2020

Data Rescue & Curation Best Practices Guide

OCUL Data Community (ODC) Data Rescue Group

Data Rescue & Curation Best Practices Guide

November 2016

Updated February 2020

Authors:

OCUL Data Community (ODC) Data Rescue Group
(Alexandra Cooper, Jane Fry, Walter Giesbrecht, Vince Gray, Vivek Jadon, Amber Leahey, Kristi Thompson, Leanne Trimble)

Table of Contents

About this guide 2
Working with at-risk datasets 2
Fundamentals: rescuing data files & metadata 3
Case study 1: Retrieving data from a hard drive 4
Case study 2: Retrieving data from an obsolete storage medium 5
Long-term preservation 6
Additional Resources 6
Data curation to improve access 6
The OCUL context 7
Data Processing Workflow 7
Completed Studies 9
Additional Resources 9
Further value-added enhancements for data reuse 9
Appendix 1: Workflow for Files from Government of Canada Open Data website 10
Appendix 2: Creating Codebooks & Data Dictionaries 13
Appendix 3: Glossary 17

About this guide

The aim of this guide is to provide an accessible and hands-on approach to handling data rescue and digital curation of at-risk data for use in secondary research. We provide a set of examples and workflows for addressing common challenges with social science survey data that can be applied to other social and behavioural research data. The goal of this guide and set of workflows presented is to improve librarians' and data curators' skills in providing access to high-quality, well-documented, and reusable research data. The aspects of data curation that are addressed throughout this guide are adopted from long-standing data library and archiving practices, including: documenting data using standard metadata, file and data organization; using open and software-agnostic formats; and curating research data for reuse.

The examples and workflows in this guide can be adapted for data rescue of any individual-level tabular data, regardless of discipline.

This guide has been compiled by the Data Rescue Group, a subcommittee of the Ontario Council of University Libraries (OCUL)[1] Data Community (ODC). The Data Rescue Group is comprised of subject librarians and technical support staff interested in improving access to data that has been deemed at-risk and of value to the academic community. Covering data in our library collections, data produced by researchers, and data made available through government or other commercial vendors, the group actively curates and makes data available through shared research data repositories including the Ontario Data Documentation, Extraction Service and Infrastructure (<odesi>)[2] and Scholars Portal Dataverse[3]. Some of the activities the group coordinates include: development and maintenance of data inventories for major research data collections; improvement of data documentation and the creation of structured metadata records for datasets; data cleaning; and making data available online so it can be easily found by our researchers.

Working with at-risk datasets

For many decades, libraries and archives have collected and curated data for use in research. This work frequently involves at-risk datasets, which may require data rescue activities such as conducting research and outreach to obtain copies of lost documentation, creating new documentation (such as machine-readable syntax and metadata for use in statistical analysis software), conducting format conversions, migrating data to a long-term storage location, and providing improved web-based access to the data.

In Canada, there are many examples of library data rescue efforts. For example, Carleton University's Data Centre has been conducting data rescue for years, and is known for their work with data from government departments and other organizations that were closing their doors, e.g., the Canadian Information Office, the Millennium Scholarship Foundation, and the Centre

[1] https://ocul.on.ca/
[2] http://odesi.ca
[3] https://dataverse.scholarsportal.info/

for Research and Information on Canada (CRIC). Their rescue work with CRIC was described in a presentation at IASSIST in 2010. At Queen's University, the library has worked with local researchers to rescue data at risk of being lost in paper files and on office computers. One example was a database of cranial measurements, covering the Arctic and Northwestern North America as well as Northeast Asia, Eurasia, Africa and the South Pacific, collected over a 30 year period. The library converted the data first to Excel, then to SPSS, and then published it in Dataverse. Another recent example occurred at the University of Alberta, where library staff rescued a weather dataset collected by the Alberta Research Council between 1957 and 1991. This data were then used in a secondary research project showing how they can be used for a map-based visualization. The code used to parse the data is also available for re-use.

Recognizing the importance of this kind of work, several organizations have formed groups to discuss the unique issues involved in data rescue and to collaboratively work on data rescue projects. For example, the Research Data Alliance (RDA)[4] has a data rescue group and in Canada, the Ontario Council of University Libraries (OCUL) Ontario Data Community formed their Data Rescue Group in 2016. These efforts are not limited to the library community. There are several data rescue efforts in the environmental sciences realm, such as Atmospheric Circulation Reconstructions over the Earth (ACRE) and the World Meteorological Organization (WMO).

