



Course for Doctoral Students

RESEARCH DATA MANAGEMENT AND OPEN DATA

25th July 2015, Social Science Data Archives,
Faculty of Social Sciences, University of Ljubljana

ECPR Summer School 2015



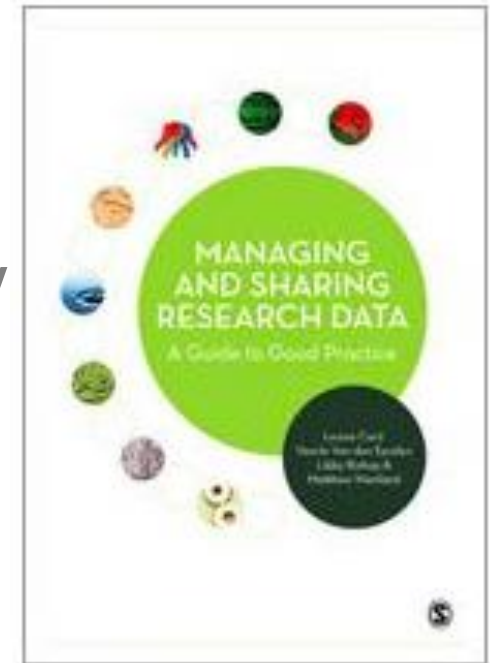
PREPARING DATA AND DOCUMENTATION FOR DIGITAL CURRATION

Irena Vipavc Brvar, Social Science Data Archives



Content

- Which things should I save and how
 - Documentation
 - Data
 - Metadata (standards)
- What tools are there



UK Data Service



SHARING MY RESEARCH

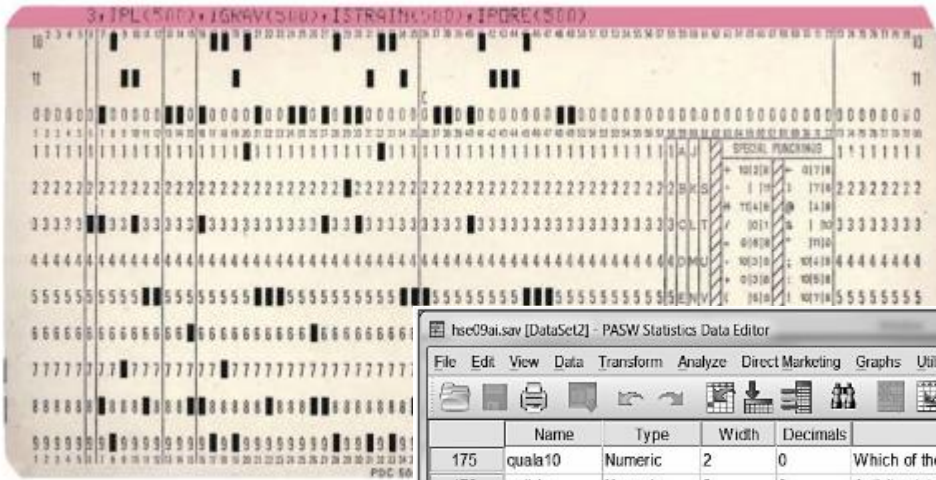
Data should be user-friendly, shareable and with long-lasting usability.

-> ensure they can be understood and interpreted by any user

This requires clear data description, annotation, contextual information and documentation.



CAN YOU UNDERSTAND/USE THESE DATA?



hsc09ai.sav [DataSet2] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 --1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 --1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 --1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 --1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 --1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 --1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 --1
182	wdlook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 --1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 --1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9.0, No a...	-99.0 --1.0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 --1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 --1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 --1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 --1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 --1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 --1
191	snumplee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 --1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 --1

Data View Variable View

Documentation

Data documentation might include:

- a survey questionnaire
- an interview schedule
- records of interviewees and their demographic characteristics in a qualitative study
- variable labels in a table
- published articles that provides background information
- description of the methodology used to collect the data

Nashville, Tennessee Music Industry Economic Impact Project

The Nashville Chamber of Commerce, in conjunction with the Music City Partnership and Belmont University, is conducting an economic impact study of the music industry in Nashville, Tennessee. This survey questionnaire is part of that study. The information received on this questionnaire will be held in strict confidence and no specific firm information will be released or published. In addition, all questionnaires and firm-specific information will be destroyed once the study is completed. If you have any questions, please contact Dr. Patrick Raines of Belmont University at 615-460-6175. Please return this survey by October 24, 2005. A stamped envelope is enclosed for your convenience.

