DATA
CURATION
NETWORK

Data Curation Primer: SPSS

Topic	Description
File Extensions	.sav .por .sps .spv/.spo
MIME Type	`application/x-spss-sav` `application/x-spss-por`
Structure	.sav: Proprietary binary format, with metadata; sections include header, dictionary, and observations. .por: Portable file format that saves data and metadata as ASCII text. .sps: Plain-text file format that records SPSS syntax, to recreate analyses. .spv/.spo: Proprietary data formats containing outputs (e.g. tables, charts, visualizations) generated by analytic functions run in SPSS (.spv = v.16 and later; .spo = v.15 and earlier).
Versions	25.0 Most recent version as of 12/21/2018. See release notes: http://www-01.ibm.com/support/docview.wss?uid=swg27049975
Primary fields or areas of use	Social sciences, psychology, education, health sciences, and survey data
Source and affiliation	SPSS Statistics is a software package initially created by SPSS Inc. It was acquired by IBM in 2009 and currently it is named IBM SPSS Statistics.
Metadata	The SPSS Dictionary (also called "code book") is part of the SPSS data file (.sav, .por) and it holds all metadata, specifically, .sav: Can contain names and labels for variables, an unformatted textual description, and an extension record with attributes. .por: Can contain names and labels for variables and an unformatted textual description.
Key questions for curation	● What version of SPSS were these files created with?

Suggested Citation: Sai Deng, Joshua Dull, Jeanine Finn, Shahira Khair. (2019). SPSS Data Curation Primer. Retrieved from the University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/202810.

This work was created as part of the Data Curation Network "Specialized Data Curation" Workshop #1 co-located with the Digital Library Federation (DLF) Forum 2018 in Las Vegas, Nevada on October 17-18, 2018. See also: Primers authored by the workshop attendees at DLF. http://datacurationnetwork.org.

review	Which is the latest version at the time of curation? Which versions are currently supported by the software producers and common operating systems? ● Are the variables well-described (with labels, etc.) within SPSS variables table? Is there an external codebook or data dictionary? ● Is there an additional README file describing project level information and data information? ● Is the data “native” to SPSS, or was it exported to SPSS from another statistical package? ● Which files (e.g., data, syntax, output) are essential to effectively sharing this data? Which are optional? ● What is the profile of potential re-users of this data? ● Is this data likely to be only of immediate use and interest, or is there a “longer tail” of potential reuse? ● Is a complete copy of the survey or interview instrument available as a separate file?
Tools for curation review	● SPSS ● Smartreader for SPSS Statistics ● PSPP ● ViewSav ● R (Haven and rio packages) ● SAS (SAS v.9.1.3 and later) ● STATA (USESPSS)
Date Created	12/21/2018
Created by	Sai Deng, University of Central Florida - sai.deng@ucf.edu Joshua Dull, Yale University - joshua.dull@yale.edu Jeanine Finn, Claremont Colleges - jeanine.finn@claremont.edu Shahira Khair, University of Victoria - skhair@uvic.ca
Date updated and summary of changes made	04/08/2019 - updated to include comments from peer review process

Table of Contents

Description of Format

While this primer primarily discusses .sav and .por files, there are other possible file formats which are typically associated with the software SPSS (See Appendix A).

1. **SPSS Statistics** (.sav): Data files saved in IBM SPSS Statistics format.
    a. Proprietary, binary format that contains both the data (observations) and the description of the data (metadata).
    b. File is divided into sections: a header (file-level metadata), a dictionary (variable names and labels), and the data itself (observations).
    c. Data files saved in .sav format cannot be read by earlier versions of SPSS prior to version 7.5. Data files saved in Unicode encoding cannot be read by releases of the software prior to version 16.0. .sav files may be created with other encodings including ASCII and UTF-8.
2. **Portable** (.por)**:** Portable format that can be read by other versions of IBM SPSS Statistics and versions on other operating systems.
    a. In most cases, saving data in portable format is no longer necessary, since .sav data files should be platform/operating system independent.
    b. Saving a file in portable format takes considerably longer than saving the file in .sav format.
    c. .por files can be opened by the free/open source version of SPSS, called PSPP**.**

Example Data

Examples below link to study in ICPSR which offers SPSS as a download option.

1. United States. Bureau of Justice Statistics. Survey of Inmates in State and Federal Correctional Facilities, [United States], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-12-19. https://doi.org/10.3886/ICPSR04572.v5
2. Cohen, Deborah (Deborah Ann). Evaluation of the Balance Calories Initiative, 2016 Baseline, Alabama and Mississippi. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-12-17. https://doi.org/10.3886/ICPSR37110.v1

Start the Conversation: Broad Questions and Clarifications on Research Data

Before going into technical details specific to the SPSS data, having a short conversation with researchers about their area of research, and data practices is helpful.

The main assumptions about SPSS are that it is used by researchers for quantitative analysis and for producing graphical representations of their data. After talking to researchers, our group now has a better understanding of how some researchers analyze, save and share their data with SPSS.

Is SPSS used by many researchers doing quantitative analysis? How is SPSS compared to similar statistical programs?
Researchers interviewed indicated that SPSS is commonly used by new or early career researchers, who are often introduced to the software in their graduate programs. Many research institutions have access to the software, so it is commonly used for instruction of students. Fields of research in which SPSS is commonly used include the social sciences, psychology, and education.

The researchers interviewed primarily used more advanced software for statistical analysis, including STATA, SAS, and R. For example, one researcher interviewed used SAS to conduct research with data from a Learning Management System (LMS). For this project, they obtained raw data from their institution's LMS describing student grades and related information, which was spread across many relational databases and other sources, like web sites, textual documents. SAS was used to organize this data into a single table for further analysis. Another researcher in statistics also used SAS for research in optimal design. In comparison, SPSS provides a more user-friendly interface.