At-risk data are often (although not always) older data, and frequently have been separated from the data creator, making it difficult to track down lost contextual information. Additionally, in some cases, some data degradation has already occurred before the data makes its way to a curating institution. Depending on factors like these, there are decisions which must be made about the level of curation that is feasible or desirable, and how accessible the data can be made for reuse. It is important that the communities involved in data rescue projects document the best practices they are developing so that others can learn from their work. Currently, there is a lack of resources and skills available to perform data rescue work in many organizations. Therefore, the availability of support resources may help make it possible for organizations to embark upon their own data rescue projects.

Fundamentals: rescuing data files & metadata

Some data rescue projects will primarily focus on making sure the data files have been moved out of an obsolete file format or storage medium, and that a minimum level of documentation is available to enable use of the file. This can be considered a base level data rescue, i.e., the minimum that must be done to ensure the data file can be accessed in the future. The exact steps needed to be taken here will depend on the initial form of the data and what state the data are in, but the goal of basic data rescue is to assemble and preserve a data collection and associated crucial information concerning the dataset while that information is still available, either directly from the dataset creators or in the form of documentation. These activities will

[4] https://www.rd-alliance.org/

stabilize a dataset to prevent future loss and will preserve all the information necessary for a future data professional (or expert user) to prepare the data for analysis.

Basic data rescue has two main components: preserving the computer files in a form that ensures they will continue to be available and readable in the future; and ensuring that sufficient documentation (in all available languages) is provided alongside the data. A third issue which may arise is that of privacy or confidentiality.

Bit-level preservation is the process of transferring datasets and associated files from unstable media (personal hard drives, CD-Roms, tapes) to long-term, stable storage that is redundant and regularly backed up, such as a trusted data repository. Hardcopy documentation should be scanned to electronic format. Bit-level preservation is necessary but not sufficient, as potential format obsolescence also needs to be addressed.

Data rescue should also involve reviewing the available documentation to determine whether it is sufficient to support secondary research. At a minimum, there should be a codebook which describes the contents of the dataset, including the number of cases and a description of all variables, including their data type and possible values. The codebook should be checked to ensure that it matches the data file. Ideally, additional documentation should also be available, including a copy of the questionnaire (for surveys), a user guide which describes data collection methodology in detail, and frequency distributions for each of the variables to allow for verifying the integrity of the dataset.

A full disclosure review[5] to assess potential privacy and confidentiality issues in the data will be beyond the expertise of most people carrying out basic data rescue, but some precautions can be taken. These precautions include checking the dataset and documentation to ensure that direct personal identifiers (e.g., names, addresses, Social Insurance Numbers) are not present, and if they are, removing them from any version of the data to be made publicly available. Indirect personal identifiers should typically be aggregated (e.g., specific geographic locations aggregated to broad geographies, and socio-economic variables like age or income aggregated into categories). A quick assessment of the sensitivity of the data should also be undertaken to determine if the data include information that should be kept confidential, such as personal health information or opinions on controversial topics. If any of these confidentiality issues are present, a fuller disclosure review should be undertaken by an expert; otherwise, it is advisable to not make the data publicly available even if direct identifiers are not present.

Case study 1: Retrieving data from a hard drive

A survey data file is currently stored on the office hard drive of a professor who is retiring. The survey data are currently saved in SAS 6.14 file format, with some variable and value labels. Documentation includes a paper copy of the original survey questionnaire, with variable names

[5] Readers looking for more information on disclosure and risk may refer to The Anonymisation Decision Making Framework (https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf).

and value codes handwritten next to the questions, a detailed description of the survey protocol in an article reporting the initial findings, and several files of output containing unpublished frequencies.

Steps for basic data rescue:

1. Import the data into a current version of SAS or other statistical software (i.e. SPSS or Stata).
   - Export as an ASCII text file, plus accompanying command codes (also as a text file) to allow the file to be read into statistical software in the future.
2. Assess and compare the data and the available documentation. Basic assessment elements include:
   - Do the variables in the survey match the questions in the codebook?
   - Does the population described match the number of cases?
3. Scan the paper copy of the original survey to PDF.
4. Create descriptive metadata to facilitate discovery and meaningful reuse.
   - Preferably to be done in a structured open format, such as XML.
   - This format can then be interpreted by a descriptive metadata standard. The primary one used by many data archives is the Data Documentation Initiative (DDI).
5. Create a file archive consisting of the data file, statistical software command codes, the PDF of the survey, the metadata file and the output frequencies.
6. Perform a basic confidentiality review and consult with the data creator before considering further steps, such as sharing with a data archive.
7. Deposit the package into a data repository, ideally one which supports the creation of basic metadata according to the Data Documentation Initiative (DDI) metadata standard.

