AIPs and documentation management: curating quantitative data

South-Eastern European Data Services

Irena Vipavc Brvar

ADP

SEEDS Workshop II

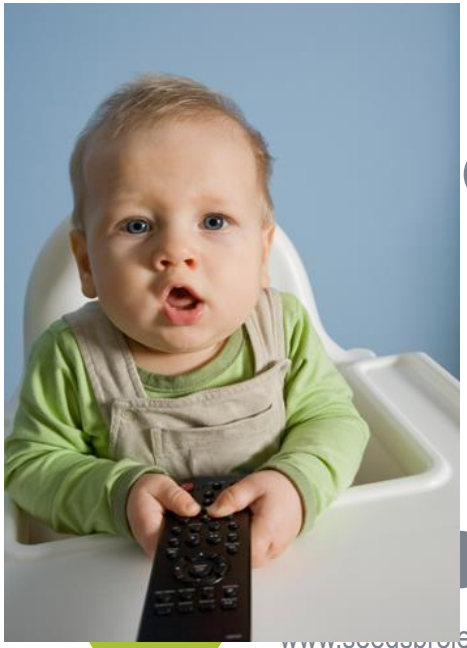
Ljubljana, 9.-11. February 2016

DATA FROM MY RESEARCH WILL BE SHARED

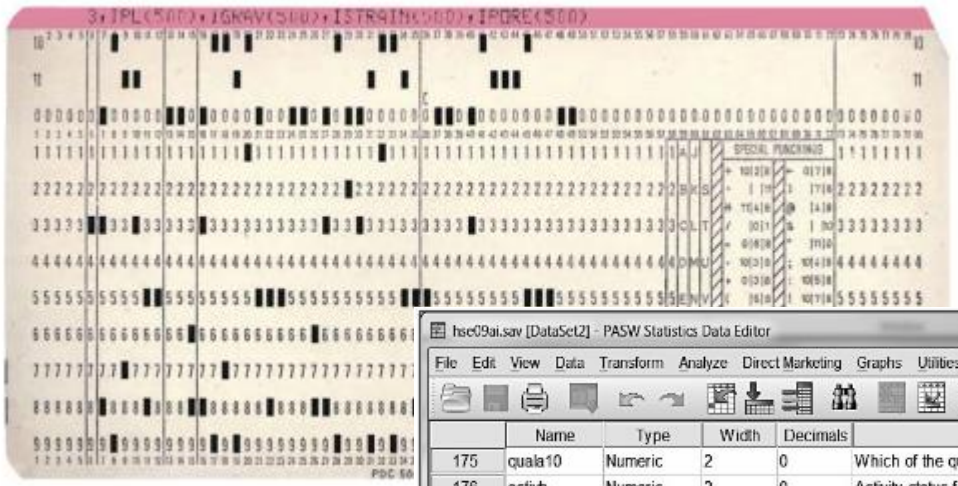
Data should be **user-friendly**, **shareable** and **with long-lasting usability**.

-> 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?



hse09ai.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	wklook4	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	snemplee	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

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

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

Source: UK Data Service

7 EU VET - Study on vocational education in seven European countries

The 7EU - VET project – Detailed Methodological Approach to Understanding the VET Education - is **a research study on vocational education and training** which builds on theoretical backgrounds and **secondary analyses** of the existing documentation as well as on national and EU data in order to conduct **quantitative and qualitative studies** and derive empirical results. The project is built upon one of the goals of the Lisbon strategy, which is the promotion and the quality of vocational education and training.

Manuals

- EUVET 12
  - Coding of Master questionnaire

- EUVET 12 (Manual for cleaning and entering data)
  - general instructions
  - defining missing variables
  - issues with specific question
  - entering data
  - quality control
  - cleaning the data
  - checking for errors.

European Social Survey

About ESS | Methodology | Data and Documentation | Resources

search

Rotating modules for ESS Round 8 announced

The ESS ERIC Scientific Advisory Board has selected the two successful modules for ESS Round 8.
More...

The 3rd International ESS Conference • 13-15th July 2016 • Lausanne, Switzerland MORE >>

Methodological Research

The European Social Survey runs a programme of research to support and enhance the methodology that underpins the high standards it pursues in every aspect of survey design, data collection and archiving.

Data and Documentation

Data and documentation can be accessed by round (year), by theme or by country. Data are available for download and online analysis.

ESS Resources

The ESS provides a series of outreach resources designed to increase the use of its data, including ESS Bibliography, Findings, Training Courses and eLearning resources.

ESS6 - integrated file
Edition 2.1

Integrated files and documents

ESS6 - integrated file, edition 2.1

ESS6 - data from Interviewer's questionnaire, edition 2.1

ESS6 - test variables from Supplementary questionnaire, edition 1.0

ESS6 - parents' occupation edition 2.1, all country files and integrated isco file

ESS6 - data from Contact forms, edition 2.0

ESS6 - interview time data, edition 1.1

ESS6 - data from Media claims, edition 4.0

Guide to weighting of ESS data

Useful hints: Combining data files, Renaming variables, Other data formats.

