NDACAN National Data Archive on Child Abuse and Neglect

Contributor’s Guide to Preparing and Archiving Quantitative Data

Third Edition

Published by:
NDACAN
Bronfenbrenner Center for Translational Research (BCTR)
Cornell University
Ithaca, NY 14853

First edition, 2002. Second edition, 2014. Third edition, 2022.

ACKNOWLEDGEMENTS

Depositing data with the National Data Archive on Child Abuse and Neglect: A handbook for contributors (1st ed.) was published in 2002 (National Data Archive on Child Abuse and Neglect, 2002). The framework for the first edition was based on *Depositing data with the Data Resources Program of the National Institute of Justice: A handbook*. (National Institutes of Justice & Sociometrics Corporation, 1991).

The framework for the second edition, published in 2014, was influenced by the *Guide to social science data preparation and archiving: Best practices throughout the data life cycle* (5th ed.) (Inter-University Consortium for Political and Social Research, 2012). This third edition provides updates to NDACAN processing policies and procedures, as well as updated links and references.

Funding for the National Data Archive on Child Abuse and Neglect (NDACAN) is provided by the Children's Bureau, an office of the Administration for Children and Families (ACF), U.S. Department of Health and Human Services.

Suggested Citation:

National Data Archive on Child Abuse and Neglect (NDACAN). (2022). *Contributor's guide to preparing and archiving quantitative data* (3rd ed.). Ithaca, New York: Cornell University.

Table of Contents

About NDACAN

Founded in 1988, the National Data Archive on Child Abuse and Neglect (NDACAN) encourages secondary analysis of research data relevant to the study of child abuse and neglect. NDACAN acquires, preserves, and disseminates data to qualified researchers to foster exploration of important issues in child maltreatment. NDACAN actively supports analysis of its datasets by providing technical support, workshops, and researcher networking opportunities.

NDACAN is funded by the Children's Bureau, an office of the Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS), and is the sole repository for a number of large, federally funded datasets, including but not limited to:

- Adoption and Foster Care Analysis and Reporting System (AFCARS)
- National Child Abuse and Neglect Data System (NCANDS)
- National Youth in Transition Database (NYTD)
- National Incidence Study of Child Abuse and Neglect (NIS)
- Longitudinal Studies on Child Abuse and Neglect (LONGSCAN)
- National Survey of Child and Adolescent Well-Being I (NSCAW)
- National Child Welfare Information Study (NSWIS)
- National Juvenile Online Victimization Incidence Study (NJOV)

In addition to federally funded research with a data archiving requirement, NDACAN accepts voluntary contributions of child welfare data from well-designed quantitative research studies.

Data Sharing and Archiving Requirements

Data sharing has long been recognized as an important part of the research process (American Statistical Association, 2022). The federal government has adopted data sharing as a standard for federal agencies and grantees (Zients & Sunstein, 2010). The requirement that research data be archived for use by secondary users has now been incorporated into grant requirements for much federally supported research (Council on Government Relations, 2012).

Since 2003, the National Institutes of Health (NIH) has promoted data sharing and required that a data sharing plan be included in all grant submissions seeking $500,000 per year in direct costs (National Institutes of Health, 2003a, 2003b; National Institutes of Health Office of Extramural Research, 2003). Beginning in January 2023, the NIH will require an approved data management plan for all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data (National Institutes of Health, 2020).

The National Science Foundation (NSF) has required that all grant proposals include a data management plan detailing compliance with the data sharing policy since 2011 (National Science Foundation Directorate for Social Behavioral and Economic Sciences, 2011). In 2020, the NSF stated that "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants…Grantees are expected to encourage and facilitate such sharing" (National Science Foundation, 2020).

Importance of Data Sharing to Child Abuse and Neglect Research

Child abuse and neglect research is a relatively young field (National Research Council, 1993), but research results over the last three decades have shown "findings that delineate a serious public health problem" (National Research Council, 2014). Many key research questions remain to be explored. Well-designed, theoretically grounded studies can inform research and policy for decades to come (Institute of Medicine & National Research Council, 2013).

Secondary analysis of such well-designed research requires an environment in which data sharing is integral to the research process. Additionally, research has demonstrated an increased citation rate for investigators who share their data (Piwowar, Day, & Fridsma, 2007; Colavizza et.al., 2020).

In recognition of the importance of data sharing to the development of the field of child abuse and neglect, the Children's Bureau adopted a policy in 1994 that requires research grantees to archive datasets with NDACAN. The purpose of this policy is to ensure that data collected with Children's Bureau funds are available to other researchers. Grantees who have additional questions about the archiving requirement should contact their Federal Project Officer.

Children’s Bureau Archiving and Data Sharing Requirements

For studies funded by the Children’s Bureau and the Administration for Children & Families, researchers should consult their websites for the most recent requirements, as well their funding agreements for specific terms. Generally, archiving requirements allow a two-year grace period during which the original investigator has exclusive use of the research data, following the conclusion of funding. The two-year grace period was designed to allow the original investigator to publish research findings prior to release of data to secondary users. NDACAN requests that original investigators contribute their data as soon as possible after the completion of the project, but prior to the expiration of funding.