Case study 2: Retrieving data from an obsolete storage medium

The Energy Statistics Handbook (ESH) is a Statistics Canada product which provides energy analysts, economists, environmentalists, researchers, and the general public with a single source of consistent energy-related data. Between 1996 and 2001, the ESH came on two 3 1/4" floppy disks, containing software which must be installed in order to access the data tables. A data extraction project was undertaken to remove the data off the floppy disks. Some of the years could be accessed on Windows XP, but others were not compatible, so it was necessary to use a Windows emulator to accomplish this project.

Steps for basic data rescue:

1. Set up an emulator for the relevant outdated operating system (e.g., Windows 98 or an earlier DOS-based version of Windows).
2. Using the emulator, install the software provided on the floppy disks.
   - Please use extra caution when the 2 disk installation is done using an ***external floppy disk drive***. The installation may not work unless you unplug and reconnect the external floppy disk drive after installing 1st floppy disk. If this step is not performed, chances are your PC may not recognize that a new floppy disk is inserted in the floppy disk drive.
3. Use the software to extract the data files and documentation.

4. If the data files themselves are extracted in an outdated file format, migrate the data to an ASCII text file.
5. Proceed with the steps described in case study 1 above.

Long-term preservation

The activities described above are part of the process of digital preservation, and are activities designed to ensure the long-term access and reuse of digital data, mitigating potential threats that may arise resulting from software, hardware, and media format obsolescence, hardware and software failures, and natural disasters. Digital preservation looks to the long-term, however, and as such, a one-time data rescue is not sufficient. Organizational preservation policies are important to ensure clarity about what future preservation activities the "rescuing" organization is committed to following.

Ensuring long-term preservation of research data can be challenging. As seen in our case studies, data are often collected and analysed using proprietary tools and software. However, open formats for data do exist and are recognized. These formats include the ASCII format - such as text (.txt) or comma separated values (.csv) - which store data in a flattened, plain text format that can be easily read into other software and tools used by researchers. Another major challenge is that data usually require detailed metadata to facilitate software readability and for meaningful reuse by another researcher. While basic data rescue may only go so far as to digitize the documentation generated by the data creator (such as user guides or codebooks), the ideal scenario would be to create formal, structured, machine-readable metadata records in an open textual format (e.g., XML). There are a variety of descriptive metadata standards for data, such as the Data Documentation Initiative (DDI), which will be described more in the next section on improving data access. Preservation metadata standards, such as PREMIS[6], are also relevant for tracking preservation actions such as data migrations.

Additional Resources

There are many resources available describing digital preservation principles, practices, and tools in much greater detail, including advice on file formats, migrations, confidentiality issues, and many other topics. Here are a few key resources:

- Sustainability of Digital Formats: Planning for Library of Congress Collections
- Digital Preservation Coalition - Digital Preservation Handbook
- MIT Libraries - Digital Preservation Management Workshop
- ICPSR Resource on Data Confidentiality

Data curation to improve access

Ideally, a data rescue project will go beyond the basics of ensuring the data file and documentation are available and can be read using the most up-to-date computers. To make data collections truly accessible requires turning documentation into a machine-readable format

[6] https://www.loc.gov/standards/premis/

that can be combined with the data files using computer technology. For example, information about variables and their values can be extracted from a PDF data dictionary and used to create code files which can be read by statistical software such as SPSS or SAS. This allows users to generate a data file that can be more meaningfully utilized within their statistical software of choice.

Data rescue projects also frequently involve selecting an access platform to use for dissemination of the data. This step will greatly increase the number of researchers who will be able to find the data. The process of loading rescued data into an access platform frequently involves the creation of machine-readable metadata records. One standard commonly supported by data repositories in the social sciences (such as Dataverse and Nesstar[7]) is the Data Documentation Initiative (DDI). Like an SPSS or SAS code file, a DDI record contains detailed information about the variables and their values, which the repository platform can read and combine with the data file in order to offer data exploration, visualization, and conversion to a range of file formats for data download.

The OCUL context

In Ontario, the ODC's Data Rescue Group endeavours to undertake significant value-added activity when conducting data rescues, while recognizing that perfection is rarely possible. The remainder of this section describes the detailed step-by-step workflow undertaken by the Data Rescue Group to rescue a sample data file from a collection of data files available on the Canadian federal government's open data portal. Basic data rescue had already been undertaken by Library and Archives Canada, which migrated the data files from tape to an ASCII text format and digitized any existing documentation. The Data Rescue Group undertook the next step in the rescue process, that is, to generate machine-readable metadata and make the data available in a more accessible and user-friendly way - in this case through OCUL's <odesi> data portal (which is based on the Nesstar data repository software).