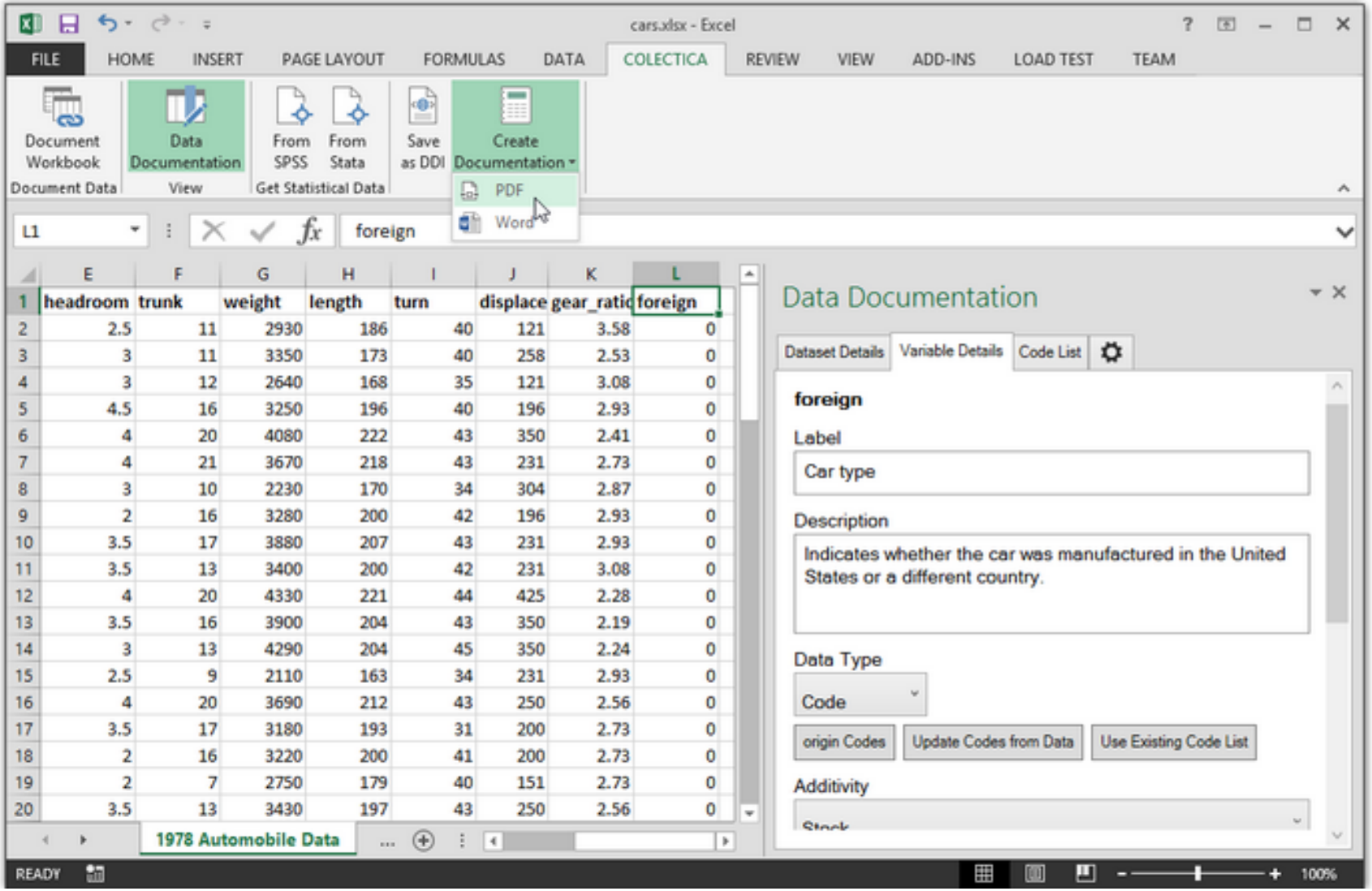
Fieldwork Summary and Deviations

European Social Survey – Data Protocol

B2. Parents' occupation .....
B3. Raw data .....
B4. Specifications and deposit of Sample design data files (SDDF) .....
B5. Common identifiers in data files .....
B6. Checking data.....
B7. Anonymised data, files 1-4 .....
B8. Deposit of indirectly identifiable data, files 5 and 6.....
C. Principles of variable definitions .....
C1. Missing values .....
C2. Multiple responses and missing values .....
C3. Formats.....
C4. Variable names.....
C5. Variable labels and categories .....
D. Standards and classifications .....
D1. International standards.....
D1.1 ISO 3166-1, Country.....
D1.2 ISO 639-2, Language .....

http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_data_protocol_e01_4.pdf

Your Dataset Deserves More than a First Row Header



	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

Colectica for Excel

Document Variables and Datasets

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

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.

Publish Documentation

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

Import Stata to Excel

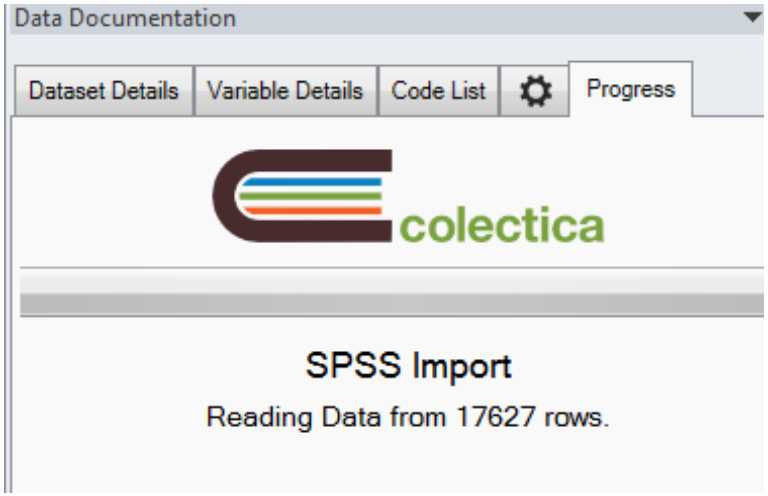
Colectica for Excel Professional allows direct importing and documenting of Stata data files, with a file extension .dta. The variable names, labels and code lists in the Stata file will also be imported and added to the stored documentation automatically.