Submitting the data while the details of the project are fresh in the contributor’s mind is extremely helpful when questions arise during the archiving process. If early submission is not possible, contributors should submit their data no later than six months before the end of the grace period. This policy allows NDACAN staff sufficient time to resolve issues specific to the dataset (e.g., confidentiality concerns, coding problems) and to prepare a user's guide.

Regardless of when materials are submitted or when archiving is complete, data will not be released to secondary users until the original investigator has approved the study documentation or the grace period has expired, whichever comes first.

Why Archive Data with NDACAN?

NDACAN is a trustworthy repository of child maltreatment data with 35 years of experience. Our archival staff members are familiar with the preservation and dissemination practices particular to child maltreatment data. These include confidentiality, data security, preservation formats, copyright, and current trends in digital archiving.

This list below highlights some of the services NDACAN provides:

Data Management Plan Assistance

- If the project is appropriate for NDACAN to archive, we offer data management plan consultation and assistance. We may also provide a letter of support describing NDACAN's commitment to archiving the data as described in a grant proposal.

Easy Submission Process

- Electronic Study Submission Forms allow the data contributor to describe the data collection being deposited.
- Arrangements can be made to deposit data with delayed dissemination. This occurs when release would pose a confidentiality risk, when other dissemination plans have been made, or when a longer exclusive use period has been negotiated prior to the deposit.

Secondary User Benefits

- Datasets are freely available to qualified secondary analysts.
- Secure electronic data transmission to secondary users.
- NDACAN produces a User's Guide, based on information supplied in the Study Submission Form, which describes the research project under which the data were collected. The User's Guide is disseminated with the data to assist users with their understanding of the data collection.
- Staff are available to respond to questions submitted by secondary analysts using archived datasets.
- The Summer Research Institute (SRI), held annually, is where scholars come to the Cornell campus to work on research projects using datasets from our repository. During the week of the Institute, scholars receive individualized help formulating research questions and analyzing archived data.
- NDACAN hosts live webinars or posts pre-recorded video lectures based on our archived datasets in order to further aid secondary users' understanding of archived datasets.

High Standards

- Staff members are professionals from various fields of study. They are well versed in data confidentiality issues and de-identification techniques.
- NDACAN staff have decades of experience with the confidentiality needs particular to

child maltreatment research.

- NDACAN 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, and licensing agreement).
- Data are preserved on a secure network drive, backed up nightly, and stored in two locations using two different methods to insure preservation and longevity.
- Each dataset is assigned a unique and persistent identifier.

Contact NDACAN (ndacansupport@cornell.edu) to discuss the unique attributes of your dataset and to prepare an archiving and dissemination plan.

Planning Ahead: Designating NDACAN to Archive Your Data

Data sharing and data management requirements are becoming more prevalent among research funders. Both the NIH and NSF have added data management planning to their proposal requirements.

If you are interested in archiving your data at NDACAN, please contact us at ndacansupport@cornell.edu to schedule a conference call to discuss your proposed research, data, and plan. If it is determined the data fit within the scope of the NDACAN data collection, we may provide a letter of support describing NDACAN's commitment to archiving the data as described in the proposal, in exchange for your written consent accepting our archiving requirements. We will also continue to offer data management consultation over the course of your project.

It is expected that a portion of your research budget be allocated to cover the archiving process. Funding agencies, such as the NIH, allow you to include data sharing expenses in your research budget. You will need to estimate the amount of work necessary to complete the NDACAN Study Submission Forms, compile the required documentation, and prepare the data files for deposit. After depositing the data, you will need to allocate staff time for responding to NDACAN questions about the dataset. It is recommended that the data and documentation are deposited with NDACAN while funding is still available to pay study staff to complete the archiving process.

Recommended Elements of a Data Management Plan

The elements for inclusion in a data management plan have not been standardized in the archiving field. For grant writers, NDACAN provides a template below to create a data management plan with language specific to archiving with us. It includes key elements as recommended by the Interagency Working Group on Digital Data (Interagency Working Group on Digital Data, 2009). However, a particular funding source may have its own guidelines.

Resources for Creating a Data Management Plan

- California Digital Library, n.d.: DMP Tool
- Cornell University. Research Data Management Services Group: Data Management Planning
- National Institutes of Health (National Institutes of Health Office of Extramural Research, 2020): Final NIH Policy for Data Management and Sharing
- National Science Foundation (National Science Foundation Directorate for Social Behavioral and Economic Sciences, 2011): NSF Dissemination and Sharing of Research Results
- Inter-University Consortium for Political and Social Research (Inter-University Consortium for Political and Social Research, 2012): Data Management Plan Guidelines

Template for NDACAN Contributor Data Management Plan

Data Description.
Provide a brief description of the digital data you plan to collect including the data type and data collection procedures.

Impact.
Discuss the impact of these data within the field and any broad societal impact. Archiving at NDACAN will facilitate secondary analysis and maximize the value of the data.

