CESSDA

Chapter 2. Organise & Document

Jindřich Krejčí <jindrich.krejci@soc.cas.cz>
Johana Chylíková <johana.chylikova@soc.cas.cz>

RDM Expert Tour Guide

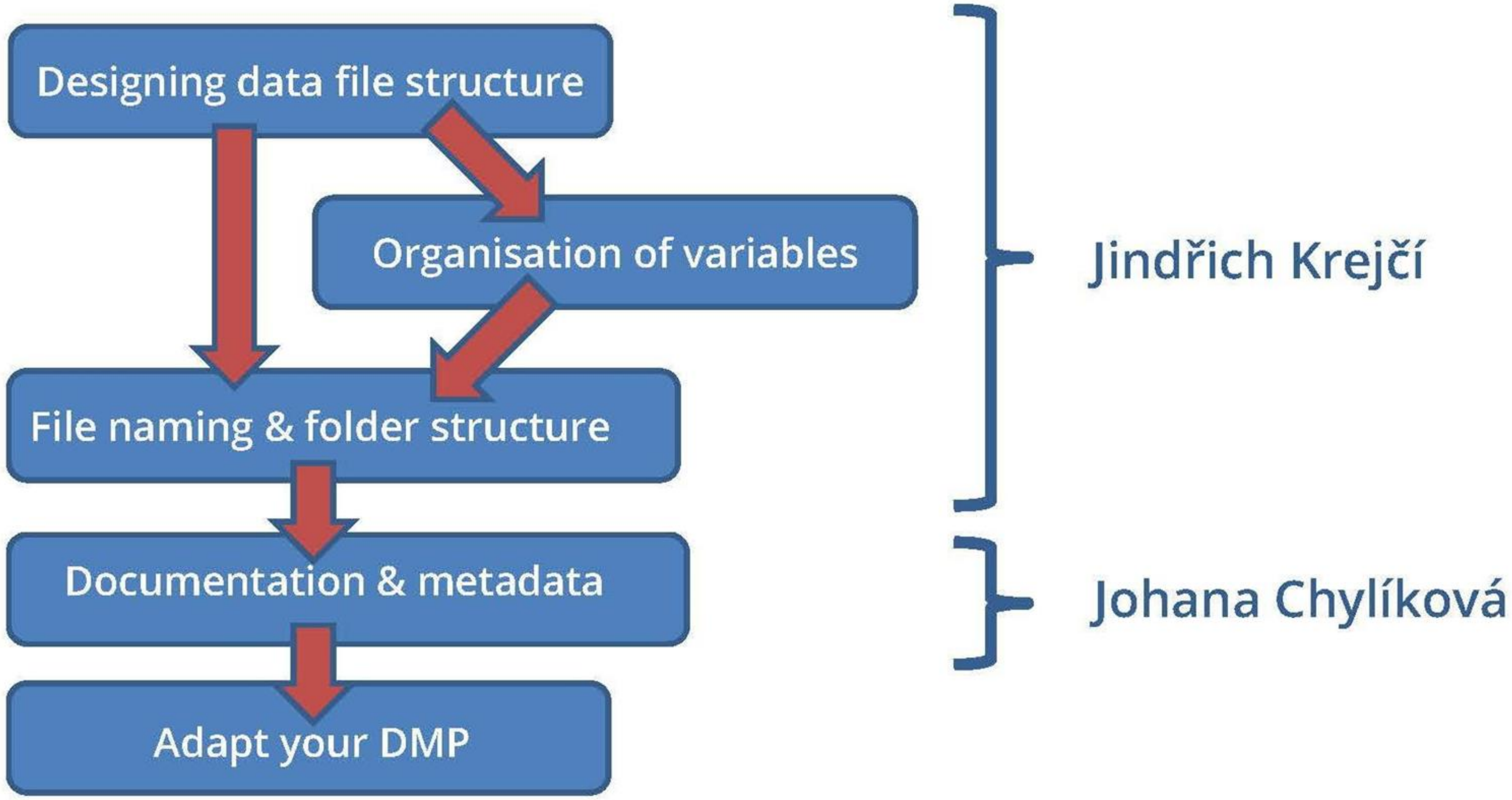
Train the Trainers event, 12-13 April 2018

cessda.eu @CESSDA_Data

Cite as: Krejci, J., J. Chylikova (2020). The CESSDA Data Management Expert Guide. Chapter 2. Organise & Document [presentation]. Bergen: CESSDA ERIC.

Chapter 2

- Authors: Jindřich Krejčí (ČSDA), Johana Chylíková (ČSDA), Katja Fält (FSD)