We are sharing these steps with others to help further knowledge exchange in this area. We have endeavoured to make these steps relatively tool agnostic. Although references to SPSS are included because that is part of our workflow, these steps could be undertaken with any statistical software package. For institutions using Nesstar software, Appendix 2 contains the process for generating codebooks from Nesstar Publisher.

Data Processing Workflow

- **Download data and documentation.** These files can include, but are not limited to, the data file, Codebook, Questionnaire, Data Dictionary, and any other information about the dataset.
- **Create syntax file.** This is done if the data file is an ASCII text file and there is an accompanying codebook or data dictionary.

[7] http://www.nesstar.com/

- Copy and paste content of Codebook (record layout section) to Excel or text editor to clean up text and create the command code file (i.e., SPSS command code file).
  Note: this could be done with any statistical software package; we are most familiar with SPSS.
- Clean up text (this can take a lot of time).
- Use cleaned-up text to create the command code file.
- Missing values – go through Codebook to determine what values are missing.
- Run syntax file against data file.
  - If there is information about frequencies in the Codebook, check these against the new data file.
  - If not, indicate in the introduction to the User Guide & Data Dictionary (see Appendix 2 for example of text) that no published frequencies exist, and that the frequencies shown herein should be used at the researcher's own risk.
- **Create metadata file** (DDI compliant XML file)
  - This step is typically accomplished using software specially designed for creating and editing metadata records using the Data Documentation Initiative (DDI) metadata standard. The Data Rescue Group uses the Nesstar software for this purpose, however, other available tools are listed on the DDI Alliance homepage.
  - With Nesstar and similar software tools, typically the first step is to load the data file into the DDI editor software, allowing the software to read all the variable information from the data file. Variable names, value labels, frequencies, and other information can all be automatically extracted from the data file and included in the DDI metadata record.
  - Once the data file had been loaded into Nesstar and the DDI metadata record begun, use the available documentation to fill out additional DDI fields with descriptive information about the survey (i.e., all the metadata that can't be extracted from the data file itself). Include references to the documentation file and related publications.
  - Group variables based on documentation. If there is no guidance for grouping, use these basic groups:
    - Administration - variables about the survey: interview number, day of interview, etc.
    - Demographic - province, sex, marital status, language, education, etc.
    - Household - household information
    - Occupation
    - Income
    - Survey questions - group survey questions by logical breaks in the questions; this may be laid out in the original Codebook
    - Derived Variables (if applicable)
    - Weight(s) - weight variable(s), always the last group listed
  - Make note of any issues in creating the dataset. These can include: not having a data dictionary or list of frequencies to compare your dataset to; problems with weights or any other variables; etc.

- **Related publications**
  Search for other publications based on the dataset - on the internet or in a library catalogue. Add any references found to the metadata record. If a copy of the document is found, link it to the metadata record.

Completed Studies

Once the data files and documentation are compiled, the study is published to the data repository (<odesi> in OCUL's case).

The following accompanying files are also published to the repository:

- User Guide
- Data Dictionary
- Questionnaire
- Original Codebook and data file (from the producer's website)
- Command code file
- any additional documents related to the dataset (related publications, documentation)

Additional Resources

For more detail on OCUL's practices for creating DDI metadata and publishing data in <odesi>, please see the following documents:

- <odesi> Best Practices Document
- How to Mark-up in Nesstar
- Data Documentation Initiative (DDI)

Further value-added enhancements for data reuse

For a data professional with significant experience working with social science survey data, it is possible to go even further in cleaning the data to help researchers avoid common pitfalls and misinterpretations of the data. Some examples of value-added tasks include:

- Declaring missing values
- Coding open-ended questions
- Restructuring multiple response questions
- Recoding variables

If the decision is made to undertake any of these activities on behalf of future researchers, it is extremely important to document all value-added work that was done, with clear indication of who did the work and the date it was done. A copy of the dataset in its original state, as provided by the data creator, should continue to be made available for examination.

Appendix 1: Workflow for Files from Government of Canada Open Data website

This workflow was taken from how the *Alcohol Consumption Survey, 1978/11*, was created and published to <odesi> . For other files from the Open Data website the files available and the process may be different.

In this workflow, the following software was used. There are other software packages available that can be used to do the same thing; these are what the Data Rescue Group used.

- SPSS: to create the command code file
- Excel and/or Notepad++[8]: to convert the PDF Codebook into a text file that can then be used with SPSS Syntax to create the SPSS data file
- Nesstar Publisher: to create the metadata record. Other tools for creating DDI compliant XML files are listed on the DDI Alliance homepage.

Download data and documentation from the Open Data website[9]

These files can include, but are not limited to, the data file (usually a text file), Codebook (usually in pdf format), Description of file (usually HTML file)

Create Syntax file (command code file)

If the Open Data website only has a data file and codebook (or data dictionary), you have to create a syntax file to create a usable data file.