Company Name _____
Contact Person _____
Telephone # _____ Fax # _____

Please specify primary nature of business _____

	Past Fiscal Yr.	Estimated Current Fiscal Yr.
Labor		
Total Number of Full-time Employees (Nashville, MSA) ^a	_____	_____
Total Number of Part-time Employees (Nashville, MSA)	_____	_____
Annual Employee Wages and Salaries	_____	_____
Professional/ Business Services, contract and additional labor	_____	_____
Operations		
Total Operating Expenditures ^b	_____	_____
Capital Improvements		
Total Capital Expenditures	_____	_____
Projected 3-Year Capital Expenditures	_____	_____
Taxes		
State Taxes	_____	_____
Local Taxes	_____	_____
Total Nashville Sales Revenues^c	_____	_____

Quality report

Source: UK Data Service



What should be captured?

Any useful documentation such as:

- final report, published reports, user guide, working paper, publications, lab books

Information on dataset structure

- inventory of data files
- relationships between those files
- records, cases...

Variable-level documentation

- labels, codes, classifications
- missing values
- derivations and aggregations

DMP



Start gathering meaningful information from as early on in the research process as possible.

What should be captured?

Contextual information about project and data

- background, project history, aims, objectives, hypotheses
- publications based on data collection

Data collection methodology and processes

- data collection process and sampling
- instruments used - questionnaires, showcards, interview schedules
- temporal/geographic coverage
- data validation - cleaning, error checking
- compilation of derived variables
- weighting: factors and variables, weighting process
- secondary data sources used

Data confidentiality, access and use conditions

- anonymisation carried out
- consent conditions/procedures
- access or use conditions of data

Source: UK Data Service

Data - level documentation

Certain types of data file may contain important information which should be preserved:

- variable/value labels; document metadata; table relationships and queries in relational databases; GIS data layers/tables

Some examples:

- SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
- MS Access: relationships between tables
- ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog
- MS Excel: document properties, worksheet labels (where multiple)

Source: UK Data Service

Data - level documentation: variable names

All structured, tabular data should have cases/records and variables adequately documented with names, labels and descriptions.

Variable names might include:

- question number system related to questions in a survey/questionnaire

e.g. Q1a, Q1b, Q2, Q3a

- numerical order system

e.g. V1, V2, V3

- meaningful abbreviations or combinations of abbreviations referring to meaning of the variable

e.g. oz%=percentage ozone, GOR=Government Office Region,

moocc=mother occupation, faocc=father occupation

- for interoperability across platforms - variable names should be max 8 characters and without spaces

Source: UK Data Service

Data - level documentation: variable labels

Similar principles for variable labels:

- be brief, max. 80 characters
- include unit of measurement where applicable
- reference the question number of a survey or questionnaire
e.g. variable 'q11hexw' with label 'Q11: hours spent taking physical exercise in a typical week' - the label gives the unit of measurement and a reference to the question number (Q11b)
- Codes of, and reasons for, missing data avoid blanks, system - missing or '0' values
e.g. '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'
- Coding or classification schemes used, with a bibliographic ref
e.g. Standard Occupational Classification 2000 - a list of codes to classify respondents' jobs; ISO 3166 alpha-2 country codes - an international standard of 2 - letter country codes

Data - level documentation: transcripts

Qualitative data/text documents:

- interview transcript speech demarcation (speaker tags)
- document header with brief details of interview date, place, interviewer name, interviewee details, context

METADATA

Metadata - data about data

Describe your survey using standard

International **standards/schemes**

- Data Documentation Initiative (DDI)
- ISO19115
- Dublin Core
- Metadata Encoding and Transmission Standard (METS)
- Preservation Metadata Maintenance Activity (PREMIS)

BASIC STRUCTURE OF DDI 2.*

- Section 1.0 - Document Description consists of bibliographic information that can be considered as the header whose elements uniquely describe the full contents of the compliant DDI file.
- Section 2.0 - Study Description consists of information about the data collection. This section includes information about who collected and who distributes the data, about the scope and coverage, sampling (if relevant), data collection methods and processing, citation requirements, etc.

Controlled Vocabulary

Multilingual

XML

Semantic and technical **interoperability**

BASIC STRUCTURE OF DDI 2.*

- Section 3.0 - Data Files Description provides information about the Data file(s).
- Section 4.0 - Variable Description provides a detailed description of variables, including (when relevant) the variable type, variable and value labels, literal questions, computation or imputation methods, instructions to interviewers, universe, descriptive statistics, etc.
- Section 5.0 - Other Study Related Materials allows for the inclusion of other materials related to the study such as questionnaires, user manuals, computer programs, interviewer manuals, maps, coding information, etc.