Researchers noted the proprietary (and expensive) nature of SPSS software as part of their motivations for increasingly working with open source platforms like R alongside of SPSS. Many times research projects involve work with outside collaborators who may not have institutional access to SPSS, so the ability to work across software platforms is key.

What kind of data do researchers generate or import into SPSS?
Researchers interviewed used a range of structured data inputs with SPSS, including tabular data and relational databases, with information from a variety of sources: web data, survey responses, model simulations, etc.

One researcher described his research using a diary study for a project in cognitive psychology. The participant diaries were coded by hand, and the codes were input as tabular data into SPSS. The researcher then used regression analysis and structural equation modeling tools within SPSS to look for patterns across the diaries.

How do researchers document their data in SPSS?
SPSS automatically generates three main outputs: the data file (*.sav), the codebook, and the syntax file. The code book describes variables and data contained in the data file, while the syntax describes the analysis process. The codebook and syntax file are important metadata to capture along with the data, but may not be sufficient to capture the context, methodology and provenance of the dataset being created..

Researchers generally described making use of SPSS functions to document their data, rather than creating separate README files. This includes creating thorough and descriptive variable and value labels, and saving complete syntax files (sometimes with additional comments).

What kind of data outputs are researchers able or willing to share from SPSS?
Researchers interviewed primarily export their data file created in SPSS into a non-proprietary format (e.g., csv), if they intend to share it or analyze it with another piece of software. Researchers

also reported sharing the syntax file to describe table structures and methods of analysis. Codebooks were also considered essential for sharing data. Based on the interviews, while the practices of researchers vary, the curators can include essential data for the datasets. If needed, the curator can generate data from the datasets the researchers provided such as variable list and data dictionary.

Key Questions

These are "reflection" questions for the curator or curation team to review while looking over the dataset. The answers (or lack of answers) will help determine what kind of clarifications might be needed from the researchers.

1. What are the file formats of the data files received? Check for a data file and codebook (.sav), and syntax file (.sps).
2. What version of SPSS were data files created with, compared to the current version at the time of curation?
3. Is use of a file-naming convention evident? Are the file names understandable?
4. Are the variables contained in the data files well-organized and sufficiently described (e.g., unique and understandable labels, etc.)? Is there an external codebook or data dictionary?
5. Is there an additional README file containing necessary metadata (e.g., project context, methodology, and data information)?
6. What does the data contained represent? Is additional documentation necessary for interpretation? (e.g., if the data represents survey results, is a complete copy of the survey or interview instrument available as a separate file?)
7. Which files (e.g., data, syntax, output) are essential to effectively sharing this data? Which are optional?
8. What is the profile of potential re-users of this data? Is there additional information they would need to know to reuse this data?

Key Clarifications

If answers to the following are not addressed in any associated metadata or documentation, follow-up with the researcher for further clarification.

Data Analysis and Curation

1. What are the dates of data collection and analysis?
2. How was missing data handled?
3. Does the dataset contain any imputed data?
4. Were data "cleaned" using a tool such as OpenRefine or R packages?
5. What other tools (if any) were used to gather or analyze this data in addition to SPSS?

Sensitive Data

6. If data pertains to human subjects, have data been sufficiently anonymized?
7. If anonymized data are being shared, what steps have been taken to prevent re-identification of participants? (e.g., are geographic aggregations used in order to help prevent re-identification of human subjects data, etc.)? Is there any data contained that could allow for re-identification?
8. Have funder and institutional human subjects protocols been followed and is the deposit of anonymized data permitted?

Other Questions

9. What are the best practices (or common practices) for this type of dataset among disciplinary or institutional repositories the researcher is aware of?

Applicable Metadata Standards, Recommended Elements and Readme File

It is recommended to document datasets created in the research lifecycle, and use software programs and tools to assist in data documentation. A dataset or project created with SPSS Statistics software need to include two levels of documentations: project level (or study level) metadata, and data level metadata.

1. **Project Level or Study Level Metadata**

    a. The project-level or study-level metadata information is separate supporting documentation. It can be saved as a README text file and/or included as record metadata in digital repositories for data archiving.
    b. The study-level metadata includes the research context and design, data collection methods, structure of data files, secondary data sources, data validation procedures and modifications made to data, and information on data confidentiality, access and use conditions (if applicable). Key documentation can serve as sources for this information, for example, project reports, lab books, questionnaires or interview guides used in surveys or interviews, as well as publications.
    c. It is recommended to assign or capture descriptive, technical, administrative, structural and preservation metadata for the dataset in digital repositories. It is also important to provide a unique identifier for the dataset (e.g., DOI, purl, handle) and make sure that the data meets citation requirement (if applicable).
    d. This metadata may be created or collected by referring to different metadata standards or schemas. Based on the examination of some datasets and a review of several metadata standards, such as Dublin Core (DC) and Data Documentation Initiative (DDI), the group would like to recommend a list of elements to be considered for documenting SPSS research data (See Appendix B).
    e. To document more detailed information on the dataset, DDI as a metadata specification for the social and behavioral sciences can be followed. It is an XML metadata standard for documenting numeric data, and has advantages such as recording variable-level information. Detailed information is available at: http://www.ddialliance.org (See Appendix C).

2. **Data Level Metadata**

    a. The IBM SPSS Statistics software can create embedded documentation for the dataset. Data files in the software contain embedded metadata that describes and defines the data in the file. This data level metadata can be exported.
    b. The most important metadata for the SPSS data includes:

        i. **Variable name**: the name assigned to the variable that acts as an identifier. (Required)

ii. **Variable label**: descriptive information of the meaning of the variable.
iii. **Variable type**: information on how the value is stored internally (e.g., numeric, string). (Required)
iv. **Value label**: descriptive information on how the variable is coded (e.g., 0 for male, 1 for female).
v. **Missing value**: information on values to be ignored in calculations.

