NDACAN National Data Archive on Child Abuse and Neglect

NDACAN Archiving Process and Steps

Background

The steps outlined in this document are explained in greater detail in the document titled, "*A Contributor's Guide to Preparing and Archiving Quantitative Data,*" referred to from this point forward as the "Contributor's Handbook." The Contributor's Handbook is found at the following link: https://www.ndacan.cornell.edu/contribute-data/A_Contributor's_Guide_to_Preparing_and_Archiving_Quantitative_Data.pdf.

This document is not intended to be an all-inclusive or exhaustive list of everything required of a Data Contributor, but rather its purpose is to provide a high-level summary of archiving processes, steps, and time frames for submission. The unique attributes of the data being archived, confidentiality protection considerations, and the quality of the data and documentation will determine the final actions and tasks associated with the general process of archiving the dataset. Information in this document should be used in conjunction with information found on the Contribute Data page of the NDACAN website (https://www.ndacan.cornell.edu/contribute-data/contribute-data-general.cfm) and the Contributor's Handbook, providing a comprehensive understanding of the archiving process.

NDACAN, in consultation with the Children's Bureau, Administration of Children and Families, reserves the right to reject any dataset that meets any one of the established archiving exclusion criteria. Also, a dataset's primary purpose may be modified if special circumstances necessitate (e.g. dissemination or preservation). The criteria are available by request.

Note: Funders, such as the National Institutes of Health (NIH), require a Data Management/ Sharing Plan to be submitted with requests for research funding. Only prospective Data Contributors who are in the process of seeking funding for a study proposal will need to create a Data Management Plan to submit to their prospective funding agency. When these instances arise, NDACAN works with the Data Contributor to draft the plan and issues a letter of acknowledgement stating that they are aware of the project and agree to accept the data upon completion of data collection, if the proposal is successfully funded.

Preparing Study Materials for Archiving

Please see the Contributor's Handbook for a more extensive description of items discussed in this section. Page 11 of the Contributor's Handbook provides a description of the information collected in the Study Submission Forms and required elements of other documents, such as the Codebooks. Data Contributors should leverage existing documentation, such as study protocols, research articles, and interim reports when completing the Study Submission Forms. The

information solicited in the forms should already exist in the other documents and should not require much in the way of creating new text.

NDACAN uses the information collected in the Study Submission Forms to create a User's Guide for the dataset. Data Contributors are welcome to create their own User's Guide document and submit it to NDACAN. NDACAN can supply Data Contributors with the required pages (cover page) and text (e.g., publication submission requirement language, acknowledgement of source, etc.) for the User's Guide. Even if a Data Contributor creates their own User's Guide documents, the Study Submission Forms are still required.

Preparation Steps

- **Complete Study Submission Form Part I** (https://www.ndacan.cornell.edu/contribute-data/Study_Submission_Form_Part_1.pdf). This document asks Data Contributor's to provide a title, abstract, and study funder information.
- **Complete an Investigator Contact Sheet** (https://www.ndacan.cornell.edu/contribute-data/SSF_InvestigatorContactSheet.pdf) for each person involved in the study. Investigator information should be submitted with Part I of the Study Submission Form. Be sure to designate one person as the "Contact person for questions about this study."
- The Study Submission Form Part I and Investigator Contact Sheet(s), should be submitted three months after study funding begins. Early contact benefits the Data Contributor by informing them of archiving requirements, next steps in the process, and timeframe for submissions. After receiving the Study Submission Form Part I, NDACAN will set up a **conference call** with the designated study contact person to discuss the unique attributes of the study/data and timeline for deposit of the study data and documentation. The Data Contributor will also have the opportunity to ask questions during the call.
- **Complete Study Submission Form Part II** (https://www.ndacan.cornell.edu/contribute-data/Study_Submission_Form_Part_II.pdf) and **Study Submission Form Part III** (The link to download the MS Excel file can be found on this webpage: https://www.ndacan.cornell.edu/contribute-data/contribute-application-process.cfm). Complete the **Instrument Information Form** (https://www.ndacan.cornell.edu/contribute-data/Study_Submission_Form_Instrument_Info.pdf) for each measure/instrument administered during the course of the study and for which there are data in the data file(s).
- Submit a **Codebook/Data Dictionary**. The Data Contributor must provide a document that contains more information about the data files that are submitted.