1. Copy and paste content of Codebook (record layout section) to Excel or Notepad++ to clean up text and create the SPSS syntax file.
2. Clean up text (this can take a lot of time).
3. Use cleaned up text to create the SPSS syntax file.
    - Missing values – go through Codebook to determine what values are missing; this is not always documented in the Codebook.
4. Run syntax file against data file.
    - If there is information about frequencies in the Codebook, check these against the new data file.
    - If not, indicate in the introduction to the User Guide & Data Dictionary (see below) that no published frequencies exist, and that the frequencies shown herein should be used at the researcher's own risk.

Create Metadata file

1. Create metadata file (DDI compliant XML file).
   Using documentation available, list all information about the survey, any related publications, etc.

[8] Notepad++ can be downloaded here - https://notepad-plus-plus.org/downloads/
[9] https://open.canada.ca/en/open-data

2. Group variables based on documentation. If there is no guidance for grouping, use these basic groups:
    - Administration - variables about the survey – interview number, day of interview, etc.
    - Demographic - province, sex, marital status, language, education, etc.
    - Household - household information
    - Occupation
    - Income
    - Survey questions - group survey questions by logical breaks in the questions; this may be laid out in the original Codebook
    - Derived Variables (if applicable)
    - Weight(s) - weight(s) variable – this is always the last group listed
3. Make note of any issues in creating the dataset. These can include: not having a data dictionary or list of frequencies to compare your dataset to; problems with weights or other variables; lack of documentation to verify; etc.

Create User Guide

User guides can be created using Word or Nesstar Publisher. Using Word you have to create the Codebook separately using the SPSS data file; with Nesstar Publisher, the Codebook is created automatically.

- Word:
  Using documentation from the Open Data website, create a User Guide listing all the information about the survey, any related publications, Topic Index, and Codebook (based on new data file). Include any notes about issues or problems encountered while recreating the data file. This can include notes about missing information (e.g. not having original frequencies to compare to the new data file frequencies), missing weight variable, or any other changes made to improve usability of the dataset. See next section, *Special Notes to add to metadata record*, for example of notes.
- Nesstar Publisher:
  If you are using Nesstar Publisher you can create a Codebook based on the metadata
    - Open file you are creating the codebook for
    - Click on *Documentation → Generate PDF…*
    - A PDF file containing all the metadata and data in the file will be created.

Special Notes to add to metadata record

If there are any issues with the data file you created, make sure to add notes in the metadata record listing the issue and any warnings.

Examples:

- Note on Data File:
  Add this note if there are no published frequencies to compare the new dataset to.
  Example - *The ASCII data file and Codebook were downloaded from the Open Data website (http://open.canada.ca/data/en/dataset/fa9def2b-37f0-4984-89f9-bb30f312e9df)*

on August 10, 2015. Using this Codebook, syntax was created and run against the ASCII data file. The SPSS data file created is provided here. Use with caution. There was no documentation to double check the frequencies created by this data file.

- Note about Open Government licence:
  Example - *The ASCII data file and Codebook are made available under the Open Government Licence – Canada (http://open.canada.ca/en/open-government-licence-canada).*
- Note about the Weight variable:
  In some datasets the weight variable does not appear to work properly. If this is the case, add a note to explain. Below is an example from the **Alcohol Consumption in Canada: A National Study, 1978/11**:
  Use with caution. When applied, the weighted sample size does not correspond to the adjusted size of the Canadian population at the time of the survey. Data from CANSIM table 051-0001 puts the population of Canada for ages 15-89 (figures for the size of the population of 90 year-olds was not available) at 18,151,792 for 1978. Reducing this figure by accounting for the exclusions yields figures of 16,881,167. The weighted population total, according to this survey, comes to 200,845.

Completed Studies

Once the data files and documentation are compiled, the study is published to <odesi>.
The following files are published (depending on the study; additional files may also be added):

- User Guide
- Data Dictionary
- Questionnaire
- Original Codebook and data file (from Open Data website)
- Syntax file
- Any additional documents related to the dataset (related publications, documentation)

Appendix 2: Creating Codebooks & Data Dictionaries

Creating a codebook from your dataset provides users an overview of the dataset. The Codebook is useful documentation to include when archiving a dataset because it gives users a quick reference to the frequencies. Codebooks can be created in Nesstar Publisher or in statistical software (such as SPSS, SAS, or Stata).

Codebooks created in Nesstar Publisher

Codebooks created in Nesstar Publisher contain the metadata and data set information that is included in the Nesstar Publisher record. Data set information includes question text, universe and notes added to individual variables, skip patterns, value labels, frequencies, and any other notes added to the record.