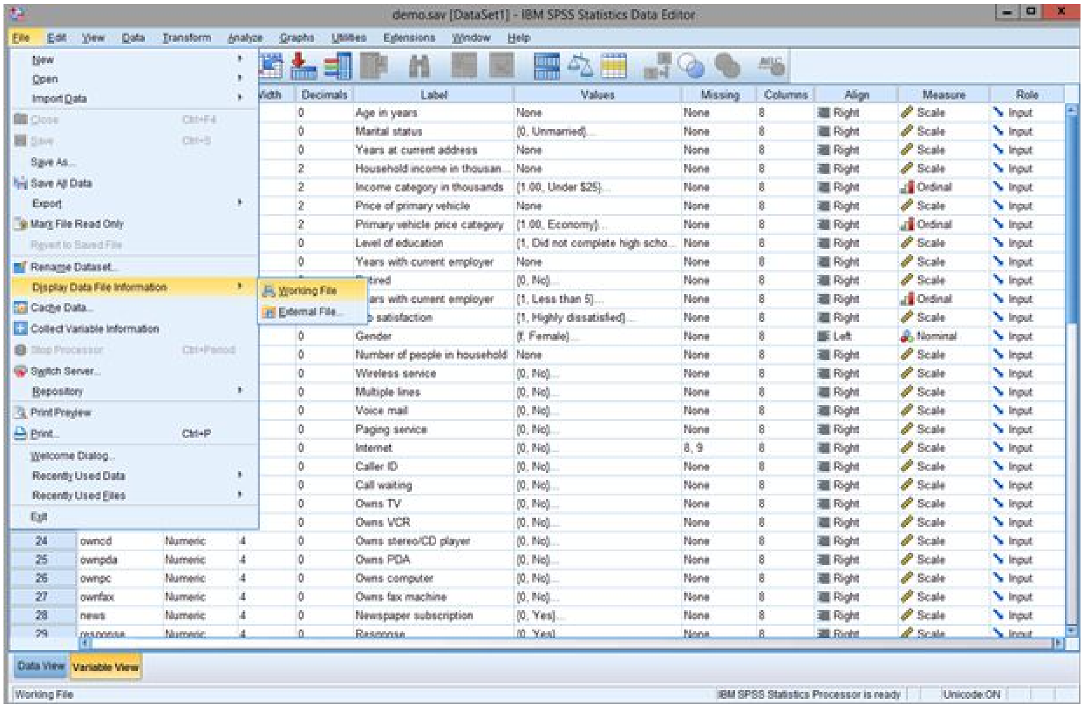
c. Other metadata information includes:

i. **Width**: the maximum number of characters that a value can have. (Required)
ii. **Decimals**: information on how to display numeric values. (Required)
iii. **Columns**: Column width for a variable. (Required)
iv. **Align**: Alignment of data values. (Required)
v. **Measure**: how the variable is measured (e.g., nominal, ordinal, scale).
vi. **Role**: the variable's supposed relation to other variables.

3. A **codebook**, or a **data dictionary** can be created from a SPSS data file.

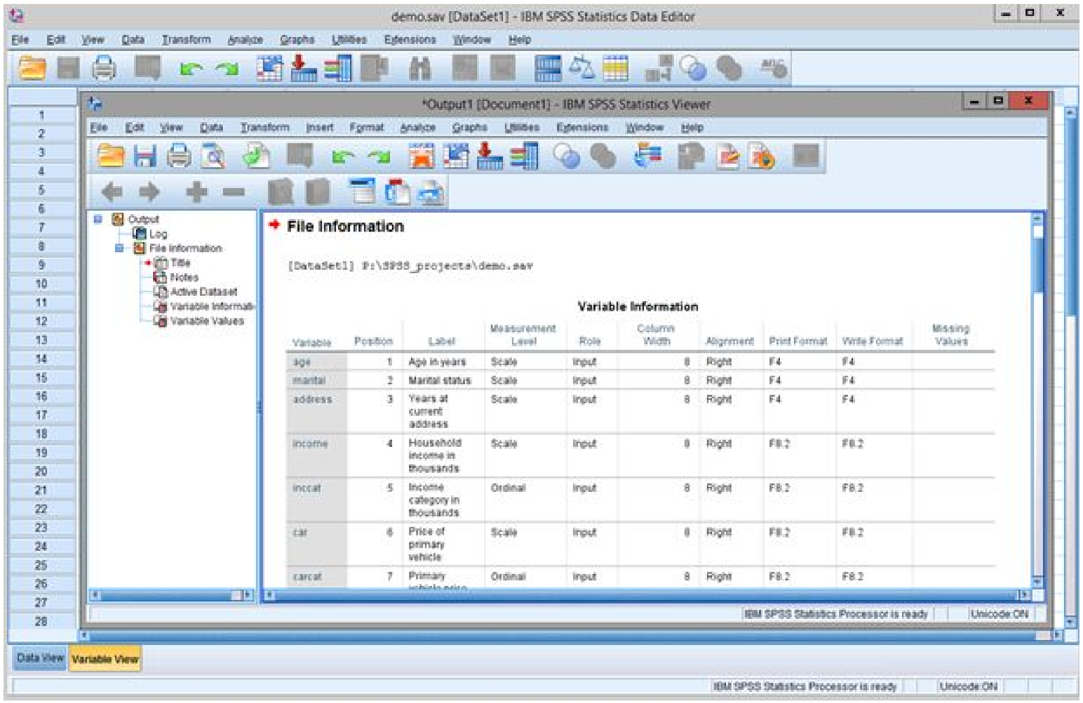
a. The data dictionary contains metadata that describes various properties of the data file.
b. Researchers are recommended to provide data dictionaries for their datasets. If this information is not provided, the curator can export one from the dataset project file. The data dictionary can be saved as a .pdf file or a text file. The following steps can be followed to export a data dictionary.