Designated Archive.
Insert: "We have consulted with the National Data Archive on Child Abuse and Neglect at Cornell University, and these data have been accepted to be archived. NDACAN has demonstrated itself to be a trustworthy repository of child maltreatment data for over 30 years and will make the data available to the research community for secondary analysis. Included with this proposal is a letter from NDACAN in which they commit to archive the data."

Content and Format.
Insert: "The data will be archived at NDACAN in formats congruent with current practices in digital archiving. The documentation and codebook will be preserved in .pdf format, and the data and metadata will be preserved in SPSS, Excel, ASCII, CSV, or .pdf format as appropriate. The data may be migrated to new formats as best practices in digital archiving are updated. The data will be distributed in multiple formats (e.g., SPSS, SAS, and Stata) to reduce barriers to analysis."

Access.
Insert: "NDACAN will make the data available to the research community. Due to the sensitive nature of child maltreatment data, the use will be restricted to licensed researchers. Information in the dataset that could directly allow study subjects to be identified will be removed. NDACAN will provide researchers with ongoing user-support."

Preservation.
Insert: "NDACAN will assure long-term preservation. Data, metadata, and documentation are stored in sustainable formats and are backed up on a secure server."

Transfer of Responsibility.
Insert: "NDACAN will be available to offer data management consultation throughout the project. The data will be provided to NDACAN for processing upon completion of the research. NDACAN will collaborate with the contributor about archiving, subject the dataset to rigorous review, process the dataset for preservation and distribution, and create additional documentation as needed. The data may be embargoed for [insert time period of 0-2 years], to allow the contributors/principal investigators time to publish before being made available to the research community."

Preparing Study Materials for Archiving

NDACAN documentation requirements are based on archiving industry standards of practice. In this section, we will describe the current standards and the types of information required in your dataset package.

Required Metadata Elements

Metadata is "data about data." The following metadata elements are required by NDACAN for archiving. Each element is addressed and can be fulfilled by completing the NDACAN Study Submission Forms and requested supplemental study documents and data files.

- Principal Investigators – Principal investigator name(s), and affiliation(s) at time of data collection.
- 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.
- Dates of Data Collection – The date data collection started and ended.
- Unit(s) of Observation – Who or what is being studied.
- Data Source(s) – Origin of the data (i.e., surveys, administrative records, etc.)
- Sample Design/Sampling Procedures – How the sample was constructed.
- Data Collection Procedures – The ways in which the data were collected.
- Response Rates – The percentage of people who participated in the study out of the total number of people solicited to participate. Attrition rates can be included here also.
- Population of Inference – To whom the results are generalizable (i.e., children in the CPS system in the US, children in foster care, emancipated youth in the Northwest, etc.)
- Weighting – The construction and appropriate application of any weight variable(s).
- Geographic Coverage of the Data Collection – Locations where the data were collected and/or where they are generalizable (i.e., rural upstate NY, city of Chicago, US, etc.).
- Variable List and Codebook*
  - Exact question wording
  - Exact meaning of codes
  - Missing data codes
  - Imputation and editing information
  - Details on constructed and weight variables

- Location in the data file when relevant
- Variable groupings

(*Much of this information is usually found in the study generated codebook, protocol, interim report, or final report.)

Other Required Documents

- **Codebook(s)**
  A codebook includes a description of each variable contained in the study's associated data file(s). Many statistical software programs can produce a basic codebook containing much of the information listed below.

  For each variable, the following information should be provided in the codebook:
  - An unambiguous variable name
  - A descriptive variable label, consisting of a textual description of the item, or a clear reference to its associated question in the data collection instrument
  - Variable data type (i.e., numeric, character, date)
  - Missing/inapplicable data codes and their meanings
  - For categorical variables, a list of valid values and corresponding labels
  - For derived variables, the derivation logic and program files used to create the variables

- **Data Collection Instruments**
  Often a study will employ the use of a multi-item measure/instrument. The measure is sometimes an existing data collection instrument developed by someone else, a set of questions developed for the current study, or a modification of an existing instrument for use in the current study. It is critical for study authors to record the following elements and whenever possible, submit copies of all instruments (along with copyright status) used in the data collection:

  - Name of the instrument
    - Provide the full name of the instrument along with any acronym.
    - If it is a project created measure, we recommend including the study name as a part of the measure name.
      - Example: LONGSCAN Witnessed Violence
    - If the measure was adapted from an existing measure, we recommend acknowledging the measure it was adapted from in addition to the naming convention mentioned in the previous bullet point.
      - Example: LONGSCAN witnessed violence [adapted from History of Witnessed Violence by Snarky 2010]
  - Instrument authors

    - Instrument version (version number or a year of publication)
    - The name of the publisher who holds the copyright (if applicable)
    - A complete APA (American Psychological Association, 2020)-formatted citation for the measure

- **Related Publications**
  Submit all citations for publications related to the archived dataset and those that are key to furthering a secondary analyst's understanding of the data. The list of relevant publication citations will be captured and preserved in the NDACAN child abuse and neglect Digital Library (canDL). The canDL is a searchable online listing of publications related to archived datasets.

- **Interviewer Guide**
  Provide copies of the instruction document distributed to data collection representatives who were involved in collecting the data (if applicable).