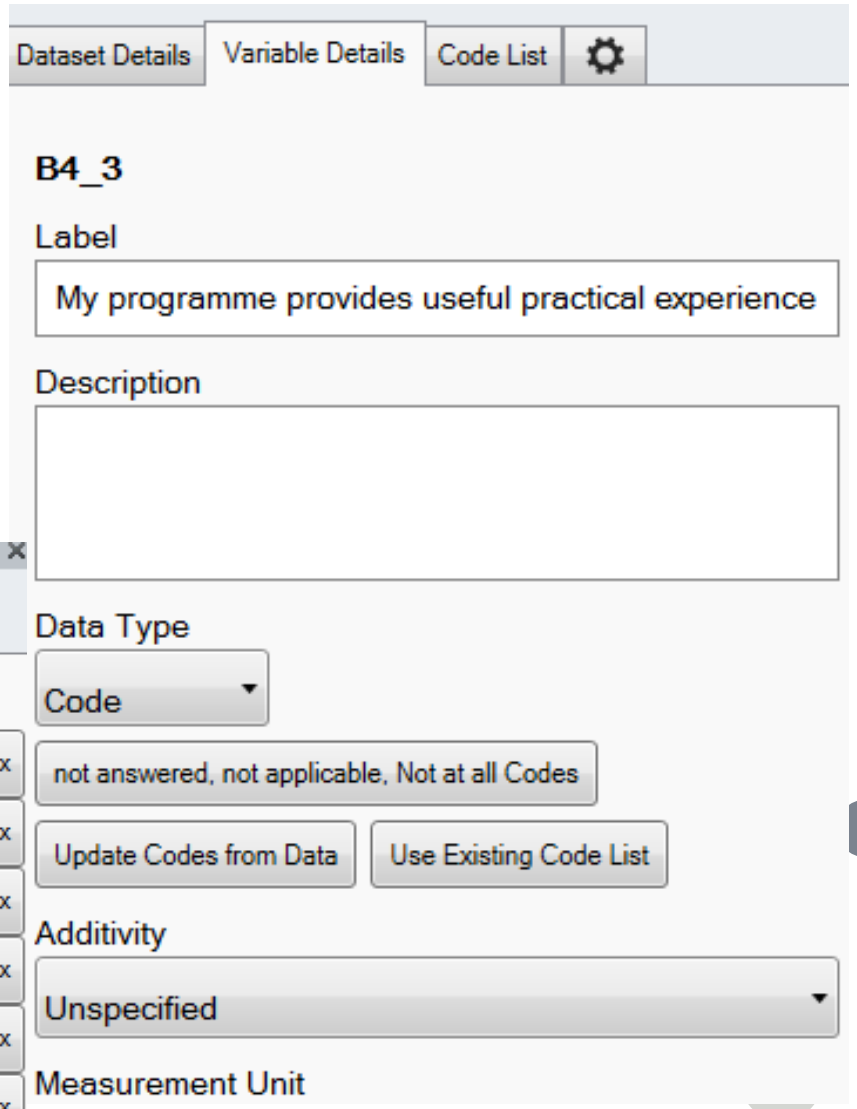
Import SPSS to Excel

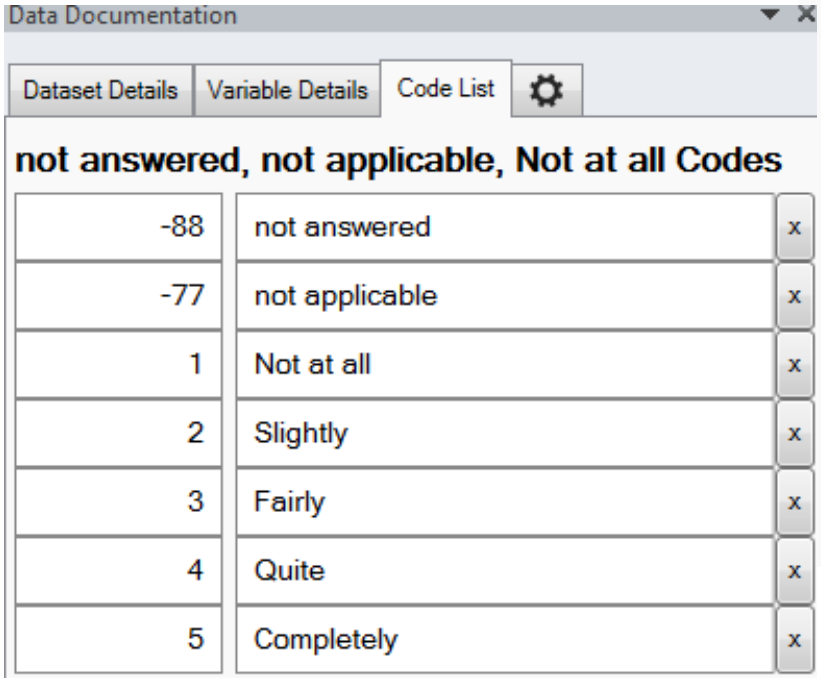
Colectica for Excel Professional allows direct importing and documenting of SPSS data files, with a file extension .sav. The variable names, labels and code lists in the SPSS file will also be imported and added to the stored documentation automatically.