In the data file, click *File* > *Display Data File Information* > *Working file*, for metadata attributes of the variables to be displayed and printed to the Output Viewer window (See Figures 1 and 2). Using "Display dictionary" command in the syntax window can produce the same result.



			Vidth	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
				0	Age in years	None	None	8	Right	Scale	Input
				0	Marital status	{0, Unmarried}...	None	8	Right	Scale	Input
				0	Years at current address	None	None	8	Right	Scale	Input
				2	Household income in thousan...	None	None	8	Right	Scale	Input
				2	Income category in thousands	{1.00, Under $25}...	None	8	Right	Ordinal	Input
				2	Price of primary vehicle	None	None	8	Right	Scale	Input
				2	Primary vehicle price category	{1.00, Economy}...	None	8	Right	Ordinal	Input
				0	Level of education	{1, Did not complete high scho...	None	8	Right	Scale	Input
				0	Years with current employer	None	None	8	Right	Scale	Input
					tired	{0, No}...	None	8	Right	Scale	Input
					ars with current employer	{1, Less than 5}...	None	8	Right	Ordinal	Input
					satisfaction	{1, Highly dissatisfied}...	None	8	Right	Scale	Input
				0	Gender	{f, Female}...	None	8	Left	Nominal	Input
				0	Number of people in household	None	None	8	Right	Scale	Input
				0	Wireless service	{0, No}...	None	8	Right	Scale	Input
				0	Multiple lines	{0, No}...	None	8	Right	Scale	Input
				0	Voice mail	{0, No}...	None	8	Right	Scale	Input
				0	Paging service	{0, No}...	None	8	Right	Scale	Input
				0	Internet	{0, No}...	8, 9	8	Right	Scale	Input
				0	Caller ID	{0, No}...	None	8	Right	Scale	Input
				0	Call waiting	{0, No}...	None	8	Right	Scale	Input
				0	Owns TV	{0, No}...	None	8	Right	Scale	Input
				0	Owns VCR	{0, No}...	None	8	Right	Scale	Input
24	owncd	Numeric	4	0	Owns stereo/CD player	{0, No}...	None	8	Right	Scale	Input
25	ownpda	Numeric	4	0	Owns PDA	{0, No}...	None	8	Right	Scale	Input
26	ownpc	Numeric	4	0	Owns computer	{0, No}...	None	8	Right	Scale	Input
27	ownfax	Numeric	4	0	Owns fax machine	{0, No}...	None	8	Right	Scale	Input
28	news	Numeric	4	0	Newspaper subscription	{0, Yes}...	None	8	Right	Scale	Input
29	response	Numeric	4	0	Response	{0, Yes}...	None	8	Right	Scale	Input

Figure 1: In the data file, click *File* > *Display Data File Information* > *Working file*, to display and print its variable information to the Output Viewer window.


demo.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Log
File Information
Title
Notes
Active Dataset
Variable Informati
Variable Values
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

File Information

```
[DataSet1] F:\SPSS_projects\demo.sav
```

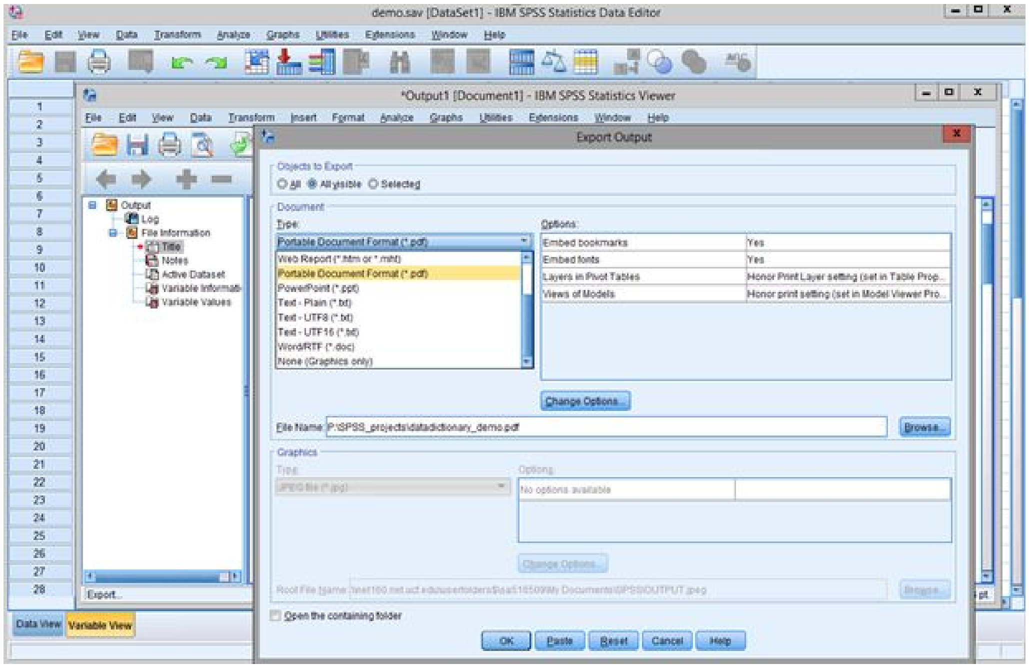
Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format	Missing Values
age	1	Age in years	Scale	Input	8	Right	F4	F4	
marital	2	Marital status	Scale	Input	8	Right	F4	F4	
address	3	Years at current address	Scale	Input	8	Right	F4	F4	
income	4	Household income in thousands	Scale	Input	8	Right	F8.2	F8.2	
inccat	5	Income category in thousands	Ordinal	Input	8	Right	F8.2	F8.2	
car	6	Price of primary vehicle	Scale	Input	8	Right	F8.2	F8.2	
carcat	7	Primary	Ordinal	Input	8	Right	F8.2	F8.2	

Figure 2: Variable information is printed to the Output Viewer window.

c. The variable information can be saved as Viewer Files (.spv), SPSS Web Report (.htm) or Cognos Active Report (.mht), by clicking *File* > *Save as* in the Output Viewer window. However, it is recommended to export it as a .pdf or text file for archiving purpose.