- **Interim and Final Project Reports**
  Include copies of any interim or final project reports produced for the study's funder or provide a citation to where those documents can be found online.

- **Original IRB Approval and Informed Consent Forms**
  Data contributors must demonstrate that they conducted the research in accordance with the U.S. Protection of Human Subjects regulations (Protection of Human Subjects, 2018) and The Belmont Report (U.S. Department of Health Education and Welfare, 1978). To comply with this requirement, a study should have been reviewed and approved by the data collector's Institutional Review Board (IRB) prior to the start of data collection. The original approval from the IRB should be saved and submitted to NDACAN as proof of review and approval. Accompanying this letter, should be copies of all versions of IRB-approved informed consent forms. The informed consent form and IRB approval must not expressly prohibit archiving and data sharing.

Preparing Data Files for Archiving

Data files must be submitted in a readable format. The best formats are those readable by standard statistical software packages such as SAS, Stata, and SPSS. NDACAN also has expertise in relational databases which enables us to accept data in common database formats (MS Access, MySQL). Below, we discuss the different data files structures, formats, data file requirements, and confidentiality protections.

File Structures

The documentation should describe and enumerate the data file structure in accordance with the following definitions:

- **Rectangular** – a data file organized with one record (row) per participant in the entire file (no duplicate ID's)
- **Hierarchical/stacked** – a data file organized with multiple records (rows) per participant
- **Relational database** – multiple data files connected to each other via a "key ID" variable
- **Longitudinal/multi-wave study files** – multiple, separate data files which can be merged together via a variable common to all files in the collection, usually the participant ID variable

File Formats

NDACAN accepts data in a variety of file formats. We currently distribute SPSS and Stata native files, and SAS programs files with text data. Some of the statistical software programs are releasing 32-bit and 64-bit versions of their programs, which can render files unreadable when created by the same software but on a different platform (e.g., files created by 32-bit SAS cannot be opened using the 64-bit version of SAS). Prior to submitting your files, please discuss with NDACAN staff the file format and also the version of the program used to save the file.

- **Native-Software Specific** – These files are constructed specifically to run in a particular software package and are rooted in the version of the software in which they were saved. This means there may be limited compatibility with earlier and later versions of the software, with the file subject to becoming obsolete.

- **Portable-Software Specific** – These files are designed to run in a particular software package but are constructed in such a way to ensure compatibility across versions of the software.

- **Text (ASCII) Data with Import Program File** – This format consists of two files. The first file is a text version of the data known as "ASCII." The second file contains the file specifications (column width, etc.) of the text data along with program commands to read it into a specific software package. This is the most robust way of preserving data for future compatibility across versions of a software program.

Required Data File Elements

For each variable, the following information should be provided in the data file or corresponding formats file:

- An unambiguous variable name
- A descriptive variable label – A textual description of the item, or a clear reference to its associated question in the data collection instrument.
- A list of valid values and corresponding labels for categorical variables
- Missing/inapplicable data codes and their meanings
- Variable data type (i.e., numeric, character, date)
- Column specifications for each variable
- Decimal settings should reflect the data contained in each variable

Protecting Confidentiality

There are established standards for de-identifying sensitive data that are used by data custodians and researchers. Below, we describe the different types of disclosure risks and some of the common methods for de-identifying data. When taking measures to reduce disclosure risks, care must be taken to preserve the research utility of the dataset and to avoid high levels of data distortion. A confidentiality disclosure review should be conducted prior to depositing data.

Direct Identifiers: These are variables that directly identify individuals (see National Archive of Criminal Justice Data, n.d. for additional examples) and must be removed from the dataset.

Examples of direct identifiers:

- Names
- Social Security Numbers
- Phone Numbers
- Medical Record Numbers
- Insurance Card Numbers
- Highly Specific Geographic Variables (i.e., Street Addresses, Geo-coordinates, Census Block)

Indirect/Quasi-Identifiers: These are variables that do not directly identify individuals but may be matched with other information to infer the identity of an individual (see National Archive of Criminal Justice Data, n.d. for additional examples).

Examples of possible quasi-identifiers:

- Gender
- Occupation
- Organizational memberships; offices held
- Date of Birth
- Ethnicity
- Rare Characteristics (disease, medications, etc.)
- Death, Date of Death, or Reason for Death

Common De-identification Strategies

There is a delicate balance between protecting participant confidentiality without compromising the richness of the dataset or inducing excessive data distortions. The researcher should carefully consider the uses of the data when contemplating which de-identification strategy to implement. Archive staff members are available to discuss and help resolve issues with variables posing a threat to disclosure. Below, we outline the commonly used de-identification strategies.

- **Variables**: Recode or remove any variable with embedded primary identifiers, such as fragments of a social security number.

- **Dates**: Recode ALL dates so that the day of the month is the 15th, while preserving month and year. If the breadth of the dataset is such that it provides extensive details about a participant, meaning the likelihood of deductive disclosure is high, removing the day while preserving year and month is advised. In some cases, generating a derived variable (e.g., replacing the birth date with the age of the participant expressed in months) in place of a partial date is more useful to researchers while maintaining participant confidentiality.