Designing data file structure

Organisation of variables

File naming & folder structure

Documentation & metadata

Adapt your DMP

Jindřich Krejčí

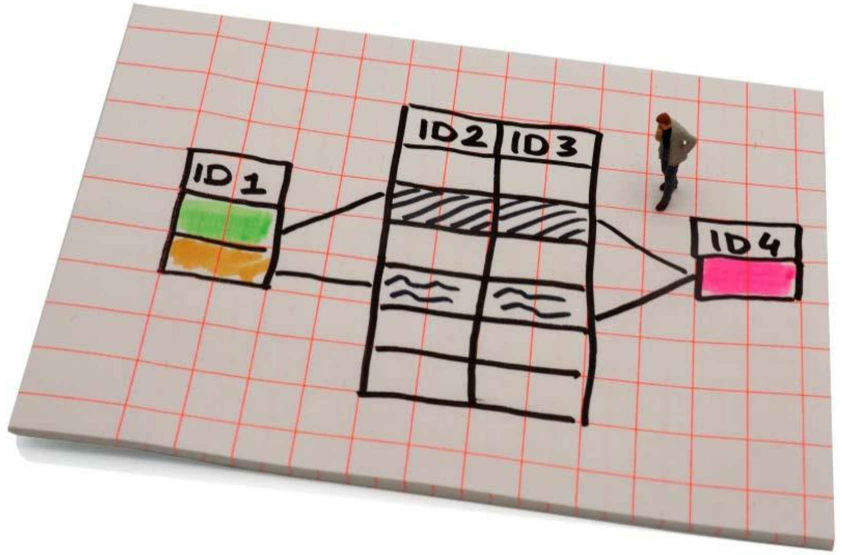
Johana Chylíková

Designing a data file structure

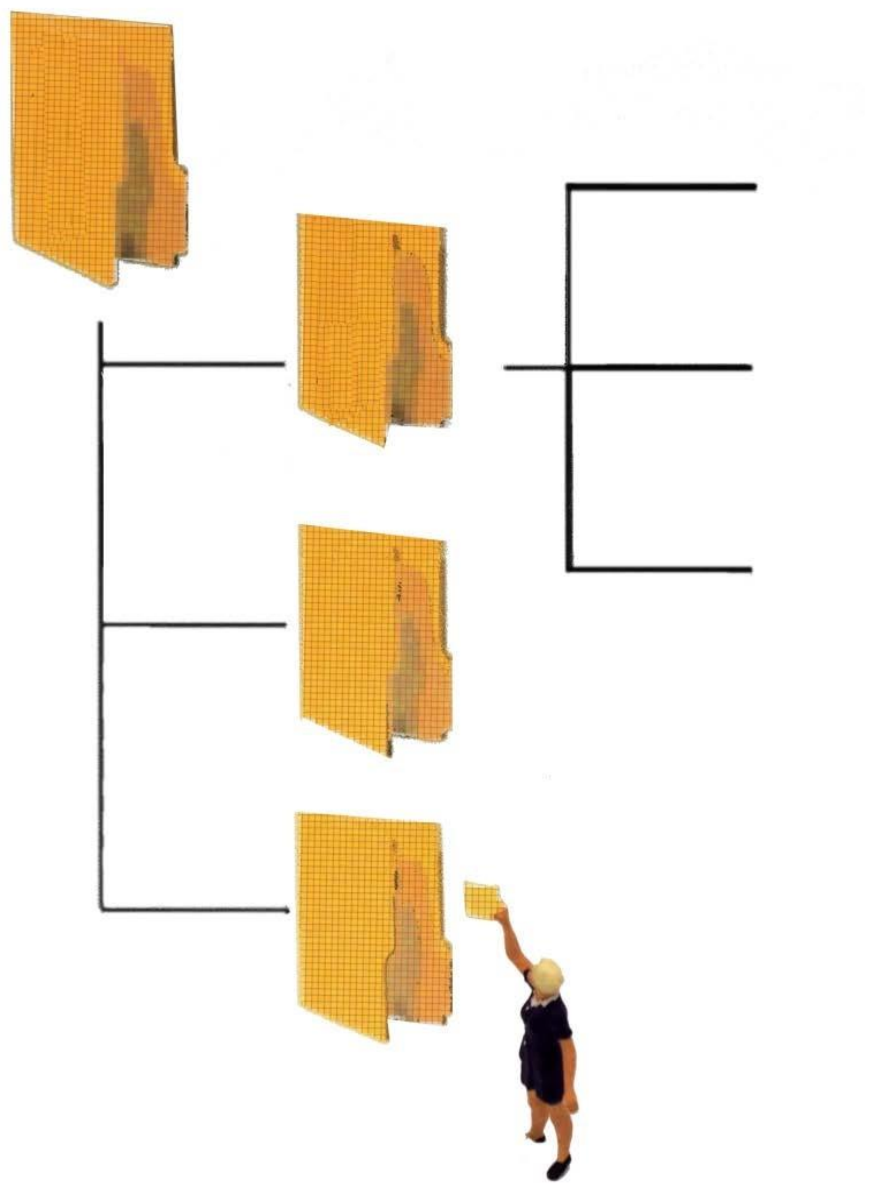
Data file structure has a huge impact on the possible ways your files can be processed and analysed. Once your structure has been filled with data, any changes to it are usually laborious and time-consuming.

Consider following:

- Units of analysis / analytical objectives / methods of analysis
- Relations: (a) between different content items; (b) to sources of your data; (c) to any other relevant external information
- Possible connections to other existing or future data files
- Strategies for version control
- Technical limitations (e.g. the size, software limitations (No. of variables & cases…)
- Software you are going to use (also flexibility for secondary analysis)



Qualitative data

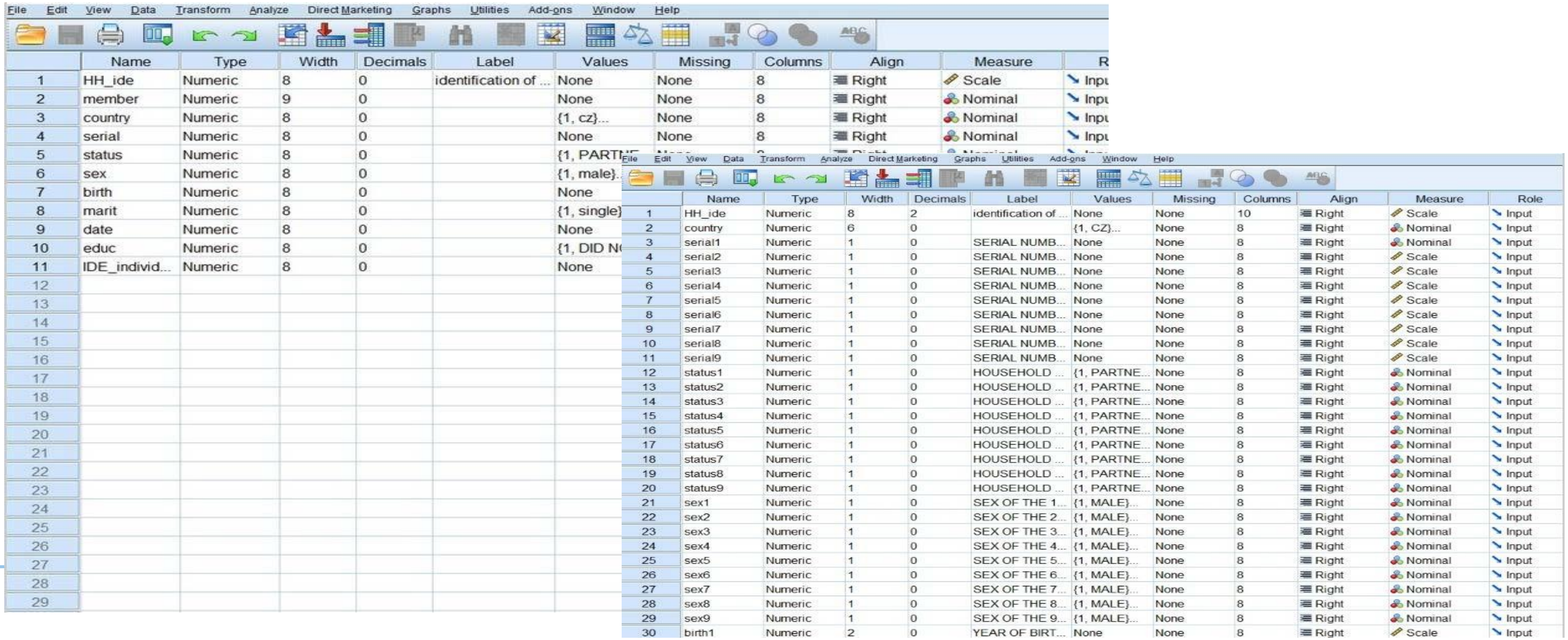


- Thinking of ways to categorise data (see 'Qualitative coding')
- Developing a file naming strategy (see 'File naming and folder structure')
- Designing a comprehensive folder structure (see 'File naming and folder structure')

Qualitative data

- Flat (rectangular) data files
- Hierarchical files
- Relational database

Hierarchical data in a flat file structure - example of household survey: either data on individuals or data on households



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	R
1	HH_ide	Numeric	8	0	identification of ...	None	None	8	Right	Scale	Inpu
2	member	Numeric	9	0		None	None	8	Right	Nominal	Inpu
3	country	Numeric	8	0		{1, cz}...	None	8	Right	Nominal	Inpu
4	serial	Numeric	8	0		None	None	8	Right	Nominal	Inpu
5	status	Numeric	8	0		{1, PART					
6	sex	Numeric	8	0		{1, male}.					
7	birth	Numeric	8	0		None					
8	marit	Numeric	8	0		{1, single}					
9	date	Numeric	8	0		None					
10	educ	Numeric	8	0		{1, DID N					
11	IDE_individ...	Numeric	8	0		None					
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	HH_ide	Numeric	8	2	identification of ...	None	None	10	Right	Scale	Input
2	country	Numeric	6	0		{1, CZ}...	None	8	Right	Nominal	Input
3	serial1	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Nominal	Input
4	serial2	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
5	serial3	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
6	serial4	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
7	serial5	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
8	serial6	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
9	serial7	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
10	serial8	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
11	serial9	Numeric	1	0	SERIAL NUMB...	None	None	8	Right	Scale	Input
12	status1	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
13	status2	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
14	status3	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
15	status4	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
16	status5	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
17	status6	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
18	status7	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
19	status8	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
20	status9	Numeric	1	0	HOUSEHOLD ...	{1, PARTNE...	None	8	Right	Nominal	Input
21	sex1	Numeric	1	0	SEX OF THE 1...	{1, MALE}...	None	8	Right	Nominal	Input
22	sex2	Numeric	1	0	SEX OF THE 2...	{1, MALE}...	None	8	Right	Nominal	Input
23	sex3	Numeric	1	0	SEX OF THE 3...	{1, MALE}...	None	8	Right	Nominal	Input
24	sex4	Numeric	1	0	SEX OF THE 4...	{1, MALE}...	None	8	Right	Nominal	Input
25	sex5	Numeric	1	0	SEX OF THE 5...	{1, MALE}...	None	8	Right	Nominal	Input
26	sex6	Numeric	1	0	SEX OF THE 6...	{1, MALE}...	None	8	Right	Nominal	Input
27	sex7	Numeric	1	0	SEX OF THE 7...	{1, MALE}...	None	8	Right	Nominal	Input
28	sex8	Numeric	1	0	SEX OF THE 8...	{1, MALE}...	None	8	Right	Nominal	Input
29	sex9	Numeric	1	0	SEX OF THE 9...	{1, MALE}...	None	8	Right	Nominal	Input
30	birth1	Numeric	2	0	YEAR OF BIRT...	None	None	8	Right	Scale	Input