To export the variable information as a .pdf file, in the Output Viewer window, click *File* > *Export*, and the “Export Output” window will open. In this opened new window, under “Objects to Export,” select “All visible” under the “Document” section, click the “Type” drop-down menu, and choose “Portable Document Format (.pdf)”, then in the “File name” section, click “Browse” to go to the directory you’d like to save the file to and create a filename (such as datadictionary_projectname.pdf), click “OK” (See Figure 3).


demo.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Extensions Window Help
Output
Log
File Information
Title
Notes
Active Dataset
Variable Information
Variable Values
Export...
Data View
Variable View
Export Output
Change Options...
Browse...
OK
Paste
Reset
Cancel
Help

Objects to Export: ○ All ● All visible ○ Selected

Document Type: Portable Document Format (*.pdf)

Web Report (*.htm or *.mht)
Portable Document Format (*.pdf)
PowerPoint (*.ppt)
Text - Plain (*.txt)
Text - UTF8 (*.txt)
Text - UTF16 (*.txt)
Word/RTF (*.doc)
None (Graphics only)

Options:	
Embed bookmarks	Yes
Embed fonts	Yes
Layers in Pivot Tables	Honor Print Layer setting (set in Table Prop...
Views of Models	Honor print setting (set in Model Viewer Pro...

File Name: P:\SPSS_projects\datadictionary_demo.pdf

Graphics Type: JPEG file (*.jpg)

Options: No options available

Root File Name: \\netl100.net.ucf.edu\userfolders$\sa516509\My Documents\SPSSOUTPUT.jpeg

☐ Open the containing folder

Figure 3: Variable information is saved as a data dictionary in PDF format.

d. Detailed customizable codebook can be generated via *Analyze > Reports > Codebook*.
e. The data dictionary can be displayed as a .pdf file, or listed under “variable” section if the digital repository provides such capability.
f. The data dictionary can also be represented as XML using the IBM created dictionary schema (http://xml.spss.com/spss/data/dictionary-1.0.xsd). The dictionary schema is an XML representation of the data dictionary (See Appendix D).

Tutorials

1. UCLA Institute for Digital Research and Education - SPSS Learning Modules
   ○ Online guide covers fundamentals of using SPSS, inputting raw data into SPSS, and data management with SPSS.
2. Kent State University Libraries - SPSS Tutorials
   ○ Four online tutorials providing an introduction to the SPSS environment, guidance on data manipulation and cleaning in SPSS, and instruction on data analysis and interpretation. Offers sample data files for users to follow along on their own.
3. SPSS Beginners Tutorials
   ○ Online guide covers beginner basics, how to prepare data, and create a data dictionary. Includes guides on conducting a range of statistical tests (e.g. T-TEst, Chi-square, Correlation, ANOVA, Linear and Multiple Regression).
4. ICPSR - A Student’s Guide to Interpreting SPSS Output for Basic Analyses
   ○ Slide deck of annotated SPSS outputs for guidance on interpretation of a range of statistical tests.

Software

1. SPSS
   - A commercial software package used for statistical analysis and data management. Outputs proprietary file formats: .sav (data files) and .sps (syntax files); as well as non-proprietary text-based formats that can be read by other statistical software.
2. Smartreader for SPSS Statistics
   - A free application developed by IBM, which allows users to view and modify SPSS output files. It does not require an SPSS installation and does not require a licence.
3. PSPP
   - A non-commercial software package for statistical analysis and data management. It is designed as a free alternative to the proprietary SPSS software, with much of the same capabilities. Outputs non-proprietary file formats.
4. ViewSav
   - An open-source program for SPSS data files, written by Karel Asselberghs (University of Amsterdam). Allows users to read .sav files, explore variables, display frequencies and summary statistics, and view code books.
5. CRC32SAV
   - A small, open-source program for SPSS data files, written by Karel Asselberghs (University of Amsterdam). Computes checksums for separate parts of a SPSS data file: file header, dictionary and data. This allows users to check how well the data and data dictionaries of separate files match.
6. R
   - A free software environment for statistical analysis, graphics, and database manipulation, based on the R programming language. Can read SPSS data saved in non-proprietary text-based formats (e.g., .csv). The Haven and rio packages can be installed in R, in order to enable R to read and write .sav files.
7. SAS
   - A commercial software platform used for advanced statistical analysis. SAS provides a graphical point-and-click user interface and more a command-line interface for more advanced use using the SAS language. SAS version 9.1.3 and later can import .sav files.
8. STATA (USESPSS)
   - A SPSS-compatible program written by Sergiy Radyakin (Development Economics Research Group, World Bank), for users of STATA, another commonly used proprietary statistical software package. Allows STATA users to load .sav files from Windows and UNIX/Mac platforms, and preserves variable and value labels.

Preservation Actions

There are advantages to keeping data files generated in SPSS in its original .sav file format; it maintains metadata, like variable labels, with the dataset and can be important for replicating analysis. However, It is also recommended that curators convert .sav files to plain-text file formats (e.g., .por, .csv) for the purpose of better long-term preservation. If possible, repositories should store the original .sav file and the plain-text copy.