- **Geographic data:** When faced with geographically specific data in large administrative datasets, NDACAN employs the following re-coding strategy: If statistical summary counts fall below a specified threshold for a specific geographical region (i.e., county), NDACAN recodes the region's identity as "other" or "unknown." Other strategies are to remove the geographical variable completely or aggregate the data to a higher level of aggregation.

- **Categorical data**: When the group counts for categorical variables are very small, especially demographic data (i.e., race), collapse very small groups into one larger group to construct a larger cell count, thereby reducing disclosure risks.

- **Numeric data**: When the variable contains sensitive information that could be manipulated to derive the identity of a participant, the numeric variable can be re-coded in the following ways:

    - **Top or bottom coding**: For variables that have distributions with long tails, collapse responses at the skewed end of the spectrum of responses into a category, starting at a designated cutoff point. For example, a continuous variable containing data in the form of age in years could implement a top coding strategy where participants age 65 and older are lumped into a single category, or participants below the age of 18 could be bottom coded into a single category.

    - **Make the data categorical**: Transform the data from continuous to several categories of responses. For the example of age in years, you would recode the data into several categories defined by an age range.

- **String data**: Verbatim responses may pose a significant disclosure risk, as the responses are not limited to pre-determined choices but are open-ended and therefore, could contain potential identifiers, such as names, birth dates, or death information. Great care should be taken to read through and remove potentially identifying information. Consideration should be given to re-coding the verbatim responses into categorical responses. (See Qualitative Data Repository, n.d.).

After you have made the initial contact with NDACAN, compiled the required documentation, and have completed the process of preparing the data files, you are now ready to archive the dataset with NDACAN!

The Archiving Process

NDACAN strives to make the process of depositing data as seamless as possible. While the bulk of the responsibility rests on the data contributor, NDACAN staff members are always available to address questions and provide assistance. In addition to the information detailed in the preceding sections, additional information and documentation is required to complete the archiving process. Figure 1 shows the archiving process of a typical dataset, from time of acquisition through dissemination.

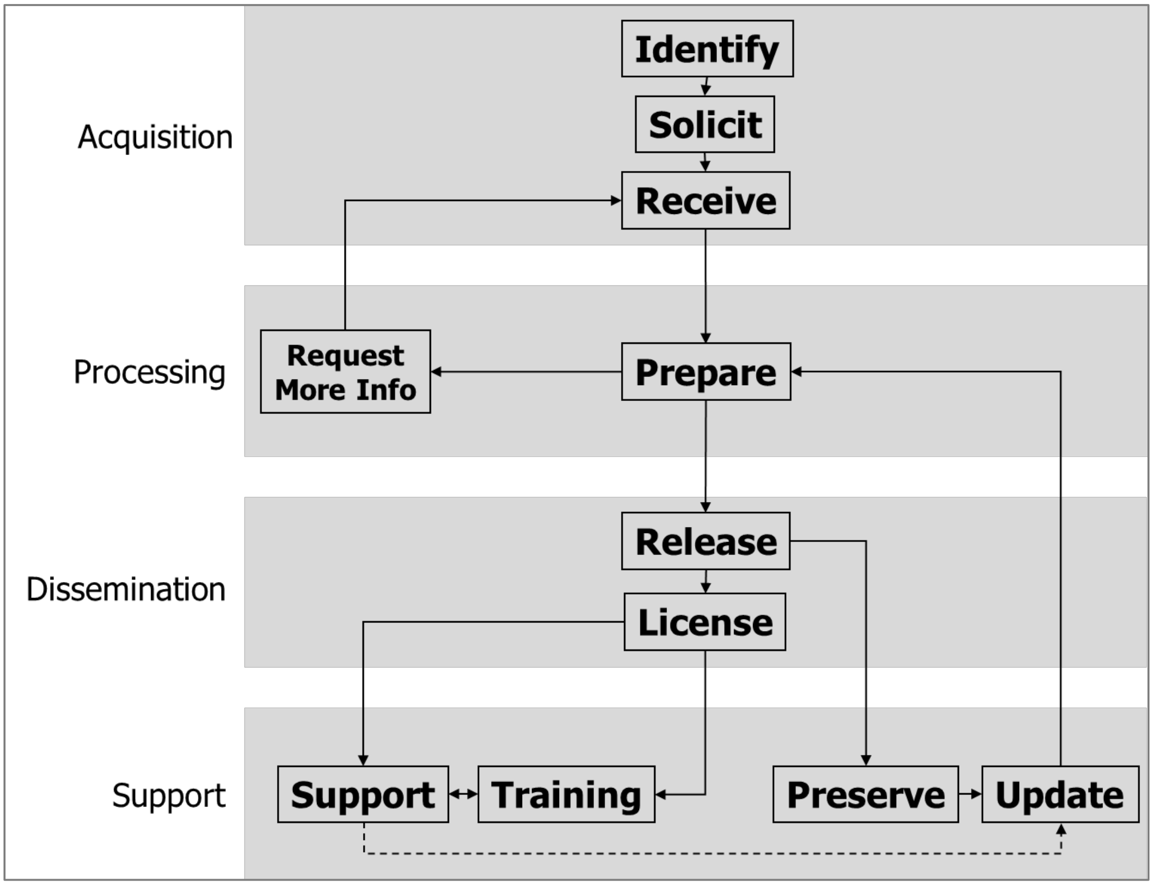

Figure 1. Archiving Cycle

Acquisition: Deposit Data with NDACAN

The Acquisition phase begins when NDACAN receives the dataset materials from the data contributor. This is the start of the "ingest" phase.