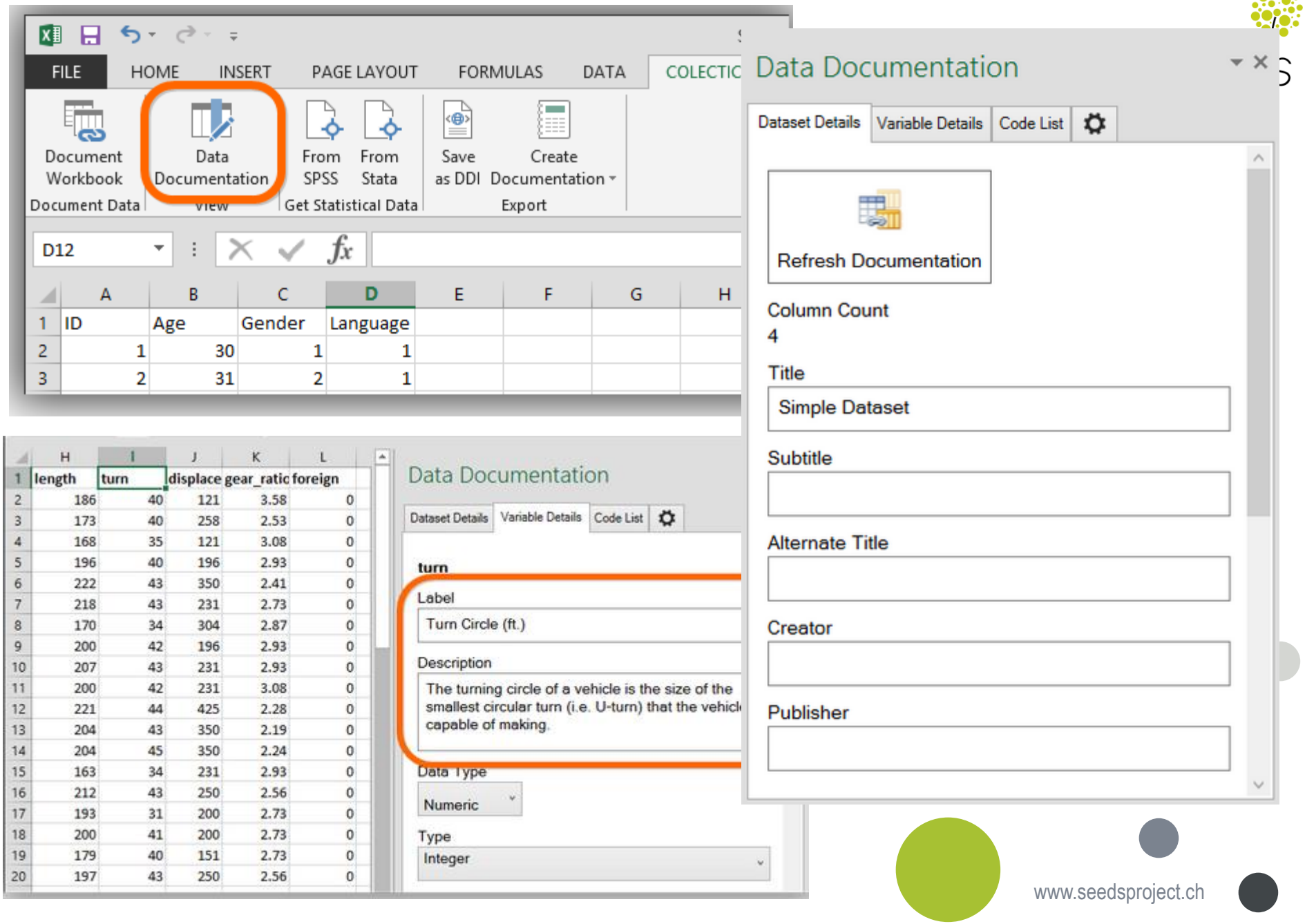
Create DDI-Lifecycle Metadata

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.


Data Documentation
Dataset Details
Variable Details
Code List
Progress
colectica
SPSS Import
Reading Data from 17627 rows.


Dataset Details
Variable Details
Code List
B4_3
Label
My programme provides useful practical experience
Description
Data Type
Code
not answered, not applicable, Not at all Codes
Update Codes from Data
Use Existing Code List
Additivity
Unspecified
Measurement Unit


Data Documentation
Dataset Details
Variable Details
Code List
not answered, not applicable, Not at all Codes
-88 not answered
-77 not applicable
1 Not at all
2 Slightly
3 Fairly
4 Quite
5 Completely


FILE
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
COLECTIC
Document Workbook
Data Documentation
From SPSS
From Stata
Save as DDI
Create Documentation
Document Data
View
Get Statistical Data
Export
D12
ID
Age
Gender
Language
Data Documentation
Dataset Details
Variable Details
Code List
Refresh Documentation
Column Count
4
Title
Simple Dataset
Subtitle
Alternate Title
Creator
Publisher
length
turn
displace
gear_ratio
foreign
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