To create a codebook in Nesstar Publisher

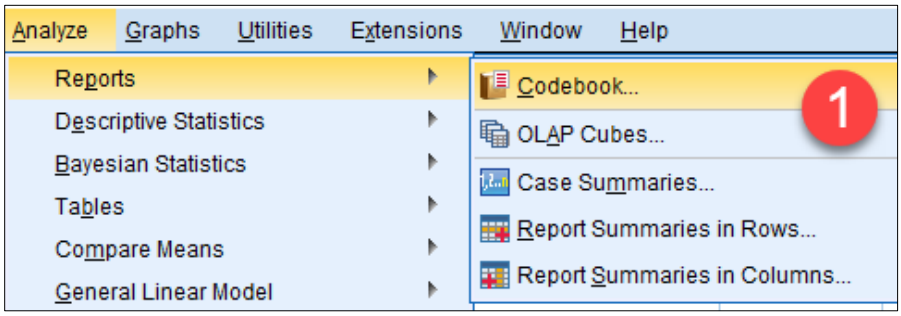
- Open file you are creating the codebook for
- Click on *Documentation → Generate PDF…*
- A PDF file containing all the metadata and data in the file will be created.

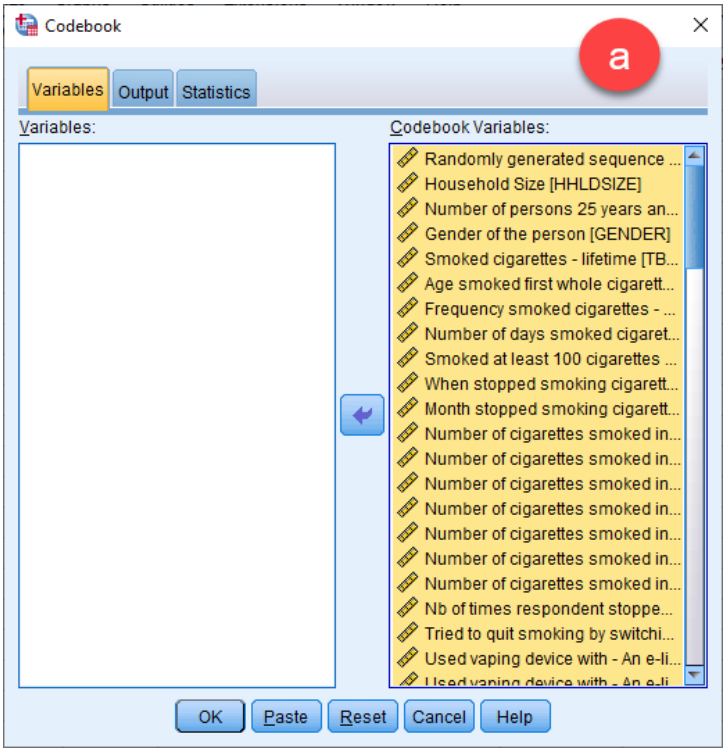
Codebooks created using SPSS

In SPSS - open dataset

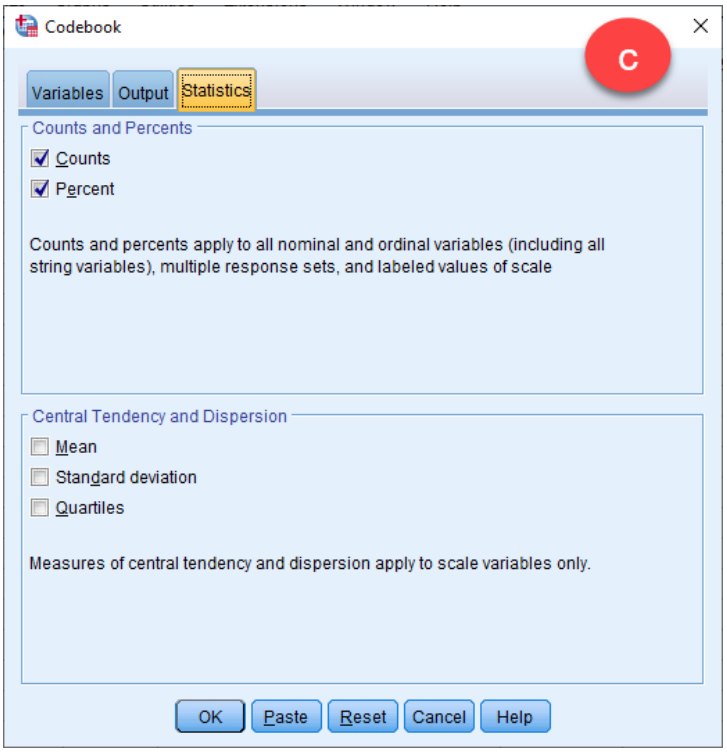
1. Analyze → Reports → Codebook…
    a. Choose the variables
    b. Choose the statistics
    c. Choose the Output
    d. Export to Excel
    With the codebook open in the output window.
        - File → Export
        - File Name – browse to where the file is to be saved (this will default to last file saved)
        - OK
    e. Export to Word
        - If your columns are too wide in Excel, the tables won't display properly in Word. We have found these to be the best sizes:
            - Column A – 10.71
            - Column B – 18.71
            - Column C – 50.29
            - Column D, E – 8.71
        - Left-align column C
        - Select the whole codebook
            - Format → AutoFit Row Height – makes sure all the rows are the same size
            - Wrap text