Study Submission Forms

Investigator Contact Sheet: Requests contact information for an investigator involved in the study. The form should be completed by each investigator and the study contact person (if not one of the investigators).

Study Submission Form Part I: Requests basic information about the study and is submitted in advance of archiving any data. The form collects the study title, abstract, funder name, award number, and award dates. After reviewing the information, NDACAN will schedule a conversation to discuss the study and determine whether the study fits within the scope of NDACAN's collection.

Study Submission Form Part II: Once the data collection is determined to be a good fit for NDACAN, the designated study contact person will complete Part II of the Study Submission Form. Part II collects information about study level details.

Study Submission Form Part III: This form consists of a spreadsheet in which the study contact person will document all data files associated with the data collection.

Instrument Information Form: This form requests information about the instruments/measures used to collect data, including how they may have been revised for the purpose of this data collection. Instrument details such as author, version, proper name of the instrument, nickname of the instrument, description, and bibliographic citation are collected.

Supporting Documentation

- Study Protocol (w/survey instruments)
- Interviewer Guide
- Data Dictionary/Codebook(s)
- Interim and Final Reports
- Citations for Published Works
- All Versions of Informed Consents
- Original IRB Approval Forms

Data Ownership and Intellectual Property Rights

Please include a clear statement of who owns the data. NDACAN requires that the data owner sign the NDACAN Contributor's Agreement upon deposit of the dataset. Additionally, for each document (instruments/measures, etc.) used in the study, the contributor will need to indicate who holds the intellectual property rights to it (provide proper bibliographic citation).

Processing

NDACAN works closely with data contributors to ensure that the documentation accurately reflects the study data and metadata. This is an iterative process, in which NDACAN staff members carefully examine the provided documentation and data in order to verify the manner in which the data were collected and subsequently recorded in the corresponding data files. When NDACAN encounters inaccuracies, inconsistencies, or has other questions, we contact the data contributor for help. It is important for data contributors to work closely with staff and respond to these requests. The more questions about a dataset are resolved prior to release, the fewer questions there will be from secondary analysts after it is released.

Replicate Statistics from Published Works

NDACAN staff will attempt to replicate basic statistical information found in published works. This is also what we instruct secondary analysts to do after they have reviewed the documentation, in order to familiarize themselves with the data. The expectation is that descriptive statistics found in the literature will match those generated using archived data. When discrepancies are discovered, Archive staff work with the data contributor to resolve inconsistencies and when an inconsistency between the archived data and published findings is not resolvable, a note about the difference is placed in the study documentation.

Conduct a Confidentiality Disclosure Review

NDACAN conducts a confidentiality disclosure review of all data files contained in the dataset in order to reduce potential threats to confidentiality, using the strategies and techniques described earlier in this handbook (Common De-Identification Strategies).

Develop a User's Guide

Using information provided in the Study Submission Form and the published literature, the Archive staff creates a User's Guide to accompany the dataset upon its release. The User's Guide contains study level metadata to aid secondary analysts' understanding of the study. Notes regarding unique attributes of the data are included in the User's Guide.

Source and Document Measures/Instruments Used in the Study

For each data collection instrument used in the study, Archive staff will document the name of the instrument, author(s), publisher (if applicable), year of publication, website where to order/download the instrument, and an APA-formatted citation. It is important that secondary analysts be able to access a copy of the instrument. We encourage archiving a copy of all instruments used during data collection. Even when copyright prevents NDACAN from including a copy of the measure under "scholarly use copyright provisions," staff will match data files to the instrument using descriptive variable and value labels. To assist prospective data users in identifying datasets of interest, we publish the name and bibliographic citations of measures used in the datasets on the "Measures Index" page of our website.

Construct a Codebook

Datasets should be contributed with a codebook relating to each data file in the dataset. When the codebook information is inadequate, staff may opt to create a new version of the codebook containing all the required elements (see “codebook” section above).

Create Multiple File Formats for Dissemination

NDACAN currently distributes data readable in the following software programs: SAS, SPSS, and Stata. Regardless of the format in which the files are contributed, staff will run the necessary conversions to create the files necessary for all supported statistical software programs.

Preservation of the Data and Metadata

NDACAN preserves all originals contributed by the investigator as well as all versions resulting from the processing of the data and documentation. These files are stored on a secure server located at Cornell University, which is backed up daily.

Dissemination

Once the dataset's data files and documentation have been finalized, it is made available to secondary analysts. The process for releasing the dataset includes announcing its availability to analysts via the Child-Maltreatment-Research-Listserv (CMRL) and adding it to the Datasets listing page of our website and other promotional materials. NDACAN actively promotes the use of its dataset holdings through conference attendance and presentations.