1. **Recommendations for preservation**
    a. CESSDA
        i. For quality assurance, CESSDA recommends curators export a plain-text version as soon as possible after receiving files so they can confer with the researchers, if necessary.
        ii. The .por format is more suitable for preservation than the .sav format.
        iii. "The only sure means of preservation for the long term is converting the binary files to plain text (.csv in ASCII or Unicode). Only plain text gives the digital archive full control over the data, without being dependent on external parties. We are recommending conversion to plain text to the CESSDA organisations."[1]
    b. Library of Congress
        i. The advantage of .sav format include incorporating metadata and variable descriptions within data file.
        ii. Disadvantages of .sav are reliance on proprietary software that might not exist in the future, the data is not transparent, and can often be compressed.
    c. ICPSR
        i. Provide data in ASCII format with appropriate setup files.
        ii. Provide data in .sav format.
        iii. Provide data in portable (.por) format.
    d. Dataverse
        i. On ingest, Dataverse will create plain-text versions of supported data files.
        ii. "Dataverse stores the raw data content extracted from such files in plain text, TAB-delimited files. The metadata information that describes this content is stored separately, in a relational database, so that it can be accessed efficiently by the application."[2]
        iii. Files created by proprietary software, like SPSS, are not ideal for archival preservation and dataverse does not guarantee the ability to process all SPSS files in the manner explained in the above section (d.i.).

2. **Options for reading & converting SAV files:**
    a. Modules exist for R to import SPSS .sav files; see rio | Import, Export, and Convert Data Files and Read SPSS (SAV & POR) files.
    b. Starting with SAS 9.1.3 SP3 (2005), SAS has had the ability to import SPSS_sav files.
    c. USESPSS is a user-written Stata module, running only on Windows and without support, to import SPSS (.sav) datasets.
    d. Stat/Transfer, a popular commercial utility for converting datasets from one format to another, can read and write SPSS_sav files.

3. **Data Archives Preferred File Formats**
    a. ICPSR
        i. ICPSR accepts and distributes datasets in .sav and .por formats.
        ii. ICPSR distributes data in plain text with added scripts (setup files) for creating binary files for the statistical software packages to work with.
    b. The UK Data Archive

[1] See the full report here: https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf

[2] "What happens during ingest?" http://guides.dataverse.org/en/4.11/user/tabulardataingest/ingestprocess.html#what-happens-during-this-ingest

        i. Recommends the .por format or delimited text file with a command file for tabular data with extensive metadata.
        ii. Lists .sav format as acceptable but not recommended for long-term preservation.
    c. GESIS archive
        i. Preferred formats for a dataset include .por and .sav.
    d. Data Archive and Networked Services)
        i. DANS lists the .sav and .por formats as preferred.
    e. CESSDA
        i. Prefers data is provided in plain text (.tab, .csv, ASCII) with setup files for SPSS.
    f. Dataverse
        i. Supports ingest of tabular data via SPSS (.sav or .por), STATA, R, CSV, and Excel files.

FAIR Principles & SPSS

1. **Findable**
    a. Stored in appropriate disciplinary and/or institutional repositories.
    b. Dataset has been assigned a DOI.
    c. Dataset is described with appropriate metadata.
2. **Accessible**
    a. Repository is well structured and accessible.
    b. SPSS datasets include human-readable data and descriptive components (that don't require SPSS to open).
3. **Interoperable**
    a. Data is saved in widely-accessible non-proprietary (non-SPSS) format, if possible.
4. **Reusable**
    a. Usage licenses for SPSS and other packages are clear.
    b. Provenance is clearly indicated.
    c. Protection of human subject information has been observed and adequately communicated.
    d. Data cleaning an other processes of manipulations of data are clearly documented to support reproducibility.

Format Use

1. Various types of quantitative analyses use SPSS software.
2. SPSS also permits export to a number of other proprietary formats -- SPSS could be used as a "translational" format.
3. Multiple datasets might be used in support of meta-analyses and longitudinal studies in social science research.

Documentation of Curation Process

1. Details of any kind of "data cleaning" performed by curators (e.g., JSON file exported from OpenRefine).
2. Notes on changes made between versions or file formats (e.g., converting .sav file to .por file), including which software was used for the conversion.

Appendix A: Other SPSS File Formats

1. **SPSS Statistics Compressed** (.zsav): Opens data files that are saved in IBM SPSS Statistics compressed format.
    a. .zsav files have the same features as .sav files, but they take up less disk space.
    b. .zsav files may take more or less time to open and save, depending on the file size and system configuration. Extra time is needed to de-compress and compress .zsav files. However, because .zsav files are smaller on disk, they reduce the time needed to read and write from disk. As the file size gets larger, this time savings surpasses the extra time needed to de-compress and compress the files.
    c. Only IBM SPSS Statistics version 21 or higher can open .zsav files.
2. **SPSS/PC+** (.sys): Opens SPSS/PC+ data files.
    a. This option is available only on Windows operating systems. If the data file contains more than 500 variables, only the first 500 will be saved. For variables with more than one defined user-missing value, additional user-missing values will be recoded into the first defined user-missing value.
3. **Syntax Files** (.sps):
    a. SPSS syntax is a programming language that is unique to SPSS. It allows you to write commands that run SPSS procedures, rather than using the graphical user interface.
    b. Syntax allows users to perform tasks that would be too tedious or difficult to do using the drop-down menus. This is the case when you are re-running the same analysis many times, or doing complex transformations on data. Syntax also provides a record of how you transformed and analyzed your data, and allows you to instantly reproduce those steps at any time.
4. **Journal Files** (.jnl):
    a. By default, SPSS conveniently records the syntax for all of the commands run in SPSS (whether you used drop-down menus or syntax) in a Journal File (extension ".jnl"). This is convenient if you wish to review what commands you ran or if you want to edit or save the syntax commands for future use.
    b. To find out where SPSS is storing this Journal File, click Edit > Options. Click File Locations and you will see the pathname for the Journal File in the Session Journal area. You can also change the location where this file is stored.

Appendix B: Project Level or Study Level Metadata

A list of elements is recommended to document project level or study level metadata in the README file and/or the metadata record in the digital repository (if available and needed). This list is compiled based on research data characteristics and several metadata standards including DC and DDI.