  For each variable, the following information must be provided in the codebook:
  - An unambiguous variable name that matches the variable name found in the data file.
  - 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.
- When not prohibited by copyright law, provide copies of the actual **data collection instruments/measures** used in the study. If a survey was administered, include a copy of the survey instrument. The instruments are important because might show skip logic for questions or qualifying logic for having received portions of the survey.
- Copies of **interim and final reports** related to the project are required.
- Copies of the **Institutional Review Board approved study protocol** specific to the data collection (if separate sites, each site will need to submit their own institution's IRB approval), and a copy of the most recent IRB approved Informed Consent Form template are required.
- At the time NDACAN receives the dataset package, they will assign the dataset a unique number and will produce the Contributor's Agreement, which will be sent to the study contact. The person who signs that form should have the authority to do so, such as a Principal Investigator. Once signed, submit the form to NDACANsupport@cornell.edu. At the point in which the dataset is done being processed and is ready for dissemination, the form will be counter-signed by the NDACAN Director and a copy sent to the Data Contributor for their records.

Preparing Data Files for Submission

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 that enables them to accept data in common database formats (MS Access, MySQL). Below is more information about confidentiality protections, different data files structures, formats, and data file requirements.

Protecting Confidentiality

A confidentiality disclosure review should be conducted by the Data Contributor prior to depositing data.

The following types of information must be removed from the dataset PRIOR to submitting the files to NDACAN:

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

Data Contributors are responsible for identifying other variables that might pose a threat to participant confidentiality, consulting with NDACAN staff (NDACANsupport@cornell.edu) and the Contributor's Handbook on how best to mitigate the risk. NDACAN will also conduct a disclosure risk analysis upon receipt of the data.

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. NDACAN currently distributes SPSS, Stata, and SAS native files, program files for SAS, SPSS, and Stata, and text data files. Prior to submitting files, discuss with NDACAN staff the file format and also the version of the program used to save the file, such as:

- 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 Data with Import Program File – This format consists of two files. Both files are text. Text can be in ASCII, UTF-8, or another specified character set. The first file is raw data, either column-specified or delimited. The second file contains the file specifications of the data file – column width for column-specified, delimiter character (usually comma or tab) for delimited files – 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/syntax file readable by the statistical software program:

- An unambiguous variable name that matches the name appearing in the codebook.
- 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.
- Decimals settings should reflect the data contained in each variable.

Note: When there are multiple data files and Codebooks, include a document that maps the data file to its respective Codebook document.

Dataset Submission

Everything described above, along with any additional documents or information deemed necessary to understanding the data, make-up what is known as the "dataset package." Once Data Contributors have completed the preparations outlined in the sections above, then the package is ready to be archived with NDACAN.

Timeframe for Archiving

Data Contributors should submit the Study Submission Form Part I and Investigator Cover Sheet(s) within the first three months after funding for the study begins. The call between NDACAN and the Data Contributor will be scheduled after receiving the documents. The actual deposit of ALL other materials and files should happen a minimum 8 months prior to study funding expiration, but can happen as soon as the Data Contributor has finished data collection and data file cleaning has concluded. The reason for this time frame is that NDACAN staff need time to process the data and documentation files. This process requires that the Data Contributor respond to questions and review the final dataset package before it can be released. NDACAN could have many datasets in the queue to be processed, so waiting too long to archive data may mean that the Data Contributor's staff will need to respond to questions after project funding has expired. If no staff are available to respond to questions, then the archiving of the data will be considered incomplete, at which point NDACAN will notify its Contracting Officer's Representative (COR) at the Children's Bureau. Archive staff with do their best to provide estimates for when processing of the dataset will begin and potentially conclude, however, the unique attributes of the datasets in the queue, ahead of the one being submitted, in addition to the unique attributes of dataset being submitted, as well as, Data Contributors response times to questions and reviews, will impact start of processing and processing completion time of datasets in unforeseeable and unpredictable ways.

The final steps, as described above, include:
Step One: Complete and submit Part I of the Study Submission Form and the Investigator Contact Cover Sheet.

Step Two: NDACAN will set-up a call to discuss the dataset (as described above). The Data Contributor will have the opportunity to ask questions. NDACAN will decide whether the data are suitable for archiving at their data archive.