Example

Survey of Health, Ageing and Retirement in Europe (SHARE): http://www.share-project.org/

- International survey; data come from different countries
- Panel survey repeating interviews with the same sample of households
- Questionnaires include both, repeated and new questions
- Refreshment: new households are added at each wave; two types of questionnaires: baseline / longitudinal quex
- Different components of the survey with different sources of information; different data collection modes
- Different types of respondents answer different parts of an interview for household: (1) family respondent, (2) financial respondent, (3) household respondent; (4) proxy respondents
- Structured by topics

SHARE

Survey of Health, Ageing and Retirement in Europe

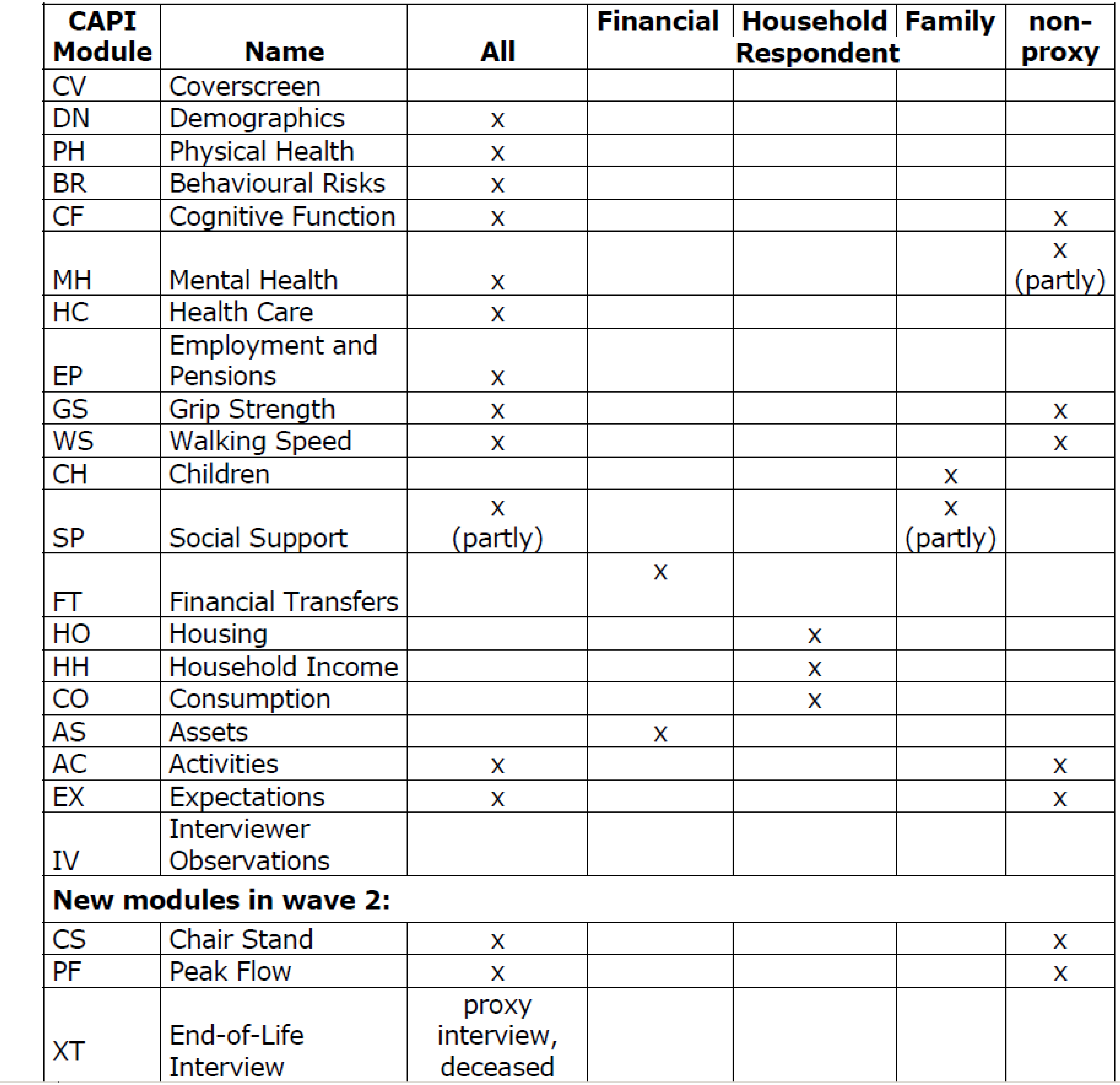
Example

SHARE - up to 30 different data modules per wave

Unique identifiers:

- **Merging data on individual level** "CC-hhhhhh- rr", "CC" = country code, "hhhhhh"= household identifier, "rr" = respondent identifier within each household (e.g. "AT-070759-01")
- **Merging data on household level** hhid`w', `w' indicates the respective wave. "CC-hhhhhh-S"(e.g. "AT-070759-A"), "CC" = country code, "hhhhhh"= household identifier, "S" identifies possible split household

Table: Who answers what in the CAPI questionnaire?



CAPI			Financial	Household	Family	non-
Module	Name	All		Respondent		proxy
CV	Coverscreen					
DN	Demographics	x				
PH	Physical Health	x				
BR	Behavioural Risks	x				
CF	Cognitive Function	x				x
MH	Mental Health	x				x (partly)
HC	Health Care	x				
EP	Employment and Pensions	x				
GS	Grip Strength	x				x
WS	Walking Speed	x				x
CH	Children				x	
SP	Social Support	x (partly)			x (partly)	
FT	Financial Transfers		x			
HO	Housing			x		
HH	Household Income			x		
CO	Consumption			x		
AS	Assets		x			
AC	Activities	x				x
EX	Expectations	x				x
IV	Interviewer Observations					
New modules in wave 2:						
CS	Chair Stand	x				x
PF	Peak Flow	x				x
XT	End-of-Life Interview	proxy interview, deceased				

Dive in deeper? Organisation of variables

Variable names and labels contribute into structuring of the data file and allow researchers to integrate part of the documentation into the data file.