Your Dataset Deserves More than a First Row Header



The screenshot shows the Microsoft Excel interface with the COLECTICA ribbon active. The ribbon includes options for Document Workbook, Data Documentation, From SPSS, From Stata, Save as DDI, and Create Documentation. The Create Documentation dropdown menu is open, showing options for PDF and Word. The spreadsheet displays data for the '1978 Automobile Data' dataset, with the 'foreign' variable highlighted in the first row. The Data Documentation pane is open on the right, showing the variable details for 'foreign'.

	E	F	G	H	I	J	K	L	
1	headroom	trunk	weight	length	turn	displace	gear_ratio	foreign	
2		2.5	11	2930	186	40	121	3.58	0
3		3	11	3350	173	40	258	2.53	0
4		3	12	2640	168	35	121	3.08	0
5		4.5	16	3250	196	40	196	2.93	0
6		4	20	4080	222	43	350	2.41	0
7		4	21	3670	218	43	231	2.73	0
8		3	10	2230	170	34	304	2.87	0
9		2	16	3280	200	42	196	2.93	0
10		3.5	17	3880	207	43	231	2.93	0
11		3.5	13	3400	200	42	231	3.08	0
12		4	20	4330	221	44	425	2.28	0
13		3.5	16	3900	204	43	350	2.19	0
14		3	13	4290	204	45	350	2.24	0
15		2.5	9	2110	163	34	231	2.93	0
16		4	20	3690	212	43	250	2.56	0
17		3.5	17	3180	193	31	200	2.73	0
18		2	16	3220	200	41	200	2.73	0
19		2	7	2750	179	40	151	2.73	0
20		3.5	13	3430	197	43	250	2.56	0

Data Documentation

Dataset Details Variable Details Code List

foreign

Label
Car type

Description
Indicates whether the car was manufactured in the United States or a different country.

Data Type
Code

origin Codes Update Codes from Data Use Existing Code List

Additivity
Stack

Colectica for Excel

Document Variables and Datasets

Colectica allows documenting of Variables, Code Lists, and Data Sets directly from within Microsoft Excel.

Import Stata to Excel

Colectica for Excel Professional allows direct importing and documenting of Stata data files, with a file extension .dta. The variable names, labels and code lists in the Stata file will also be imported and added to the stored documentation automatically.

Metadata is Embedded

Colectica saves your standards-based metadata directly in the Microsoft Excel file. If you email or share your file, the metadata will still be attached.

Import SPSS to Excel

Colectica for Excel Professional allows direct importing and documenting of SPSS data files, with a file extension .sav. The variable names, labels and code lists in the SPSS file will also be imported and added to the stored documentation automatically.

Publish Documentation

Colectica for Excel can generate documentation for your Variables, Code Lists, and dataset in PDF, Word, HTML, and XSL-FO.

Create DDI-Lifecycle Metadata

Export your data documentation to an XML file in the DDI metadata format, the standard for data documentation. Open and edit it from Colectica Designer, Colectica Express, or other DDI applications.

Microsoft Excel ribbon showing the 'Data Documentation' button highlighted with an orange circle. The ribbon includes FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, and COLECTIC. The 'Data Documentation' button is located in the 'DATA' tab, between 'Document Workbook' and 'From SPSS'.

	A	B	C	D	E	F	G	H
1	ID	Age	Gender	Language				
2	1	30	1	1				
3	2	31	2	1				

Data Documentation

Dataset Details | Variable Details | Code List | ⚙️

Refresh Documentation

Column Count
4

Title
Simple Dataset

Subtitle

Alternate Title

Creator

Publisher

	H	I	J	K	L
1	length	turn	displace	gear_ratio	foreign
2	186	40	121	3.58	0
3	173	40	258	2.53	0
4	168	35	121	3.08	0
5	196	40	196	2.93	0
6	222	43	350	2.41	0
7	218	43	231	2.73	0
8	170	34	304	2.87	0
9	200	42	196	2.93	0
10	207	43	231	2.93	0
11	200	42	231	3.08	0
12	221	44	425	2.28	0
13	204	43	350	2.19	0
14	204	45	350	2.24	0
15	163	34	231	2.93	0
16	212	43	250	2.56	0
17	193	31	200	2.73	0
18	200	41	200	2.73	0
19	179	40	151	2.73	0
20	197	43	250	2.56	0

Data Documentation

Dataset Details | Variable Details | Code List | ⚙️

turn

Label
Turn Circle (ft.)