Title: Title of the data collection. Mapped to dc:title.
Principal Investigator(s): The person, corporate body, or agency responsible for the work's intellectual content. Mapped to dc:creator.
Publisher: The person or organization responsible for the physical processes of the document. Mapped to dc:publisher.
Funding Agency: The source(s) of funds for production of the work. Mapped to dc:description or dc:description.sponsorship (if available).
Grant Number: The grant or contract number of the project. Mapped to dc:description or dc:description.sponsorship (if available).
Identifier: Unique string or number (producer's or archive's number), such as doi, handle number. Mapped to dc:identifier.
Rights: Copyright statement for the data collection. Mapped to dc:rights.
Citation: The citation information for the dataset. Mapped to dc.identifier.citation.
Subjects: The topic or broad category classification of the dataset. Mapped to dc:subject.
Description: Summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents including major variables, subject areas covered, and what questions the PIs attempted to answer when they conducted the study. Mapped to dc:description or dc:description.abstract.
Geographic Coverage: Geographic coverage of the dataset including the geographic scope of the data, and geographic coding provided in the variables. Mapped to dc:coverage.
Time Period: The time period covered by the dataset. Mapped to dc:coverage.
Date of Collection: Date when the data were collected. Mapped to dc.date.created.
Data Collection Notes: Methodology used in data collection. Mapped to dc:description.
Data Type(s): Types of data such as survey data, experimental data, psychological test, textual data, coded textual etc. Mapped to dc:type.
Methodology: Study purpose, study design, sample, time method, universe, unit(s) of observation, data source, data type(s), mode of data collection, description of variables, response rates, presence of common scales. Mapped to dc.description.
Data Source: The source of the data collection. Mapped to dc:source.
Other Study Description Materials: Other materials that are related to the study description, including appendices, sampling information, weighting details, methodological and technical details, publications based upon the study content, related studies or collections of studies. Mapped to dc:relation.
Language: Language of the study as well as the dataset. Mapped to dc:language.
Format: Type of data file (e.g., .sav, .sps., .spv, .por, .txt, .pdf, .doc, .xls, .xml, .jpg). Mapped to dc:format.
Original Release Date: The original release date of the dataset. Mapped to dc:date or dc:date.issued.
Data Update Information: Information on data updates, transformation, versioning, summarization, descriptions of migration and replication, and information about other events that have affected the files. Some of these administrative metadata are generated by the system. Can also include a description field for this information.
Data Preservation Information: More information on properties of data, the technical environment and fixity information. Can also include a description field for this information.
Data Files Description: Other technical information such as compression or encoding algorithms, encryption and decryption keys, software, hardware on which the data were, operating systems, application software, as well as file relationships. Can also include a description field for this information.

Appendix C: DDI Metadata

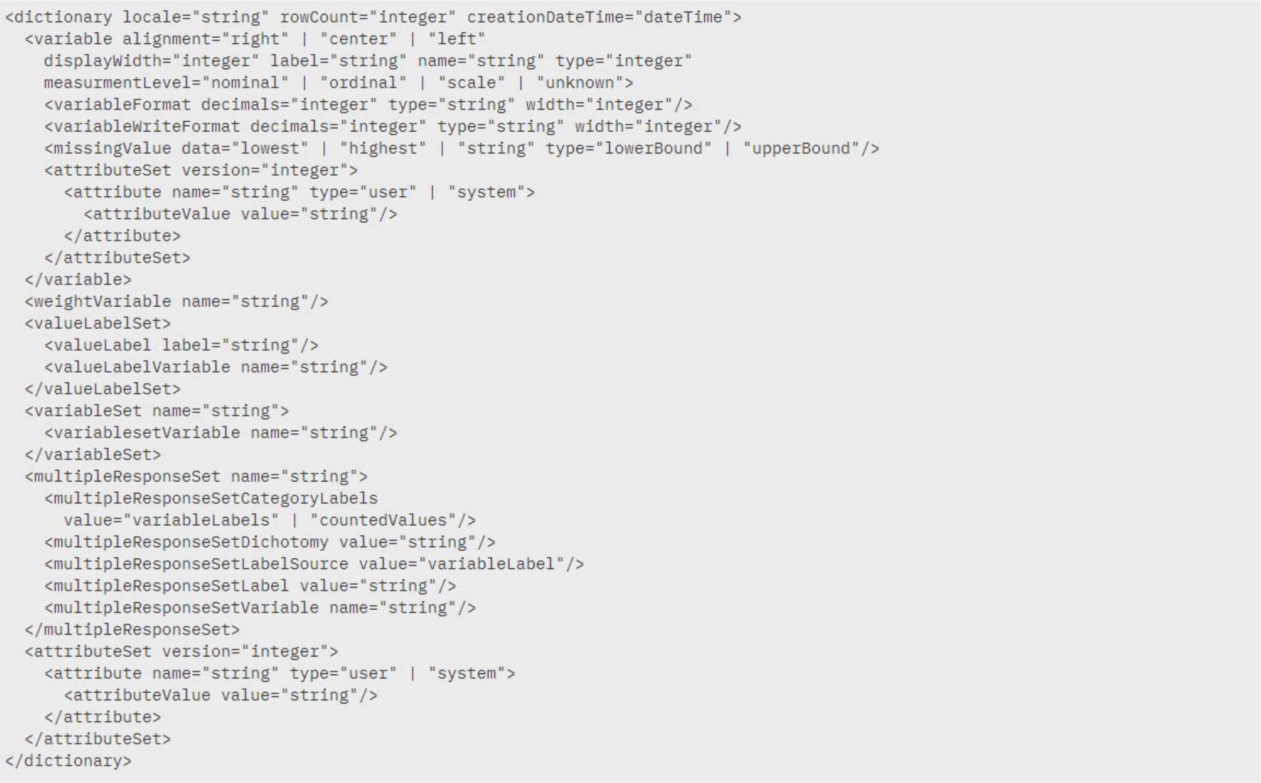
There are two DDI standards: DDI Lifecycle & DDI Codebook. Both are compatible as XML-Schema and include all the Dublin Core elements. Most DC elements can map directly to DDI (See DC-DDI Mapping Table at https://www.ddialliance.org/resources/ddi-profiles/dc). For information on creating DDI metadata *from SPSS, see* http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/paper/viewFile/144/120.