Reflect following:

- Relations between variables
- Links to elements of the study and sources of the data
- Types of variables

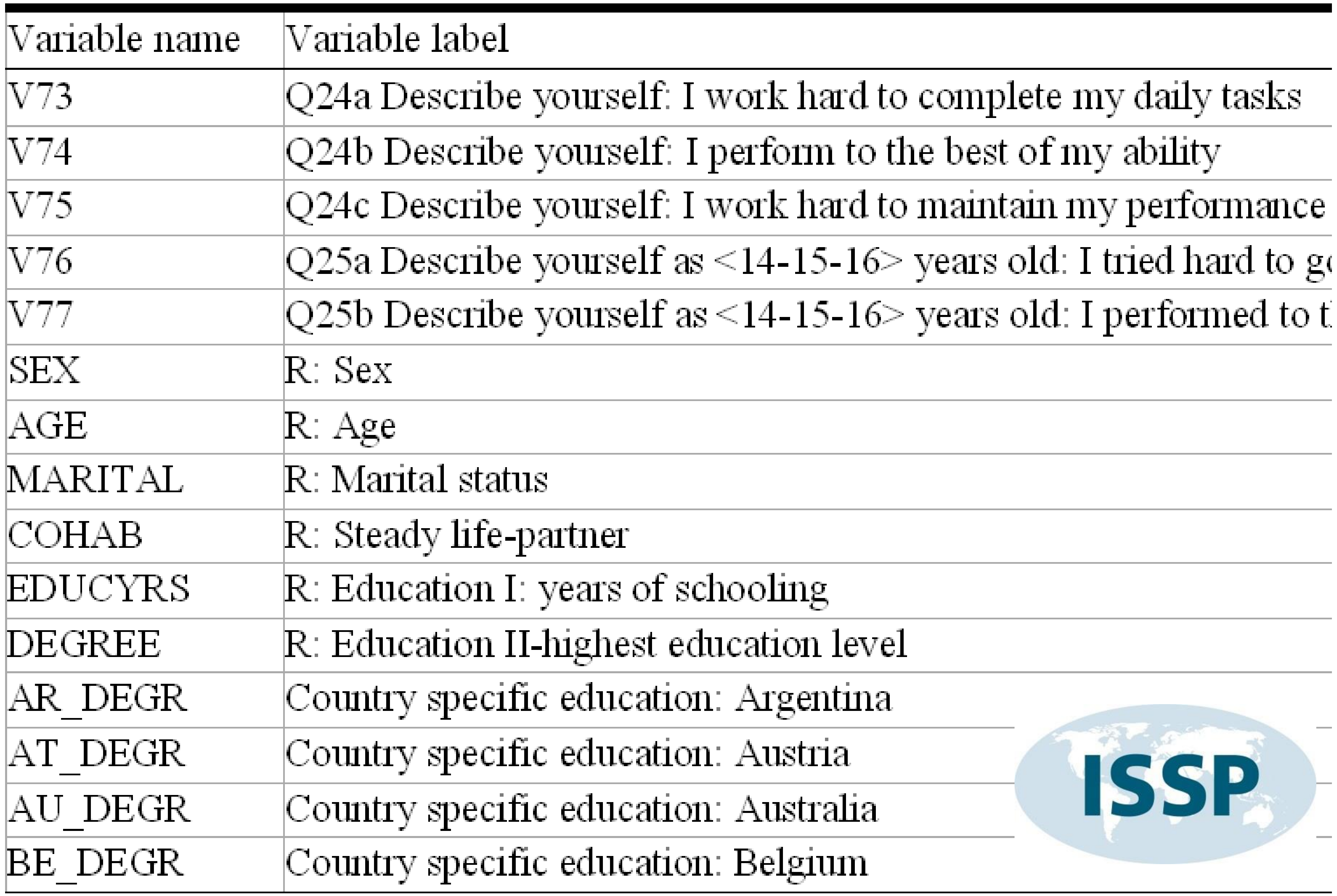
Basic rules of naming variables

- Do not start with a number
- Do not use special characters
- No spaces
- Be short (eight characters)
- Do not use national specific characters
- Make them meaningful

Three approaches to naming:

- Using numeric codes that reflect the variable’s position in a system (e.g. V001, V002, V003...);
- Using codes that refer to the research instrument (e.g. question number in a questionnaire: Q1a, Q1b, Q2, Q3a...);
- Using mnemonic names that refer to the content of variables (e.g. BIRTH for the year of birth, AGE for respondent’s age etc.). The word mnemonic means “memory aid”.

International Social Survey Programme (ISSP)



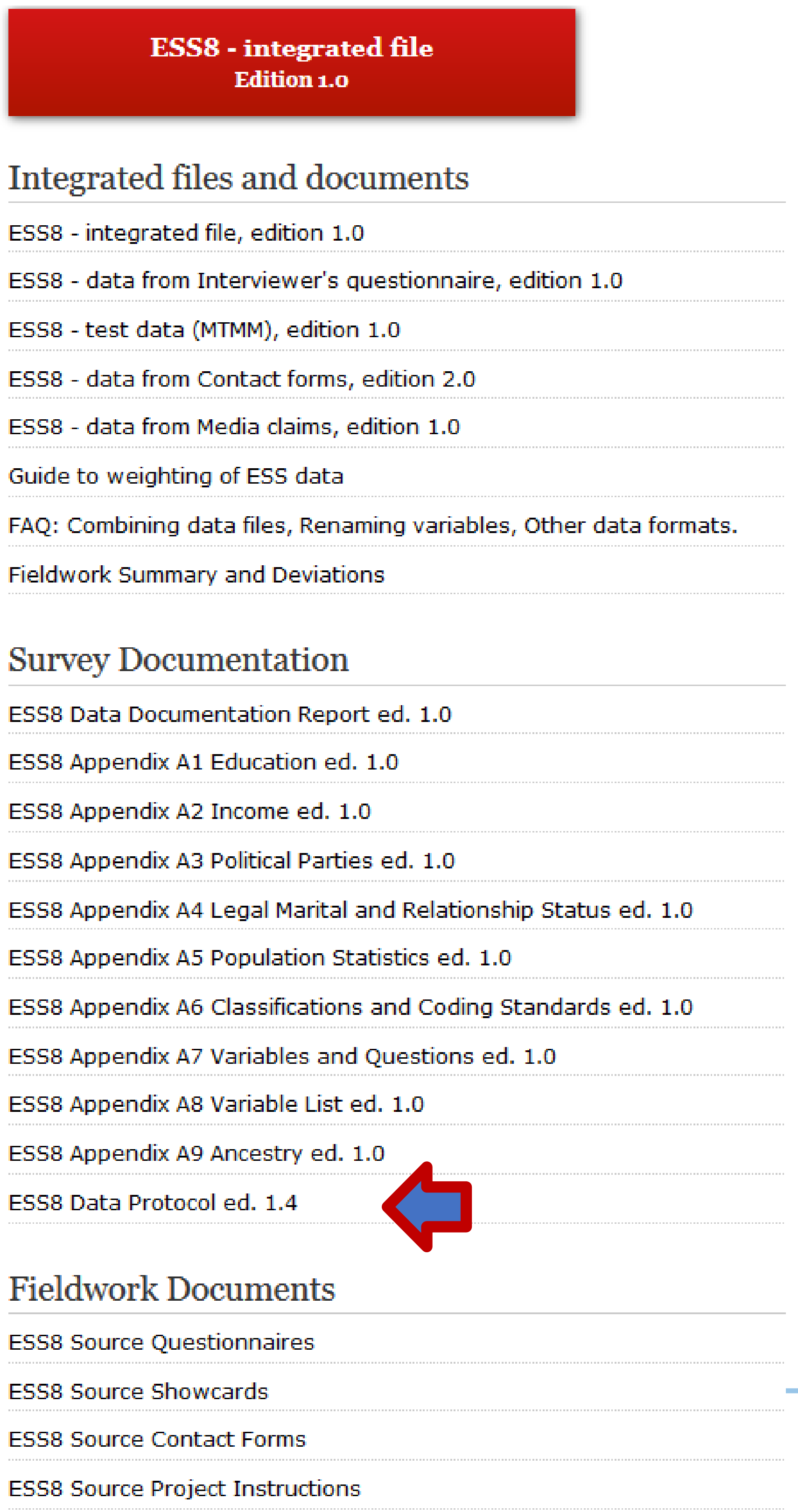
Variable name	Variable label
V73	Q24a Describe yourself: I work hard to complete my daily tasks
V74	Q24b Describe yourself: I perform to the best of my ability
V75	Q24c Describe yourself: I work hard to maintain my performance
V76	Q25a Describe yourself as <14-15-16> years old: I tried hard to g
V77	Q25b Describe yourself as <14-15-16> years old: I performed to t
SEX	R: Sex
AGE	R: Age
MARITAL	R: Marital status
COHAB	R: Steady life-partner
EDUCYRS	R: Education I: years of schooling
DEGREE	R: Education II-highest education level
AR_DEGR	Country specific education: Argentina
AT_DEGR	Country specific education: Austria
AU_DEGR	Country specific education: Australia
BE_DEGR	Country specific education: Belgium

ISSP combine numeric codes that reflect the variable’s position and mnemonic approach while ESS uses only the mnemonic names.