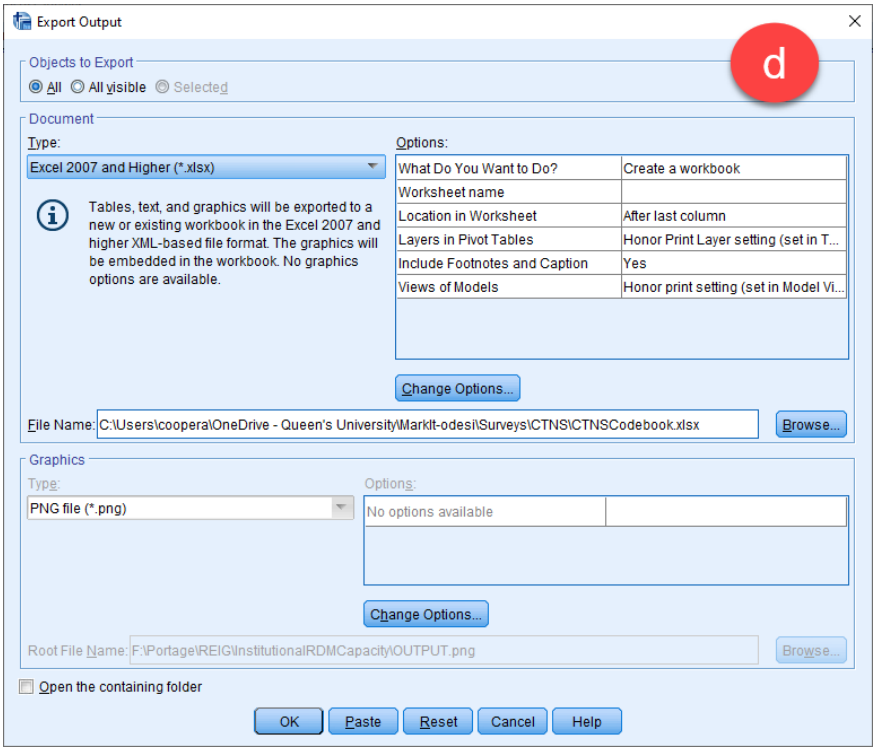
- In Word, set all margins to 0.5 inches
- Select all in Excel and paste into Word
- Add section headings to match the variable grouping in the documentatio






Codebook
b
Variables
Output
Statistics
Variable Information
Position
Label
Type
Format
Measurement level
Role
Value labels
Missing values
Custom attributes
Reserved attributes
You must select at least one Variable Information option.
File Information
File name
Location
Number of cases
Label
Documents
Weight status
Custom attributes
Reserved attributes
Variable Display Order
Display order
Variable List
Ascending
Descending
Maximum Number of Categories
Suppress display of value labels, counts and percents for variables with more than n categories
Maximum number of categories:
200
OK
Paste
Reset
Cancel
Help


Codebook
c
Variables
Output
Statistics
Counts and Percents
Counts
Percent
Counts and percents apply to all nominal and ordinal variables (including all string variables), multiple response sets, and labeled values of scale
Central Tendency and Dispersion
Mean
Standard deviation
Quartiles
Measures of central tendency and dispersion apply to scale variables only.
OK
Paste
Reset
Cancel
Help


Export Output
d
Objects to Export
All
All visible
Selected
Document
Type:
Excel 2007 and Higher (*.xlsx)
Tables, text, and graphics will be exported to a new or existing workbook in the Excel 2007 and higher XML-based file format. The graphics will be embedded in the workbook. No graphics options are available.
Options:
What Do You Want to Do?
Create a workbook
Worksheet name
Location in Worksheet
After last column
Layers in Pivot Tables
Honor Print Layer setting (set in T...
Include Footnotes and Caption
Yes
Views of Models
Honor print setting (set in Model Vi...
Change Options...
File Name:
C:\Users\coopera\OneDrive - Queen's University\MarkIt-odesi\Surveys\CTNS\CTNSCodebook.xlsx
Browse...
Graphics
Type:
PNG file (*.png)
Options:
No options available
Change Options...
Root File Name:
F:\Portage\REIG\InstitutionalRDMCapacity\OUTPUT.png
Browse...
Open the containing folder
OK
Paste
Reset
Cancel
Help