Step Three: If NDACAN determines the data are suitable for archiving, prepare the remaining elements of the dataset package in accordance with the Contributor's Handbook and as summarized above.

Step Four: Once the dataset package is assembled, create a compressed .zip folder which contains the entirety of the dataset package. Notify NDACANsupport@cornell.edu that the dataset is ready for submission.

Step Five: When NDACAN receives the request to submit the dataset package, they will evaluate and choose one of the following file sharing methods for the Data Contributor to send the dataset package, based on what is known about the data at that time (or use filing sharing methods with similar security features as those described below):

- **Cornell Enterprise Box**

Cornell Enterprise Box provides a web interface for uploading, downloading, sharing, and editing files. It uses high-grade TLS encryption in transfer, multi-layered encryption at rest with 256-bit AES, and works on Macs, Windows, and mobile devices. A Cornell netID is required to use Cornell Enterprise Box. Cornell assigns netID's with strong passwords to employees and students for using IT services. To view the rules that employees and students must follow visit this page (https://it.cornell.edu/device-security/strong-passwords-your-computer-netid-and-othercornell-services). To receive data, NDACAN creates secure folders in Cornell Enterprise Box and grants time-limited access to Data Contributors. For delivering data, NDACAN creates secure folders in Cornell Enterprise Box and grants time-limited access to data recipients. Data Contributors and data recipients only need to sign up for the free box.com individual account to send or receive data through this method.

- **Cornell Dropbox (not affiliated with dropbox.com)**
  NDACAN uses Cornell's proprietary secure file transfer portal DropBox to disseminate and receive data requiring a two-step authentication. The system will soon be renamed "Cornell Secure File Transfer." Cornell Dropbox encrypts data both in transit (TLS) and at rest using modern encryption standards. Files exist only for the amount of time specified, at which point they are securely deleted from the system. Dropbox is hosted on-premise at Cornell in the most secure (network, physical access) portion of its datacenter. For recipients of data, NDACAN creates a 3-day window to download the encrypted restricted data .zip package. NDACAN tells the recipient a password by phone and supplies a unique link generated by Cornell DropBox. The Cornell DropBox system requires the recipient to be present at their computer when downloading, and tests this by providing an access PIN which expires in 10 minutes.

Step Six: Once NDACAN retrieves the file from one of the systems discussed in Step Five, then they will conduct a quick review to be sure that the files received match what is required or was discussed in prior conversations. Processing the dataset package may not occur right away if other datasets were in the queue ahead of the dataset submitted.

Step Seven: NDACAN will process the dataset in the order in which it was received in the queue of datasets waiting to be processed. This requires a study contact person to be available to respond to questions and review the final dataset package once it has been prepared. More information about the tasks undertaken during processing can be found starting on page 21 of the Contributor's Handbook.

Dataset 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 NDACAN's 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 the website (http://www.ndacan.cornell.edu/). 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 (https://www.ndacan.cornell.edu/datasets/request-dataset.cfm). The steps for most datasets include the following:

- The analyst must submit their contact information.
- The analyst must print, complete, sign, and submit a Term of Use Agreement for each dataset requested (https://www.ndacan.cornell.edu/datasets/order_forms/TermsofUseAgreement.pdf).

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

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 (e.g., 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, NDACAN staff are open to discussing the matter. The general stance of NDACAN is to reduce barriers to data access for qualified secondary analysts. NDACAN strives to make every dataset fall under its General Terms of Use Agreement data access process.

Below are the links to the restricted access data license process and forms:

- Restricted Data Order Instructions: https://www.ndacan.cornell.edu/datasets/request-restricted-data.cfm
- License for Restricted Data: https://www.ndacan.cornell.edu/datasets/order_forms/LicenseRestrictedDataset.pdf
- Application for Restricted Data: https://www.ndacan.cornell.edu/datasets/order_forms/ApplicationRestrictedData.pdf

Once the required forms and steps have been completed by a secondary analyst interested in using the dataset, and their application materials have been reviewed and approved, then NDACAN staff use one of the file sharing methods described in Step Five of the Data Submission section of the document, to send them the dataset package.

For more information, visit our website at https://ndacan.acf.hhs.gov or email us at NDACANsupport@cornell.edu.