Why? What is better in relation to different purposes?

European Social Survey (ESS)

http://www.europeansocialsurvey.org/



ESS8 - integrated file
Edition 1.0

Integrated files and documents

- ESS8 - integrated file, edition 1.0
- ESS8 - data from Interviewer's questionnaire, edition 1.0
- ESS8 - test data (MTMM), edition 1.0
- ESS8 - data from Contact forms, edition 2.0
- ESS8 - data from Media claims, edition 1.0
- Guide to weighting of ESS data
- FAQ: Combining data files, Renaming variables, Other data formats.
- Fieldwork Summary and Deviations

Survey Documentation

- ESS8 Data Documentation Report ed. 1.0
- ESS8 Appendix A1 Education ed. 1.0
- ESS8 Appendix A2 Income ed. 1.0
- ESS8 Appendix A3 Political Parties ed. 1.0
- ESS8 Appendix A4 Legal Marital and Relationship Status ed. 1.0
- ESS8 Appendix A5 Population Statistics ed. 1.0
- ESS8 Appendix A6 Classifications and Coding Standards ed. 1.0
- ESS8 Appendix A7 Variables and Questions ed. 1.0
- ESS8 Appendix A8 Variable List ed. 1.0
- ESS8 Appendix A9 Ancestry ed. 1.0
- ESS8 Data Protocol ed. 1.4

Fieldwork Documents

- ESS8 Source Questionnaires
- ESS8 Source Showcards
- ESS8 Source Contact Forms
- ESS8 Source Project Instructions

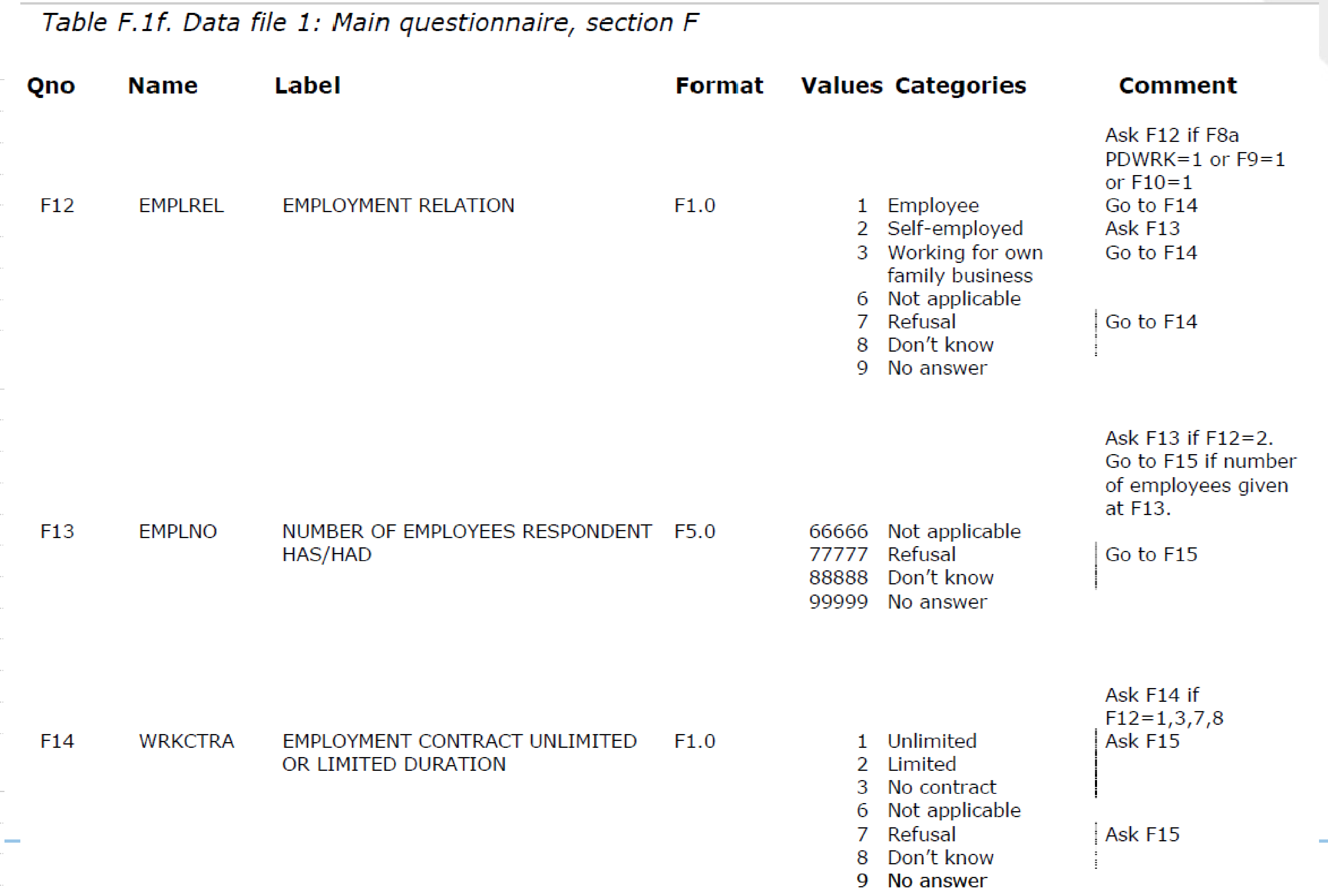


Table F.1f. Data file 1: Main questionnaire, section F

Qno	Name	Label	Format	Values	Categories	Comment
						Ask F12 if F8a PDWRK=1 or F9=1 or F10=1
F12	EMPLREL	EMPLOYMENT RELATION	F1.0	1	Employee	Go to F14
				2	Self-employed	Ask F13
				3	Working for own family business	Go to F14
				6	Not applicable	
				7	Refusal	Go to F14
				8	Don't know	
				9	No answer	
						Ask F13 if F12=2. Go to F15 if number of employees given at F13.
F13	EMPLNO	NUMBER OF EMPLOYEES RESPONDENT HAS/HAD	F5.0	66666	Not applicable	
				77777	Refusal	Go to F15
				88888	Don't know	
				99999	No answer	
						Ask F14 if F12=1,3,7,8
F14	WRKCTRA	EMPLOYMENT CONTRACT UNLIMITED OR LIMITED DURATION	F1.0	1	Unlimited	Ask F15
				2	Limited	
				3	No contract	
				6	Not applicable	
				7	Refusal	Ask F15
				8	Don't know	
				9	No answer	

File naming strategy

Principal identifier: provide useful clues to the content, status and version of a file, uniquely identify a file; help in classifying and sorting files; be consistent in time and among different people.