Codebooks Created using SAS

- Self-generating Codebooks Using SAS

Codebooks Created using Stata

- Stata manual

Codebooks Created using R

- CRAN Codebook tutorial
- R Markdown from R Studio

Appendix 3: Glossary

At-risk data

Without curation, all data are at-risk of becoming unavailable for future use. Risks include obsolescence of storage media or file format, data degradation, or lack of documentation.

Bit-level preservation

The process of transferring datasets and associated files from unstable media (personal hard drives, CD-Roms, Tapes) to long-term, stable storage that is redundant and regularly backed up, such as a trusted data repository.

Codebooks

Codebooks are typically prepared and used in Social Science research. They include information on research/survey design and methodology, response and non-response codes for each variable, missing values, original questionnaire/data collection instrument.

Command files

Syntax files or command code files are text files that contain the program/code to run against the set of data values in order to analyse them. Creating syntax or command code is an excellent way to document your data analysis process. The syntax can be recycled to run similar analysis for other projects and can be easily modified or customised based on project requirements.

Data curation

“A managed process, throughout the data lifecycle, by which data/data collections are cleansed, documented, standardized, formatted and inter-related. This includes versioning data, or forming a new collection from several data sources, annotating with metadata, adding codes to raw data (e.g., classifying a galaxy image with a galaxy type such as "spiral"). Higher levels of curation involve maintaining links with annotation and with other published materials. Thus a dataset may include a citation link to publication whose analysis was based on the data. The goal of curation is to manage and promote the use of data from its point of creation to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Special forms of curation may be available in data repositories. The data curation process itself must be documented as part of curation. Thus curation and provenance are highly related.” (CASRAI).

Data dictionary

Data dictionaries provide definitions or descriptions of data fields present in a data file. A data dictionary contains information on variables, subjects or observations as well as identifies data elements and their attributes such as names, definitions and units of measure and other relevant information.

Data repositories

Repositories “preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse” (CASRAI)

Some examples of repositories that OCUL uses include:

- <odesi> (Ontario Data Documentation, Extraction Service and Infrastructure).
    - Uses Nesstar as its underlying repository software. Nesstar’s Publisher software is a freely available tool for editing DDI metadata records.
- Scholars Portal Dataverse.
    - Uses the The Dataverse Project as its underlying repository software.

Data rescue

Data rescue is the process of acting to ensure that at-risk data is not permanently lost. Data rescue may involve conducting research and outreach to obtain copies of lost documentation, creating new documentation (such as machine-readable syntax and metadata for use in statistical analysis software), conducting format conversions, migrating data to a long-term storage location, and providing improved web-based access to the data.

Digital preservation

“A term that encompasses all of the activities required to ensure that the digital content designated for long-term preservation is maintained in usable formats, for as long as access to that content is needed or desired, and can be made available in meaningful ways to current and future users.” (ICPSR)

Documentation & metadata

“A crucial part of ensuring that research data can be used, shared and reused by a wide-range of researchers, for a variety of purposes, is by taking care that those data are accessible, understandable and usable. This requires clear data description, annotation, contextual information and documentation that explains how data were created or digitised, what data mean, what their content and structure are, and any manipulations that may have taken place. Creating comprehensive data documentation is easiest when begun at the onset of a project and continued throughout the research.” (UK Data Service)

Data documentation can be prepared at 3 different levels -- project level; file or dataset level; and variable or item level. Types of documentation include codebooks, data dictionaries, readme files, syntax files, and metadata records.

Metadata

Metadata may include elements of the information contained in all of the above documentation types, but the information is stored in a standardized, structured, machine-readable form (e.g., XML). Metadata are “typically used for resource discovery, providing searchable information that

helps users to find existing data, as a bibliographic record for citation, or for online data browsing" (UK Data Archive).

Readme files

Readme files are typically simple text files that describe the core documentation about a research and its data files. A readme file may describe the individual file(s) with file(s) format and/or an entire dataset as a whole and may contain any relevant information about specific software used in data collection and dissemination.

Syntax files

Syntax files or command code files are text files that contain the program/code to run against the set of data values in order to analyse them. Creating syntax or command code is an excellent way to document your data analysis process. The syntax can be recycled to run similar analysis for other projects and can be easily modified or customised based on project requirements.