Requests for Access to a Dataset

Secondary analysts interested in using a dataset will first review the dataset title, abstract, and documentation available from the Datasets page of our website (https://www.ndacan.acf.hhs.gov/). In order to gain access to the dataset, the analyst will complete the steps involved in requesting a dataset, as detailed on the dataset's Order Dataset page. The steps for most datasets include the following:

- The analyst must submit their contact information.
- The analyst must print, complete, sign, and submit a Terms of Use Agreement for each dataset requested.

NDACAN staff reviews the request for dataset access to ensure eligibility for data licensing.

Restricted Access Data Licensing

For some of NDACAN's datasets, additional documentation is required before an analyst can gain access. Datasets requiring these additional measures are those with highly sensitive data (vulnerable populations, availability of small-scale geographical data, contractually required data access restrictions, etc.). NDACAN decides the nature of the access protocol used for each deposited dataset. If data contributors believe that their data requires these additional measures of protection, Archive staff are open to discussing the matter. The general stance of NDACAN is to reduce barriers to data access for qualified secondary analysts. We strive to make every dataset fall under our General Terms of Use Agreement data access process.

Preservation

Contributed data are preserved at several points in time. First, when they arrive at NDACAN as Contributor Originals, then again after the dataset and documentation has been processed and the data are ready for initial dissemination. And again, each time a dataset undergoes revisions. NDACAN uses versioning to document dataset changes across time.

Data Security

The NDACAN servers, where contributed data is stored, are hosted within the College of Human Ecology at Cornell University in a secure server room. The room is environmentally controlled and physical access is limited and monitored by the College's Computing Services Group (CSG). NDACAN uses an enterprise class backup software, VEEAM community edition, to backup all of our servers and workstations. These backups can be encrypted, de-duplicated, and compressed as the data being backed up requires.

For redundancy and security, copies of those backups are picked up by the EZ-Backup software, provided by Cornell, each night and copied to the Tivoli storage manager for off-site long-term backup and archive storage. Only authorized administrators with the proper NDACAN active directory domain authentication and group membership have access to these backups and can restore data sets to the servers. EZ-Backup uses IBM Tivoli Storage Manager (TSM), which can store backups from all platforms. The files in their compressed backup form in the Tivoli system do not allow any type of file level access by individuals. The backups have to be restored to the servers they were taken from by the VEEAM software in order to have access to the files.

This approach provides a multitier or layered security format. Anyone attempting to access the files must have administrative access at four different system levels to reach readable data.

Technical Support

Technical support for secondary analysts begins when NDACAN releases the dataset.

Secondary Analyst Inquiries

Secondary analysts are instructed in the study documentation and on the website to submit questions about a dataset to Archive staff via the ndacansupport@cornell.edu. We consider ourselves to be a buffer between you and the secondary analyst. Archive staff will attempt to respond to inquiries first before we reach out to the data contributor for help.

Special Data Requests

Interested agencies and organizations who lack the statistical expertise, research experience, or who are not qualified to receive the data according to the Terms of Use may submit special data requests. NDACAN staff work with the requestor to clarify what is needed, run the statistical analysis, and provide the requestor with summary information, usually in the form of tables. In the past, these types of requests have resulted in research briefs and have populated websites with child maltreatment statistics.

Summer Research Institute

In addition to responding to technical support requests from secondary analysts, NDACAN also hosts the Summer Research Institute (SRI) annually that provides selected participants with intensive user support while they work on research projects using datasets from the Archive. The Institute is designed to help secondary analysts complete and publish their findings in the peer-reviewed literature.

Online Resources for Secondary Analysts

In order to assist secondary analysts further with their research plans and dataset identification, NDACAN maintains two searchable online databases accessible from our website, the canDL and the Measures Index.

- *child abuse and neglect Digital Library (canDL):*
  The canDL is a collection of published works, in the form of bibliographic citations, from both data contributors and secondary analysts and is organized by dataset. The digital library assists users in becoming familiar with the types of research questions already addressed in the literature, as well as furthering their understanding of how the data were collected and used by the original author.
  https://www.ndacan.acf.hhs.gov/candl/candl.cfm

- *Measures Index:*
  The Measures Index is a database of bibliographic citations for each measure used by the studies in the Archive. The Measures Index is searchable by keyword or dataset number. Keyword searches enable prospective secondary analysts to identify datasets of interest based on the presence of measures of a particular construct.
  https://www.ndacan.acf.hhs.gov/measures-index/measures-index.cfm

Abbreviations

AFCARS – Adoption and Foster Care Analysis and Reporting System

APA – American Psychological Association

canDL – child abuse and neglect Digital Library

CSG – Computing Services Group, located within the College of Human Ecology at Cornell University

CMRL – Child-Maltreatment-Research-Listserv

IRB – Institutional Review Board

LONGSCAN – Longitudinal Studies on Child Abuse and Neglect

NCANDS – National Child Abuse and Neglect Data System

NDACAN – National Data Archive on Child Abuse and Neglect

NIH – National Institutes of Health

NIS-4 – Fourth National Incidence Study of Child Abuse and Neglect

NSF – National Science Foundation

NSCAW I – The National Survey of Child and Adolescent Well-Being I

NSCAW II – The Second National Survey of Child and Adolescent Well-Being

NYTD – National Youth in Transition Database

SRI – NDACAN's Summer Research Institute

Bibliography

American Psychological Association. (2020). *Publication Manual of the American Psychological Association, (7th ed.).* Washington, DC: American Psychological Association.