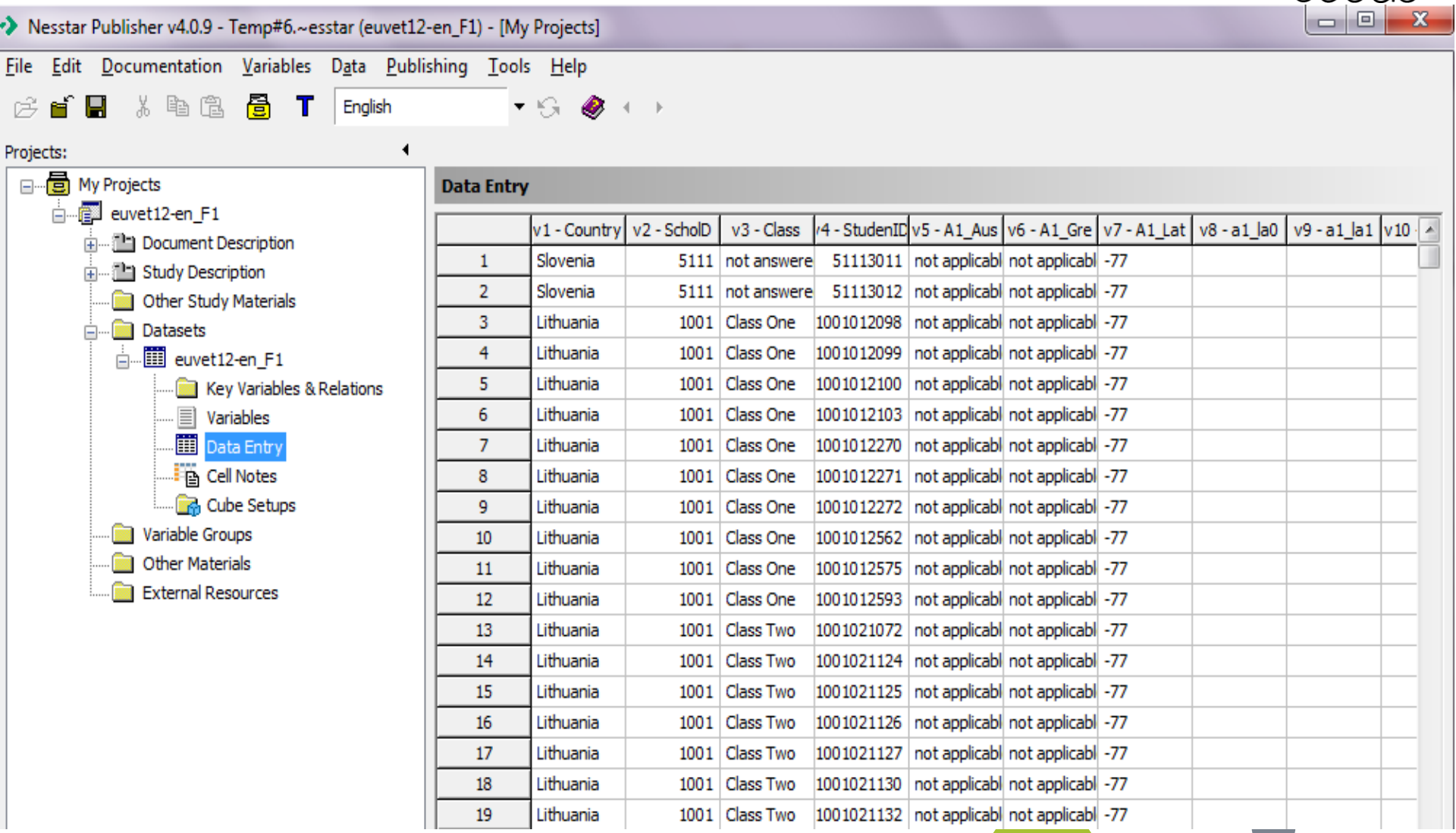
Nesstar Publisher – a sophisticated authoring environment that can publish data from a variety of sources (including SPSS, SAS, Excel etc.). The tool includes a specialised metadata editor, data and metadata validation routines and metadata templates that provide standardisation and control.

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.

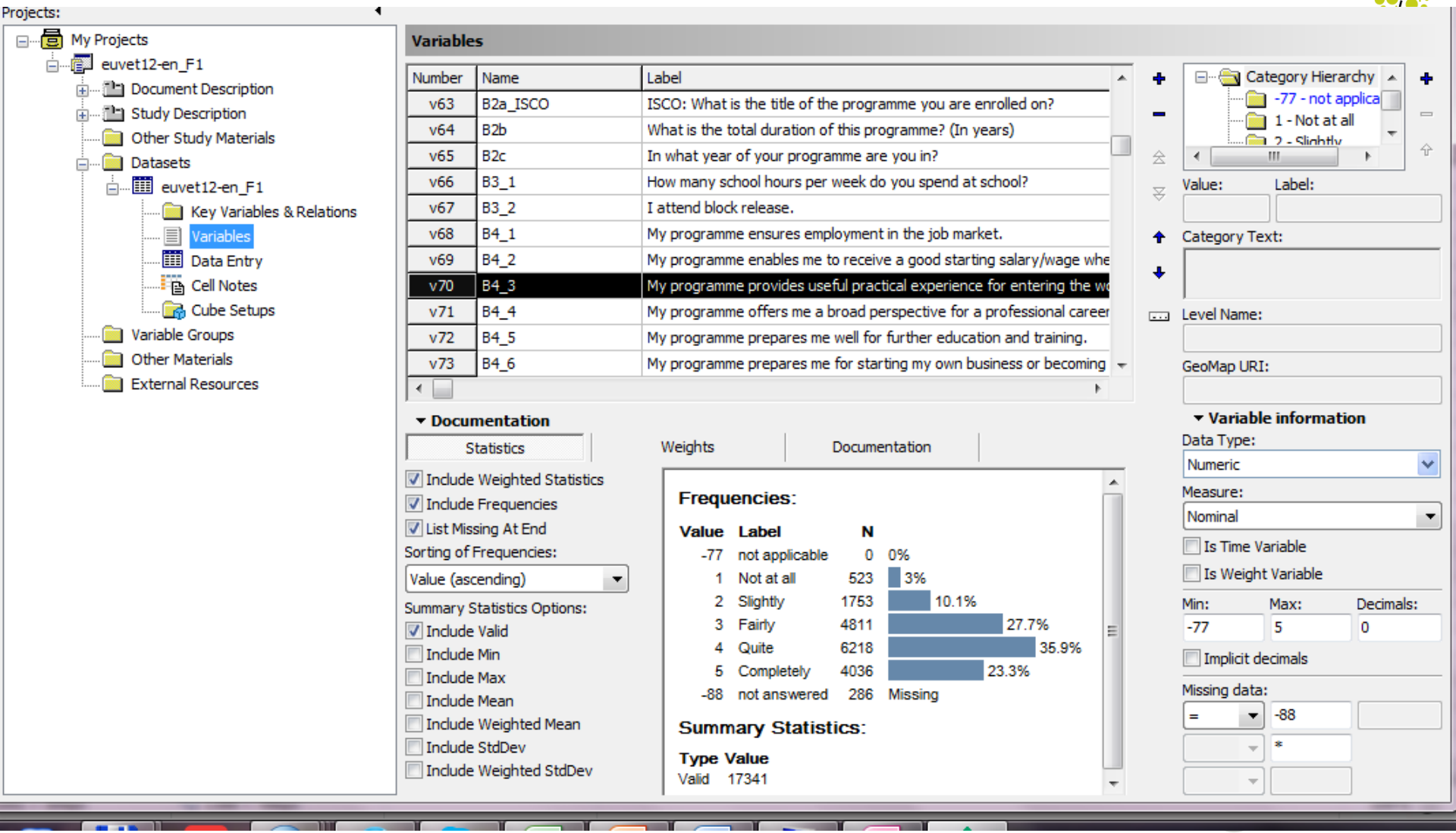
File formats currently supported:

- 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.

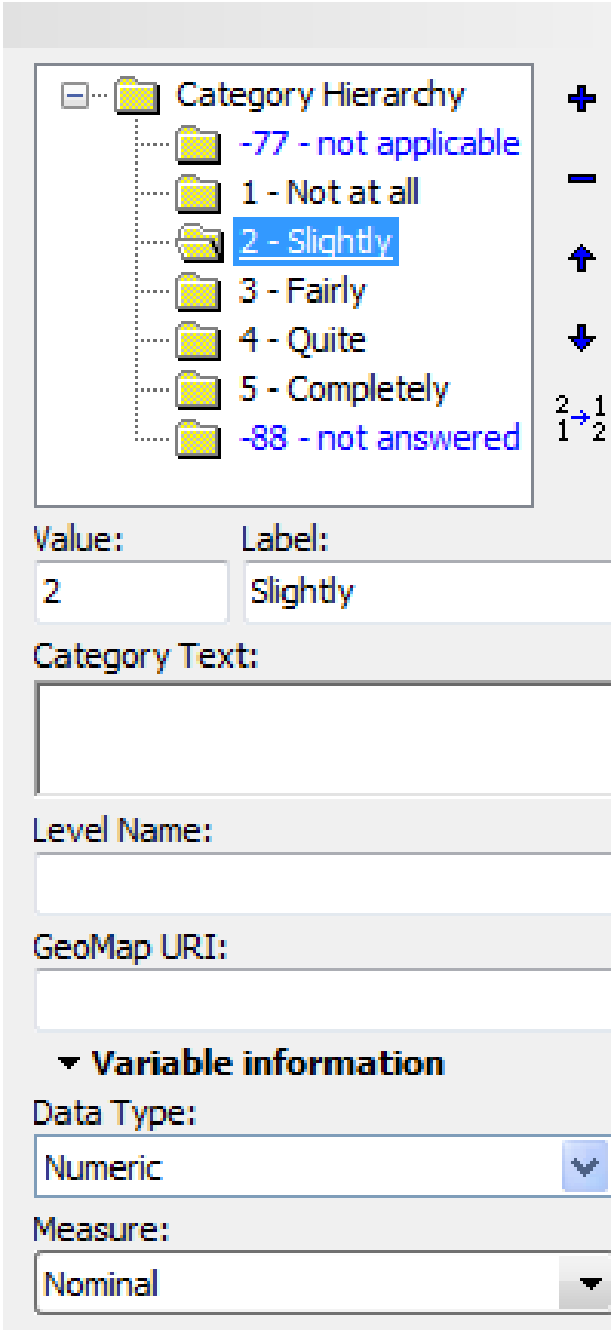


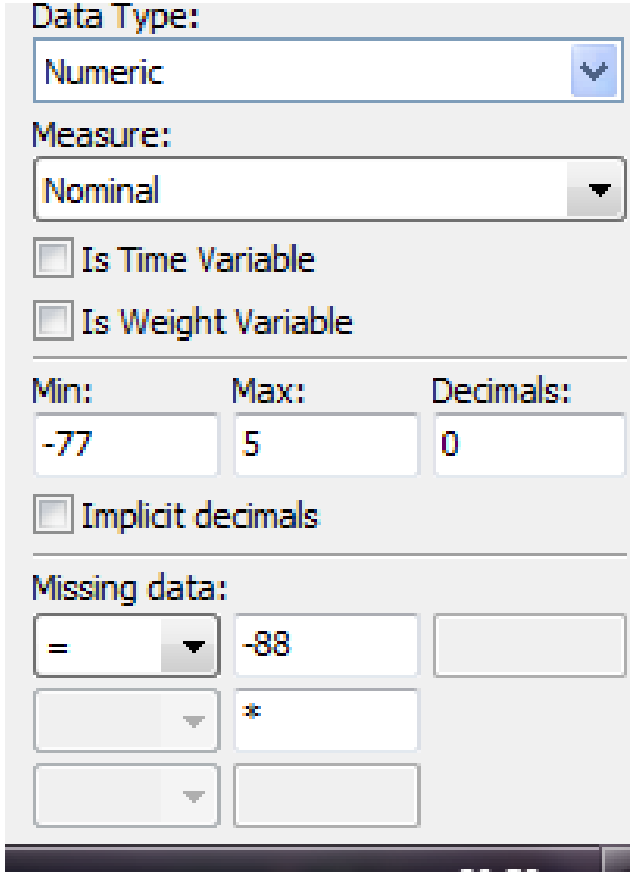
	v1 - Country	v2 - SchoID	v3 - Class	v4 - StudenID	v5 - A1_Aus	v6 - A1_Gre	v7 - A1_Lat	v8 - a1_la0	v9 - a1_la1	v10
1	Slovenia	5111	not answere	51113011	not applicabl	not applicabl	-77			
2	Slovenia	5111	not answere	51113012	not applicabl	not applicabl	-77			
3	Lithuania	1001	Class One	1001012098	not applicabl	not applicabl	-77			
4	Lithuania	1001	Class One	1001012099	not applicabl	not applicabl	-77			
5	Lithuania	1001	Class One	1001012100	not applicabl	not applicabl	-77			
6	Lithuania	1001	Class One	1001012103	not applicabl	not applicabl	-77			
7	Lithuania	1001	Class One	1001012270	not applicabl	not applicabl	-77			
8	Lithuania	1001	Class One	1001012271	not applicabl	not applicabl	-77			
9	Lithuania	1001	Class One	1001012272	not applicabl	not applicabl	-77			
10	Lithuania	1001	Class One	1001012562	not applicabl	not applicabl	-77			
11	Lithuania	1001	Class One	1001012575	not applicabl	not applicabl	-77			
12	Lithuania	1001	Class One	1001012593	not applicabl	not applicabl	-77			
13	Lithuania	1001	Class Two	1001021072	not applicabl	not applicabl	-77			
14	Lithuania	1001	Class Two	1001021124	not applicabl	not applicabl	-77			
15	Lithuania	1001	Class Two	1001021125	not applicabl	not applicabl	-77			
16	Lithuania	1001	Class Two	1001021126	not applicabl	not applicabl	-77			
17	Lithuania	1001	Class Two	1001021127	not applicabl	not applicabl	-77			
18	Lithuania	1001	Class Two	1001021130	not applicabl	not applicabl	-77			
19	Lithuania	1001	Class Two	1001021132	not applicabl	not applicabl	-77			


