSN technical note on metadata standards
Provides an overview of metadata standards in general and DDI specifically.
http://www.surveynetwork.org/HOME/sites/default/files/resources/DDI_SDMX_IHSN_DRAFT.pdf

JISC's Policy-making for Research Data in Repositories: A Guide
Outlines areas of concern in creating repository policies for data, including content and formats, metadata, ingest, access and reuse, preservation, and succession plans.
http://www.disc-uk.org/docs/guide.pdf

Starting the Conversation: University-wide Research Data Management Policy
Considers institutional policy issues for research data management in academic settings, such as data ownership, access to and preservation of data, ethical considerations, and costs.
http://www.oclc.org/content/dam/research/publications/library/2013/2013-08.pdf

UK Data Service Data management costing tool and checklist
Provides a breakdown of data management activities; does not provide sample costs.
http://www.data-archive.ac.uk/media/247429/costingtool.pdf