1. **DDI-Lifecycle** (DDI-L) is meant to record metadata across the entire research project including metadata on the project/study-level, publication, analysis, data files, and variables. DDI-L is also referred to as DDI 3 (current version 3.2).

2. **DDI-Codebook** (DDI-C) is a simplified version of DDI-L which works best for simple survey data. DDI-C is also referred to as DDI 2 (current version 2.5).

Appendix D: Dictionary Schema

An XML representation for the data dictionary can be exported from the software using its programming capability (if needed). The dictionary schema (http://xml.spss.com/spss/data/dictionary-1.0.xsd) is installed with the software (IBM Knowledge Center) and is shown in Figure 4.



```
<dictionary locale="string" rowCount="integer" creationDateTime="dateTime">
  <variable alignment="right" | "center" | "left"
    displayWidth="integer" label="string" name="string" type="integer"
    measurmentLevel="nominal" | "ordinal" | "scale" | "unknown">
    <variableFormat decimals="integer" type="string" width="integer"/>
    <variableWriteFormat decimals="integer" type="string" width="integer"/>
    <missingValue data="lowest" | "highest" | "string" type="lowerBound" | "upperBound"/>
    <attributeSet version="integer">
      <attribute name="string" type="user" | "system">
        <attributeValue value="string"/>
      </attribute>
    </attributeSet>
  </variable>
  <weightVariable name="string"/>
  <valueLabelSet>
    <valueLabel label="string"/>
    <valueLabelVariable name="string"/>
  </valueLabelSet>
  <variableSet name="string">
    <variablesetVariable name="string"/>
  </variableSet>
  <multipleResponseSet name="string">
    <multipleResponseSetCategoryLabels
      value="variableLabels" | "countedValues"/>
    <multipleResponseSetDichotomy value="string"/>
    <multipleResponseSetLabelSource value="variableLabel"/>
    <multipleResponseSetLabel value="string"/>
    <multipleResponseSetVariable name="string"/>
  </multipleResponseSet>
  <attributeSet version="integer">
    <attribute name="string" type="user" | "system">
      <attributeValue value="string"/>
    </attribute>
  </attributeSet>
</dictionary>
```

Figure 4: Documentation of dictionary schema provided by IBM.

Bibliography

Digital Curation Centre | because good research needs good data. (n.d.). Retrieved January 16, 2019, from http://www.dcc.ac.uk/

Digital Preservation Coalition. (n.d.). Quantitative File Formats for Preservation - Digital Preservation Coalition. Retrieved January 2, 2019, from https://www.dpconline.org/blog/quanititative-file-formats-for-preservation

Document your data. (n.d.). Retrieved January 16, 2019, from https://www.ukdataservice.ac.uk/manage-data/document

File formats — English. (n.d.). Retrieved January 2, 2019, from https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats/file-formats

GESIS - Leibniz Institute for the Social Sciences. (n.d.). Retrieved January 2, 2019, from https://www.gesis.org/en/services/archiving-and-registering/data-archiving/preparing-data-for-submission/

Guide to Social Science Data Preparation and Archiving, Phase 6: Depositing Data. (n.d.). Retrieved January 2, 2019, from https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter6.html

Harvey, D. R. (2010). *Digital curation: a how-to-do-it manual*. New York: Neal-Schuman Publishers.

IBM Compatibility of SPSS files (.sav, .sps, .spv, .spo) between different versions - United States. (2016, September 7). [CT741]. Retrieved January 16, 2019, from https://www-01.ibm.com/support/docview.wss?uid=swg21480797

IBM Knowledge Center - Data file types. (2014, October 24). Retrieved January 2, 2019, from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_25.0.0/statistics_mainhelp_ddita/spss/base/data_file_types.html

IBM Knowledge Center - Dictionary Schema Overview. (2014, October 24). Retrieved January 16, 2019, from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/programmability_option/dictionary_schema_intro.html#dictionary_schema_intro

IBM Knowledge Center - IBM SPSS Statistics V25.0 documentation. (2014, October 24). Retrieved January 16, 2019, from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_25.0.0/statistics_kc_ddita/spss/product_landing.html

IBM Knowledge Center - Saving data: Data file types. (2014, October 24). Retrieved January 16, 2019, from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/base/savedatatypes.htm

Institute for Quantitative Social Science. (2019, February 19). *User Guide*. Retrieved April, 1 2019, from http://guides.dataverse.org/en/4.11/user/index.html

Integrating Colectica with IBM SPSS Data Collection. (n.d.). Retrieved from http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi14/paper/viewFile/144/120

Recommended formats. (n.d.). Retrieved January 2, 2019, from https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

SPSS Dictionary. (n.d.). Retrieved January 16, 2019, from https://www.spss-tutorials.com/spss-dictionary/

SPSS Portable File, ASCII encoding. (2017, May 22). [web page]. Retrieved January 2, 2019, from https://www.loc.gov/preservation/digital/formats/fdd/fdd000468.shtml

SPSS System Data File Format Family (.sav). (2017, June 4). [web page]. Retrieved January 2, 2019, from https://www.loc.gov/preservation/digital/formats/fdd/fdd000469.shtml

Yeager, K. (n.d.). LibGuides: SPSS Tutorials: Using SPSS Syntax. Retrieved January 16, 2019, from https://libguides.library.kent.edu/SPSS/Syntax