Projects:
My Projects
euvet12-en_F1
Document Description
Study Description
Other Study Materials
Datasets
euvet12-en_F1
Key Variables & Relations
Variables
Data Entry
Cell Notes
Cube Setups
Variable Groups
Other Materials
External Resources
Variables
Number Name Label
v63 B2a_ISCO ISCO: What is the title of the programme you are enrolled on?
v64 B2b What is the total duration of this programme? (In years)
v65 B2c In what year of your programme are you in?
v66 B3_1 How many school hours per week do you spend at school?
v67 B3_2 I attend block release.
v68 B4_1 My programme ensures employment in the job market.
v69 B4_2 My programme enables me to receive a good starting salary/wage whe
v70 B4_3 My programme provides useful practical experience for entering the wo
v71 B4_4 My programme offers me a broad perspective for a professional career
v72 B4_5 My programme prepares me well for further education and training.
v73 B4_6 My programme prepares me for starting my own business or becoming
Category Hierarchy
-77 - not applica
1 - Not at all
2 - Slightly
Value:
Label:
Category Text:
Level Name:
GeoMap URI:
Documentation
Statistics
Weights
Documentation
Include Weighted Statistics
Include Frequencies
List Missing At End
Sorting of Frequencies:
Value (ascending)
Summary Statistics Options:
Include Valid
Include Min
Include Max
Include Mean
Include Weighted Mean
Include StdDev
Include Weighted StdDev
Frequencies:
Value Label N
-77 not applicable 0 0%
1 Not at all 523 3%
2 Slightly 1753 10.1%
3 Fairly 4811 27.7%
4 Quite 6218 35.9%
5 Completely 4036 23.3%
-88 not answered 286 Missing
Summary Statistics:
Type Value
Valid 17341
Variable information
Data Type:
Numeric
Measure:
Nominal
Is Time Variable
Is Weight Variable
Min:
Max:
Decimals:
-77
5
0
Implicit decimals
Missing data:
=
-88
*

Variables

Number	Name	Label	Width	StartCol	EndCol	Record	Dec
v63	B2a_ISCO	ISCO: What is the title of the programme you are enrolled on?	8	*	*	*	
v64	B2b	What is the total duration of this programme? (In years)	11	*	*	*	
v65	B2c	In what year of your programme are you in?	11	*	*	*	
v66	B3_1	How many school hours per week do you spend at school?	11	*	*	*	
v67	B3_2	I attend block release.	11	*	*	*	
v68	B4_1	My programme ensures employment in the job market.	11	*	*	*	
v69	B4_2	My programme enables me to receive a good starting salary/wage when su	11	*	*	*	
v70	B4_3	My programme provides useful practical experience for entering the workf	11	*	*	*	
v71	B4_4	My programme offers me a broad perspective for a professional career.	11	*	*	*	
v72	B4_5	My programme prepares me well for further education and training.	11	*	*	*	
v73	B4_6	My programme prepares me for starting my own business or becoming self	11	*	*	*	
v74	B4_7	My programme is recognised within society as having a good reputation.	11	*	*	*	
v75	B4_8	My programme prepares me for a job that is important for society.	11	*	*	*	
v76	B5_1	Satisfaction: Most of my classes are interesting.	11	*	*	*	
v77	B5_2	Satisfaction: Most of my teachers are usually well prepared when teaching	11	*	*	*	


Category Hierarchy
-77 - not applicable
1 - Not at all
2 - Slightly
3 - Fairly
4 - Quite
5 - Completely
-88 - not answered
Value:
2
Label:
Slightly
Category Text:
Level Name:
GeoMap URI:
Variable information
Data Type:
Numeric
Measure:
Nominal


Data Type:
Numeric
Measure:
Nominal
Is Time Variable
Is Weight Variable
Min:
Max:
Decimals:
-77
5
0
Implicit decimals
Missing data:
=
-88
*

▼ **Documentation**

Statistics | Weights | Documentation

- [x] Include Weighted Statistics
- [x] Include Frequencies
- [x] List Missing At End

Sorting of Frequencies:

Value (ascending)

Summary Statistics Options:

- [x] Include Valid
- [ ] Include Min
- [ ] Include Max
- [x] Include Mean
- [ ] Include Weighted Mean
- [ ] Include StdDev
- [ ] Include Weighted StdDev

Frequencies:

Value	Label	N	
-77	not applicable	0	0%
1	Not at all	523	3%
2	Slightly	1753	10.1%
3	Fairly	4811	27.7%
4	Quite	6218	35.9%
5	Completely	4036	23.3%
-88	not answered	286	Missing

Summary Statistics:

Type	Value
Valid	17341
Mean	3.663

Variable Groups

- Variable Groups
  - A_F1: Preliminary program and the tra
  - B_F1: Current programme
  - C_F1: Acquired Knowledge
  - D_F1: About yourself and your career
  - E_F1: Acquired skills and abilities
  - F_F1: Information and communication
  - G_F1: You and your family
  - Other_F1
  - A_F2: Preliminary program and the tra
  - B_F2: Current programme
  - C_F2: Acquired Knowledge
  - D_F2: About yourself and your career
  - E_F2: Acquired skills and abilities
  - F_F2: Information and communication
  - G_F2: You and your family
  - Other_F2

Description | Variables

Dataset	Number	Name	Label
euvet12_f1	v148	D1_1	Goals: Obtaining solid occupational proficiencies
euvet12_f1	v149	D1_2	Goals: Receiving a high income
euvet12_f1	v150	D1_3	Goals: Gaining job security
euvet12_f1	v151	D1_4	Goals: Having responsibility at work
euvet12_f1	v152	D1_5	Goals: Having opportunities to learn new things at work
euvet12_f1	v153	D1_6	Goals: Undertaking interesting tasks in the workplace
euvet12_f1	v154	D1_7	Goals: Having a job that makes me happy
euvet12_f1	v155	D1_8	Goals: Having a good relationship with colleagues
euvet12_f1	v156	D1_9	Goals: Advancing to a high level of status in society
euvet12_f1	v157	D1_10	Goals: having enough spare-time to do other things in life
euvet12_f1	v158	D1_11	Goals: Making and maintaining relationships with others (e.g. family and friends)
euvet12_f1	v159	D2_1	Role: A woman should be prepared to cut down on her paid work for the sake of her family (not asked
euvet12_f1	v160	D2_2	Role: When jobs are scarce, men should have more right to a job than women (not asked in UK)
euvet12_f1	v161	D2_3	Role: There should be many more women in political and public leadership roles (not asked in UK)
euvet12_f1	v162	D2_4	Role: Men should take as much responsibility as women for the home and children (not asked in UK)
euvet12_f1	v163	D2_5	Role: A man who stays at home and runs the household is not a real man (not asked in UK)
euvet12_f1	v164	D2_6	Role: When there are children in the home, parents should stay together even if they do not get along
euvet12_f1	v165	D2_7	Role: A persons family ought to be his or her main priority in life (not asked in UK)
euvet12_f1	v166	D3	Do you think men and women have the same opportunities to get a job in your aspired occupation? (no
euvet12_f1	v167	D4	What kind of job do you expect to have when you are about 30 years old?
euvet12_f1	v168	d40	What kind of job do you expect to have when you are about 30 years old?
euvet12_f1	v169	d41	What kind of job do you expect to have when you are about 30 years old?
euvet12_f1	v170	d42	What kind of job do you expect to have when you are about 30 years old?
euvet12_f1	v171	D4_ISCO	ISCO: What kind of job do you expect to have when you are about 30 years old?
euvet12_f1	v172	D5_1	Sector: Industry (e.g. producing industry, steel, motor, oil)
euvet12_f1	v173	D5_2	Sector: Services (e.g. nursing, policing, hairdressing)

v31	A3_3a_2	Previous grade: first foreign language_comparable	8	*	*	1
v32	A4_1	Choice: The programme offered good job prospects	11	*	*	1
v33	A4_2	Choice: My previous examination grades prevented me being able to enrol	11	*	*	1

Documentation

Statistics | Weights | Documentation

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

Literal Question

A4. How important were the following aspects to you when you were choosing your current programme? (Please tick only one box in each row) The programme offered good job prospects

Cleaning a data file

One think is to correct label a data file, yet another think is to check the data.

See if data file needs to be anonymized. Level of anonimization is defined by user conditions of a file it self.

You might decide to prepare files for different communities:

- ScUF – Secure Use File
- SUF – Scientific Use File
- PUF – Public Use File
- CUF – Campus Files (usually specific file prepared for training)

Basic anonymisation of archives distributed microdata

- deleting variables

Direct identifiers (telephone numbers, addresses etc.) are removed.

- recoding indirect identifiers

But still allowing serious researchers to receive datasets with indirect identifiers non-recoded). Recoding includes removing values and bracketing – combining the categories of a variable (aggregation).

Anonymisation of Eurostat files
(the case of Eurostat Labor Force Survey)

- deleting variables: indirect identifiers and unneeded variables are removed (municipality, wave nr. etc.)
- bracketing: age, marital status, education, years of residence, age of establishment of residence, duration of search of employment, professional status, country & nationality
- classification: income numbers are not given, respondents are divided into classes based on their income
- aggregation: economic activity and occupation values are aggregated at 1-digit level
- top-coding: restricting the upper range of a variable (no. of hours worked)

You can find more in

- UKDA – Create & Manage Data

http://www.data-archive.ac.uk/create-manage

- ICSPR – Guide to Social Science Data Preparation and Archiving

http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html

IHSN – Data archiving and dissemination

http://www.ihsn.org/home/archiving

MANTRA – Research Data Management Training

http://datalib.edina.ac.uk/mantra/

www.seedsproject.ch