Consider following elements:

- Version number
- Date of creation (date format YYYY-MM-DD)
- Name of creator
- Description of content
- Name of research team
- Publication date
- Project number

Example:

20130311_interview2_audio.wav

20130311_interview2_trans.rtf

20130311_interview2_image.jpg

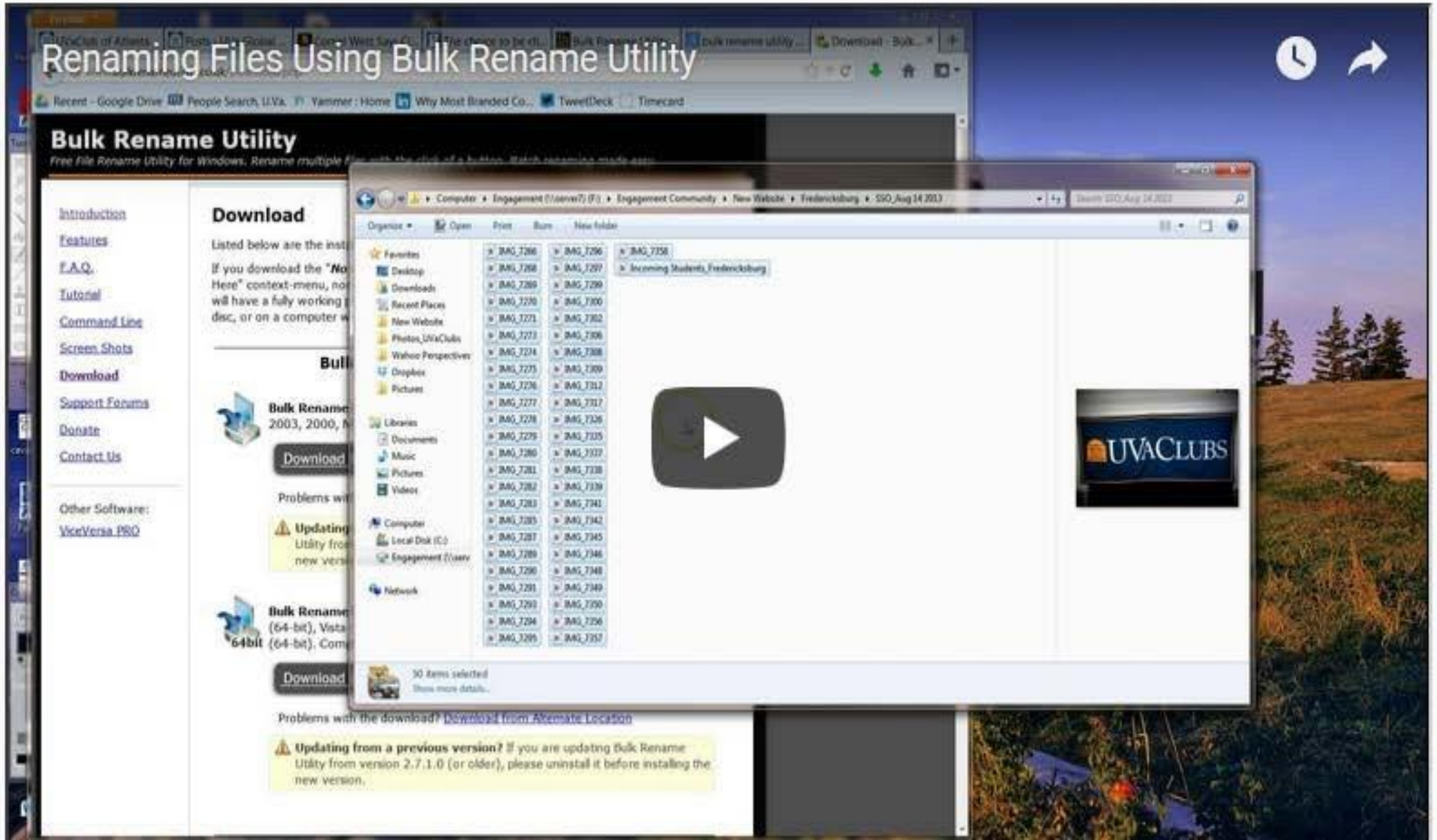
Setting up file naming convention:
<date><type><ID1><gender><age><municipality><datatype><ID2>

Expert Tour Guide on Data Management

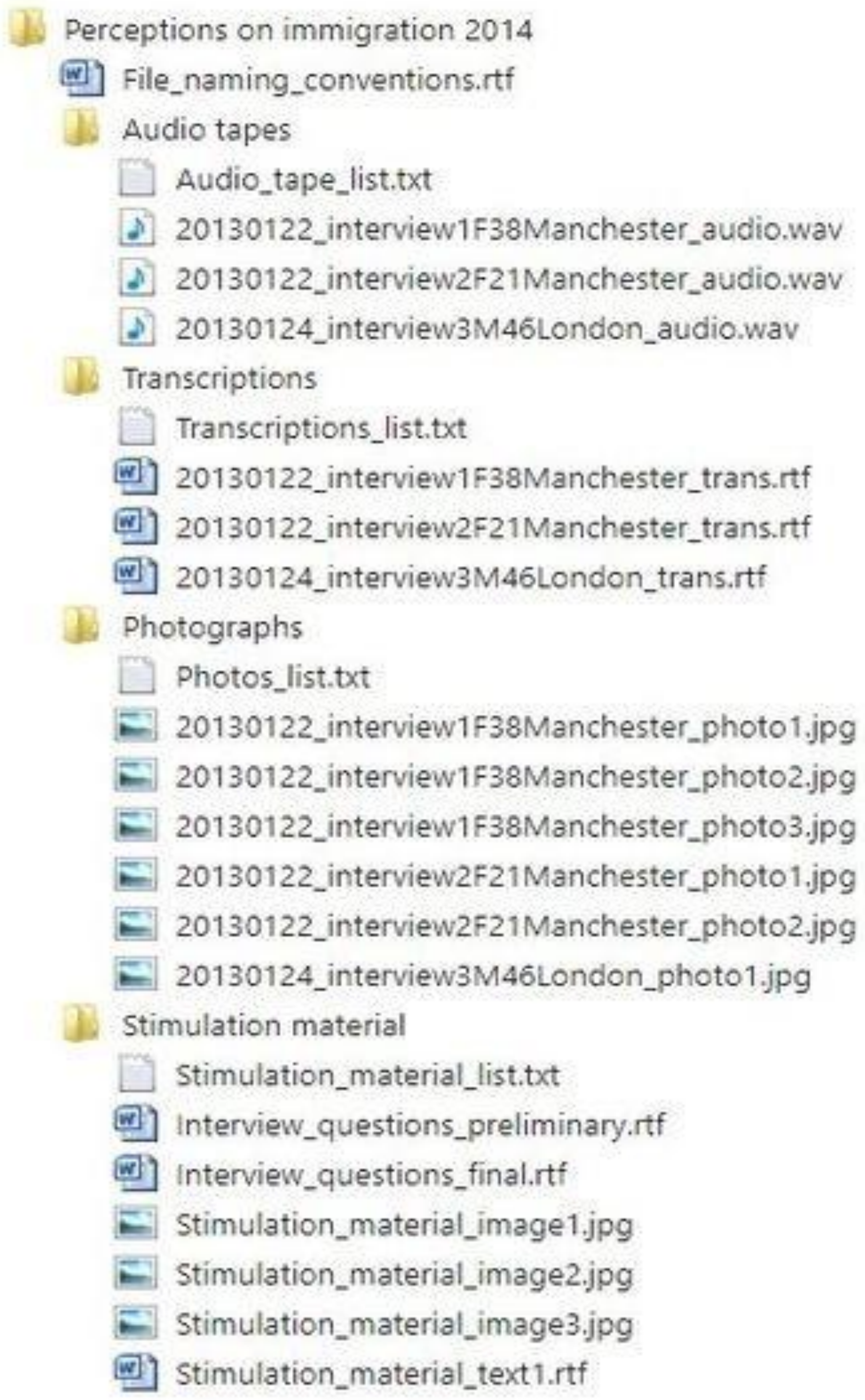
1. Plan
2. Organise & Document
   - Designing a data file structure
     - Organisation of variables
   - **File naming and folder structure**
   - Documentation and metadata
   - Adapt your DMP: part 2
   - Sources and further reading
3. Process
4. Store
5. Protect
6. Archive & Publish