American Statistical Association. (2022). Ethical guidelines for statistical practice. Retrieved from https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice. Accessed 1 APR 2022.

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020). The citation advantage of linking publications to research data. PLoS ONE 15(4): e0230416. https://doi.org/10.1371/journal.pone.0230416

Council on Government Relations. (2012). Access to, sharing and retention of research data: Rights and responsibilities. Available from https://www.cogr.edu/sharing-and-retention-research-data-rights-and-responsibilities. Accessed 1 APR 2022

Institute of Medicine, & National Research Council. (2013). *New directions in child abuse and neglect research*. Washington, DC: The National Academies Press.

Inter-University Consortium for Political and Social Research (ICPSR). (2009). Guidelines for effective data management plans. Retrieved February 3, 2014, from http://www.icpsr.umich.edu/files/datamanagement/DataManagementPlans-All.pdf. Accessed 1 APR 2022.

Inter-university Consortium for Political and Social Research (ICPSR). (n.d.). *Guide to social science data preparation and archiving: Best practices throughout the data life cycle* (6th ed.). Ann Arbor, MI: Institute for Social Research, University of Michigan. Available at https://www.icpsr.umich.edu/web/pages/deposit/guide/. Accessed 1 APR 2022.

Interagency Working Group on Digital Data. (2009). Harnessing the power of digital data for science and society: Report of the interagency working group on digital data to the committee on science of the national science and technology council. Available from https://www.nitrd.gov/pubs/Report_on_Digital_Data_2009.pdf. Accessed 1 APR 2022.

National Archive of Criminal Justice Data (NACJD). (n.d.) *Data with confidential content*. Available at https://www.icpsr.umich.edu/web/pages/NACJD/archiving/confidential-content.html. Accessed 4 APR 2022.

National Data Archive on Child Abuse and Neglect (NDACAN). (2002). *Depositing data with the National Data Archive on Child Abuse and Neglect: A handbook for contributors* (1st ed.). Ithaca, New York: Cornell University.

National Institutes of Health. (2003a). *NIH data sharing policy*. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/. Accessed 1 APR 2022.

National Institutes of Health. (2003b, February 26). *Final NIH statement on sharing research data*. Retrieved from http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html. Accessed 1 APR 2022.

National Institutes of Health. (2020). *Final NIH Policy for Data Management and Sharing.* Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html#:~:text=October%2029%2C%202020-,January%2025%2C%202023,-Related%20Announcements. Accessed 31 MAR 2022.

National Institutes of Health Office of Extramural Research. (2003, March 5). *NIH data sharing policy and implementation guidance*. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm. Accessed 1 APR 2022.

National Institutes of Justice, & Sociometrics Corporation. (1991). *Depositing data with the Data Resources Program of the National Institute of Justice: A handbook*. Available at https://nij.ojp.gov/library/publications/depositing-data-data-resources-program-national-institute-justice-handbook. Accessed 1 APR 2022

National Research Council. (1993). *Understanding Child Abuse and Neglect*. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id=2117#.VFzZ4n87KOk.mailto. Accessed 1 APR 2022.

National Research Council. (2014). *New Directions in Child Abuse and Neglect Research.* Washington, DC: The National Academies Press. Available from https://nap.nationalacademies.org/catalog/18331/new-directions-in-child-abuse-and-neglect-research. Accessed 4 APR 2022.

National Science Foundation Directorate for Social Behavioral and Economic Sciences. (2011). *Data archiving policy*. Retrieved from http://www.nsf.gov/sbe/ses/common/archive.jsp. Accessed 1 APR 2022.

National Science Foundation. (2020). *Proposal & Award Policies & Procedures Guide. Chapter XI - Other Post Award Requirements and Considerations*. Retrieved from https://www.nsf.gov/pubs/policydocs/pappg20_1/pappg_11.jsp#XID4. Accessed 31 MAR 2022.

Piwowar H.A., Day R.S., & Fridsma D.B. (2007). Sharing Detailed Research Data is Associated with Increased Citation Rate. *PLoS ONE 2*(3): e308. doi: 10.1371/journal.pone.0000308.

Protection of Human Subjects, 45 C.F.R. § 46(2018). Available at https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html. Accessed 1 APR 2022.

Qualitative Data Repository. (n.d.) *De-Identification.* Available at https://qdr.syr.edu/guidance/human-participants/deidentification. Accessed 1 APR 2022.

U.S. Department of Health Education and Welfare. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Zients, J. D., & Sunstein, C. R. (2010). *Sharing data while protecting privacy*. Retrieved from https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2011/m11-02.pdf. Accessed 1 APR 2022.