Description
The turning circle of a vehicle is the size of the smallest circular turn (i.e. U-turn) that the vehicle capable of making.

Data Type
Numeric

Type
Integer

Nesstar Publisher

Nesstar Publisher – a sophisticated authoring environment that can publish data from a variety of sources (including SPSS, SAS, Excel etc.). The tool includes a specialised metadata editor, data and metadata validation routines and metadata templates that provide standardisation and control.

Easy editing/creation and export of DDI documented datasets with XML experience needed.

Tools to compute/recode/label new, or existing, variables to be added to a dataset before publishing.

Tools to validate metadata and variables.

The ability to import and export data to the most common statistical formats, including delimited files.

The ability to include automatically generated frequency and summary statistics for each variable.





















Multilingual - Arabic, Chinese, English, French, Portuguese, Russian and Spanish and more.

File formats currently supported:

- Nesstar (.Nesstar)
- NSDstat (.NSDstat)
- DDI Document (*.xml)
- SPSS (*.sav)
- SPSS Portable (*.por)
- SPSS Syntax (*.sps)
- STATA (*.dta)
- Statistica (*.sta)
- NSDstat (*.nsf),
- dBase (*.dbf)
- DIF (*.dif)
- Delimited Text (*.txt, *.csv, *.sdv, *.cdv, *.prn)
- PC-Axis (*.px)
- Excel (*.xls)
- Hierarchy Definition File (*.NSDstatHDef)

File size limitations: The maximum size of file that can be imported is approximately 10 Gigabytes, with a limitation within a file to 260 million cases. However, using files of this size will affect response times.

Projects:

- [-] Citation - Production Statement
 -  Copyright
 -  Producers
 - Study Description**
 - [-] Citation
 -  Title
 -  ID Number
 -  Authoring Entity / Primary I
 -  Distributors
 -  Version
 - [-] Citation - Production Statement
 -  Producers
 -  Fundings
 - [-] Scope - Subject Information
 -  Keywords
 -  Topic Classifications
 - [-] Abstract
 -  Abstract
 - [-] Scope - Summary Data Descripti
 -  Countries
 -  Geographic Coverage
 -  Unit of Analysis
 -  Universe
 - [-] Methodology - Data Collection
 -  Time Method
 -  Sampling Procedure
 -  Mode of Data Collection
 -  Weighting
- [-] Other Study Materials

Study Description**Title**

[STUDKR12]3 Study circles 2012 : Monitoring activities

Authoring Entity / Primary Investigator

Name	Affiliation
Bogataj, Nevenka	Slovenian Institute for adult education

Distributors

Name	Abbreviation	Affiliation
Social Science Data Archives	ADP	University of Ljubljana

Producers

Name	Abbreviation
Slovenian Institute for adult education	ACS

Fundings

Agency	Abbreviation
Ministry of Education, Science and Sport	MIZS

Keywords

Text
adult educating
life-long learning
group learning
community learning

Variables

Number	Name	Label	Width	StartCol	EndCol	Record	Decimals
v42	V1_dr_r	drugo	200	*	*	1	0
v43	V2	Študijski krožek je del ne	33	*	*	1	0
v44	V2_2	ce odgovor da	8	*	*	1	0
v45	V2_2_dr_r	drugo	200	*	*	1	0
v46	V3	Kako ste nacrtovali izobr	8	*	*	1	0
v47	V3_dr_r	drugo	200	*	*	1	0
v48	V4	Kako ste nacrtovali akcij	7	*	*	1	0
v49	V4_dr_r	drugo	200	*	*	1	0
v50	V5_1	viru ustanove	8	*	*	1	0
v51	V5_2	viru ljudje	8	*	*	1	0
v52	V5_3	viru literatura	8	*	*	1	0
v53	V5_4	viru avdio posnetki	8	*	*	1	0
v54	V5_5	viru splet	8	*	*	1	0
v55	V5_6	viru okolje	8	*	*	1	0
v56	V5_7	drugo	8	*	*	1	0

Documentation

Statistics

Weights

Documentation

- Question
 - Pre-Question Text
 - Literal Question**
 - Post-Question Text
 - Interviewer Instructions
- Description
 - Variable Text
 - Concepts

Literal Question

3. Kako ste načrtovali IZOBRAŽEVALNE cilje? (Označite izbrani odgovor.) 5. drugo (Opišite.)

Category Hierarchy

Value: Label:

Category Text:

Level Name:

GeoMap URI:

Variable information

Data Type:

Measure:

Is Time Variable

Is Weight Variable

Min: Max: Decimals:

Implicit decimals