How to ... use Bulk Rename Utility

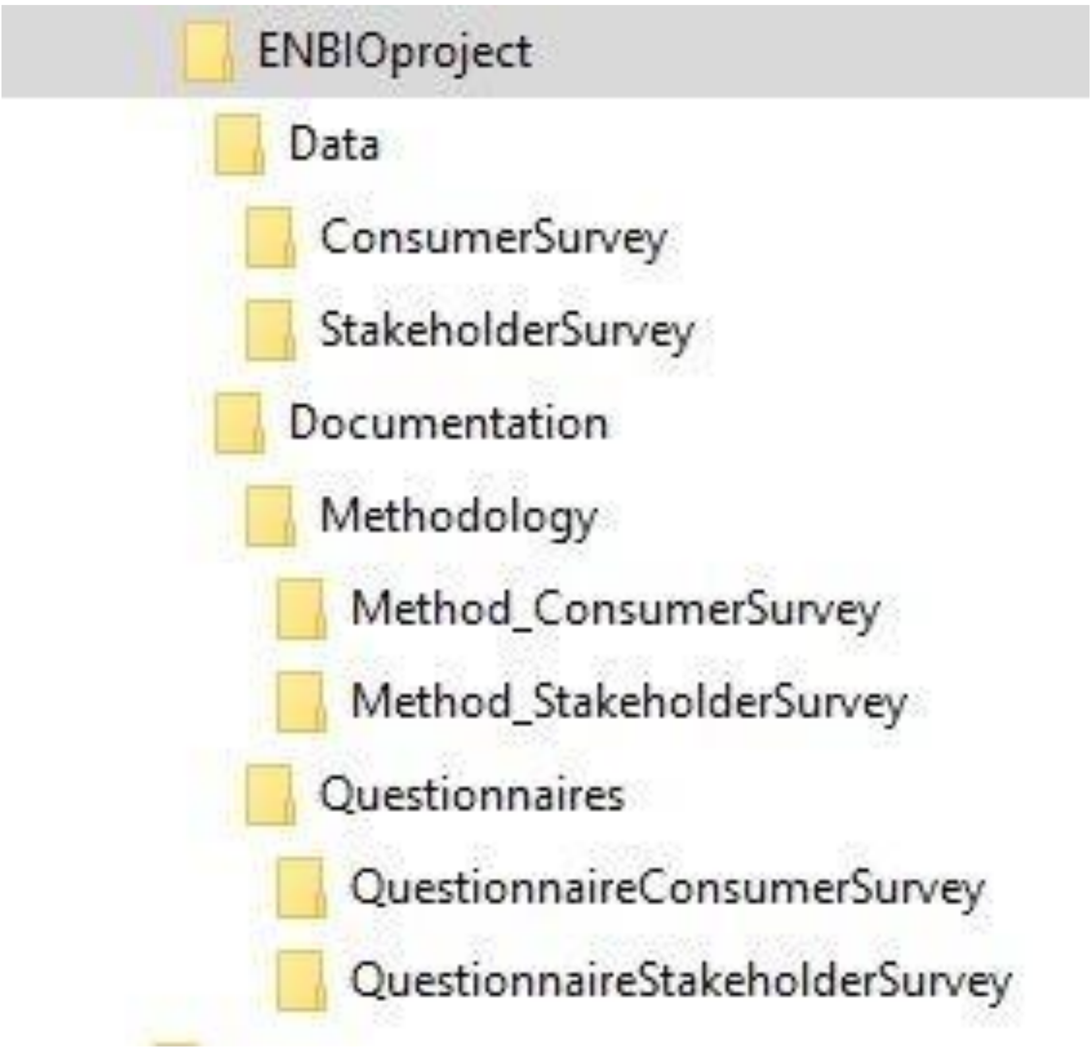
Follow the steps in the video to use Bulk Rename Utility to batch rename your files.


Renaming Files Using Bulk Rename Utility
Bulk Rename Utility
Free File Rename Utility for Windows. Rename multiple files with the click of a button. Batch renaming made easy.
Introduction
Features
F.A.Q.
Tutorial
Command Line
Screen Shots
Download
Support Forums
Donate
Contact Us
Other Software:
ViceVersa PRO
Download
Download
Download
Problems with the download? Download from Alternate Location
Updating from a previous version? If you are updating Bulk Rename Utility from version 2.7.1.0 (or older), please uninstall it before installing the new version.
30 items selected
UVACLUBS

Folder structure



- Perceptions on immigration 2014
  - File_naming_conventions.rtf
  - Audio tapes
    - Audio_tape_list.txt
    - 20130122_interview1F38Manchester_audio.wav
    - 20130122_interview2F21Manchester_audio.wav
    - 20130124_interview3M46London_audio.wav
  - Transcriptions
    - Transcriptions_list.txt
    - 20130122_interview1F38Manchester_trans.rtf
    - 20130122_interview2F21Manchester_trans.rtf
    - 20130124_interview3M46London_trans.rtf
  - Photographs
    - Photos_list.txt
    - 20130122_interview1F38Manchester_photo1.jpg
    - 20130122_interview1F38Manchester_photo2.jpg
    - 20130122_interview1F38Manchester_photo3.jpg
    - 20130122_interview2F21Manchester_photo1.jpg
    - 20130122_interview2F21Manchester_photo2.jpg
    - 20130124_interview3M46London_photo1.jpg
  - Stimulation material
    - Stimulation_material_list.txt
    - Interview_questions_preliminary.rtf
    - Interview_questions_final.rtf
    - Stimulation_material_image1.jpg
    - Stimulation_material_image2.jpg
    - Stimulation_material_image3.jpg
    - Stimulation_material_text1.rtf



- ENBIOproject
  - Data
    - ConsumerSurvey
    - StakeholderSurvey
  - Documentation
    - Methodology
      - Method_ConsumerSurvey
      - Method_StakeholderSurvey
    - Questionnaires
      - QuestionnaireConsumerSurvey
      - QuestionnaireStakeholderSurvey

Exercise

Describe a model research situation and ask your students to develope appropriate DMP including elements regarding data organisation.

E.g.: Survey of households; different survey instruments: F2F household questionnaire, F2F personal questionnaire; drop off personal questionnaire; proxy questionnaire. Required analysis on both levels, household and individual...

- How will you organise your data? Will the data be organised in simple files or more complex databases?
- Strategy in variable naming and labeling?
- File naming strategy, convention?

Why to document your data?

- Documentation = information about your data
- Makes the data publishable, discoverable, citable and reusable
- The data quality becomes clear = using data without information about them is„no quality"

How to start with documentation?

- Documentation connected to your research plan = you already have some of it.

Imagine that other researchers use your data; what would they need to know?

- Data archives have their standards, you can adopt them to guide you => Plan where to deposit your data
- Want others to use your data? Make your documentation international = use English.

Documentation – Two levels

1) Project level documentation 2) Data level documentation

1. Project-level documentation includes following information:

- For what purpose data was created
- What does the dataset contain
- How was data collected
- Who collected the data and when
- How was the data processed
- What manipulations were done
- How can data be accessed
- For more see ETG

2. Data-level (or object-level) documentation

Qualitative data:

- Information at the level of individual objects such as pictures, videos or interview transcripts
  - Interview transcripts, diaries, observations etc.: Background and contextual information described at the beginning of a file as a header or summary page = useful
- Textual data file (for example, interview) - Key information in the separate file

- QUESTION: Which are the key information for the qualitative interview?

Data level documentation: Qualitative data

Example:

- Interview date: 08.02.2013
- Interviewer: Matt Miller
- Pseudonym of interviewee: Ian (not the real first name of the interviewee)
- Occupation of interviewee: Journalist
- Age of interviewee: 32
- Gender of interviewee: Male?

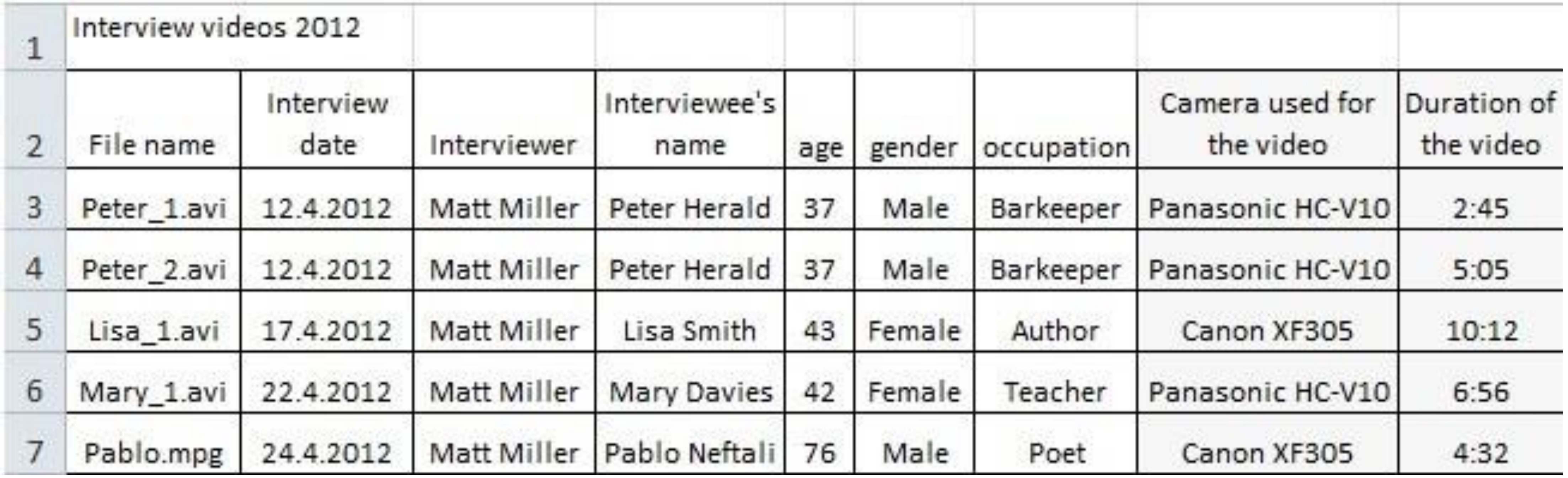
Data level documentation: Qualitative data

Qualitative data collections

(for example image or interview collections)

Concise list

Prepare a data list, aggregated information about each data item, in XLS (age, gender, occupation or location, and identifying details of the data items).



1	Interview videos 2012								
2	File name	Interview date	Interviewer	Interviewee's name	age	gender	occupation	Camera used for the video	Duration of the video
3	Peter_1.avi	12.4.2012	Matt Miller	Peter Herald	37	Male	Barkeeper	Panasonic HC-V10	2:45
4	Peter_2.avi	12.4.2012	Matt Miller	Peter Herald	37	Male	Barkeeper	Panasonic HC-V10	5:05
5	Lisa_1.avi	17.4.2012	Matt Miller	Lisa Smith	43	Female	Author	Canon XF305	10:12
6	Mary_1.avi	22.4.2012	Matt Miller	Mary Davies	42	Female	Teacher	Panasonic HC-V10	6:56
7	Pablo.mpg	24.4.2012	Matt Miller	Pablo Neftali	76	Male	Poet	Canon XF305	4:32

Same for = Audiovisual data files (image, audio or video), Image collections.

Data level documentation: Qualitative data

Material collected from online periodicals:

- Save references to web resources, like URLs (!! web content may change over time)
- Copy articles into a word processing program;

Materials from periodicals: articles, photographs and other materials:

- Detailed bibliographic info (author(s), title, date of publication etc.);
- Make a list of articles, photos etc., sort alphabetically or chronologically.

Data level documentation: Qualitative data

Storing documentation:

- Write the documentation into a separate, well-structured file, and associate that with the data file.

Use the same filename stem:

- 20130311_interviews_audio,
- 20130311_interviews_trans,
- 20130311_interviews_image,
- 20130311_interviews_metadata.

- The latter part of the name = the specifics of the file (audio, transcription, image, metadata).

2. Data-level (or object-level) documentation

Quantitative data

Information about the data file:

- Data type, file type and format, size, data processing scripts (syntax).

Information about the variables in the file:

- Descriptions of variables: variable names, variable values, value labels
- Description of derived variables
- If applicable, frequencies, basic contingencies etc. Exact original wording of the question
- Variables in a database, SPSS - information in the Variable view, not in the XLS
- See Expert tour

Data level documentation: Quantitative data

Variable labels should:

- Be brief with a maximum of 80 characters;
- Indicate the unit of measurement, where applicable;
- Reference the question number of a survey or questionnaire, where applicable.

Example:

- Variable: 'Q11eximp'
- Variable label: 'Q11: How important is exercise for you?
- Value labels: 1: Very unimportant. 2. Unimportant. 3. Neutral. 4. Important. 5. Very important.

Data level documentation: Quantitative data

Information about the cases in the file

- A specification of each case (units of research like e.g. a respondent, a household etc.)

- Description of the missing values at each variable – code?
- Description of the weighting variable – how it was constructed
- Explanation or definition of codes and classification schemes used – e.g. ISCO

Metadata: Templates and Standards

- Metadata= data documentation, data about data
- Machine readable metadata: used by data archives, not necessarily produced by data creators

- You provide the archive with data documentation document
- Use **metadata template** so you include all important information

Metadata templates (for easy starting)

- Simple
- Basic necessary information
- Help to create basic documentation
- See different metadata templates in Expert Tour

Example: York University Library Metadata Template

1.Title

2.Creator

3.Subject

4.Description

5.Publisher

6.Contributor

7.Date

8.Type

9.Format

10.Identifier

11.Source

12.Language

13.Relation

14.Coverage

15.Rights

Metadata standards

(for when you need your metadata to be very structured, archives)

- At first look seem quite scary – too complex
- They are used by data archives
- When you submit your dataset at a trusted data repository, these standards are automatically applied.

Metadata standards are discipline specific:

- DDI for social sciences
- See accordion
- Example of DDI >> Next slide

Example of DDI:

Datafile: Galanakis, Michail (University of Helsinki): Intercultural Urban Public Space in Toronto 2011-2013 [dataset]. Version 1.0 (2014-02-13). Finnish Social Science Data Archive [distributor]

The machine-readable XML file of DDI documentation of a data file looks like this:

See link: https://services.fsd.uta.fi/catalogue/FSD2926/DDI/FSD2926e.xml

DDI information - Expert tour (it is too long for the PPP)

Metadata for new data types – new standards still under development

- Emerging new data types (big data, social media data etc.) also would require metadata and documentation
- To provide metadata for social media data and transaction data with metadata, theo metadata standards by the Data Documentation Initiative (DDI) should serve as the guiding framework.
- However, the DDI standard is still insufficient to document all the details necessary for reproducing a social media dataset
  - For example, the DDI format does not allow describing biases caused by data mining interfaces of social media platforms, changes in data availability and formats, explanations about code and scripts used in collection, cleaning and analysis etc. Such information can be described only in an unstructured manner as an additional comment in the standard's form.

cessda

Thank you for your attention!

Jindřich Krejčí <jindrich.krejci@soc.cas.cz>
Johana Chylíková <johana.chylikova@soc.cas.cz>

cessda.eu

@CESSDA_Data