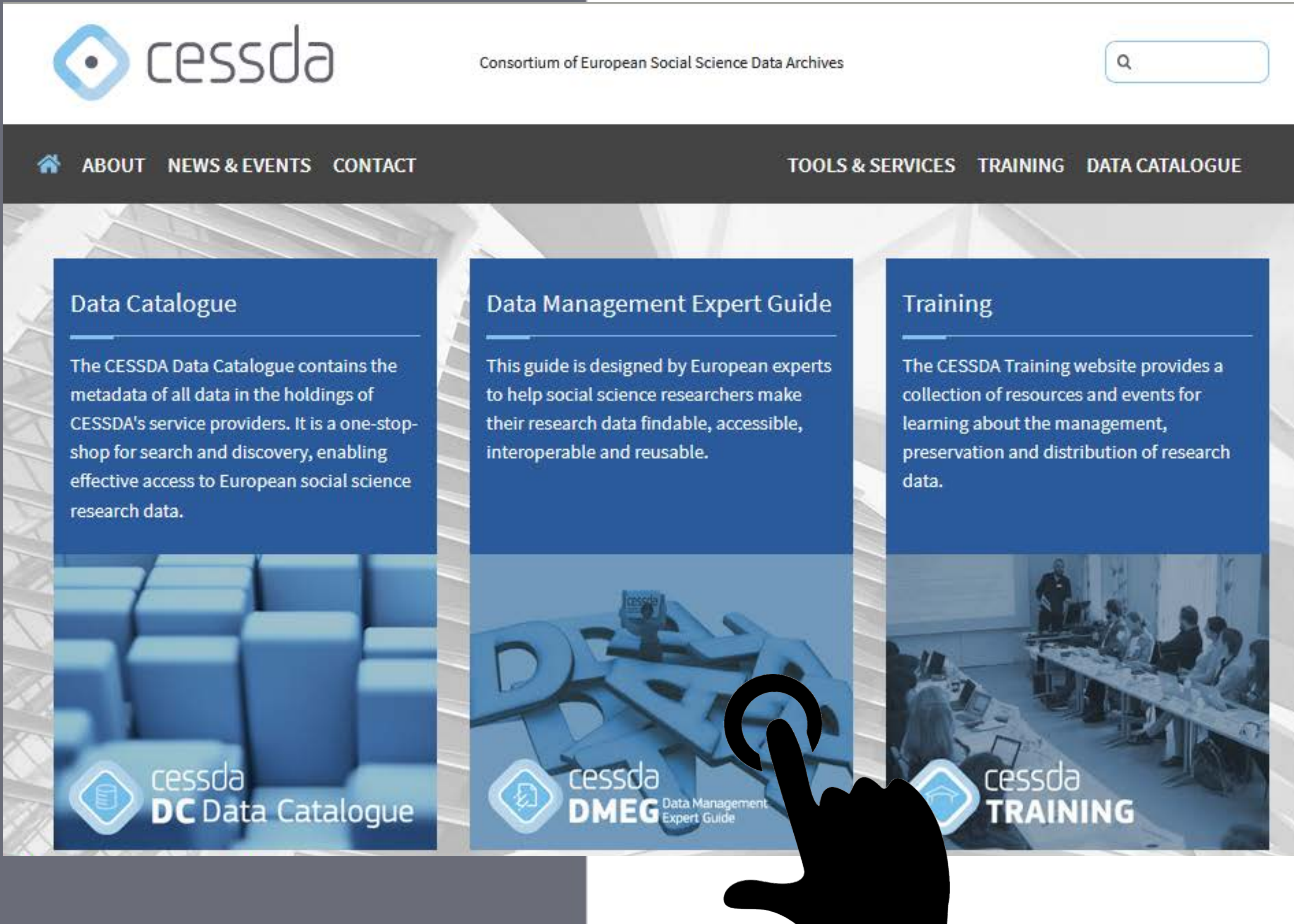
Data Management Expert Guide Chapters 1, 2, 3 & 4

Jindřich Krejčí | CSDA
<jindrich.krejci@soc.cas.cz>

2nd CESSDA Widening 2018
Belgrade | 14 - 15 November 2018

CESSDA DMEG
at www. cessda.eu



cessda

Consortium of European Social Science Data Archives

Data Catalogue

The CESSDA Data Catalogue contains the metadata of all data in the holdings of CESSDA's service providers. It is a one-stop-shop for search and discovery, enabling effective access to European social science research data.

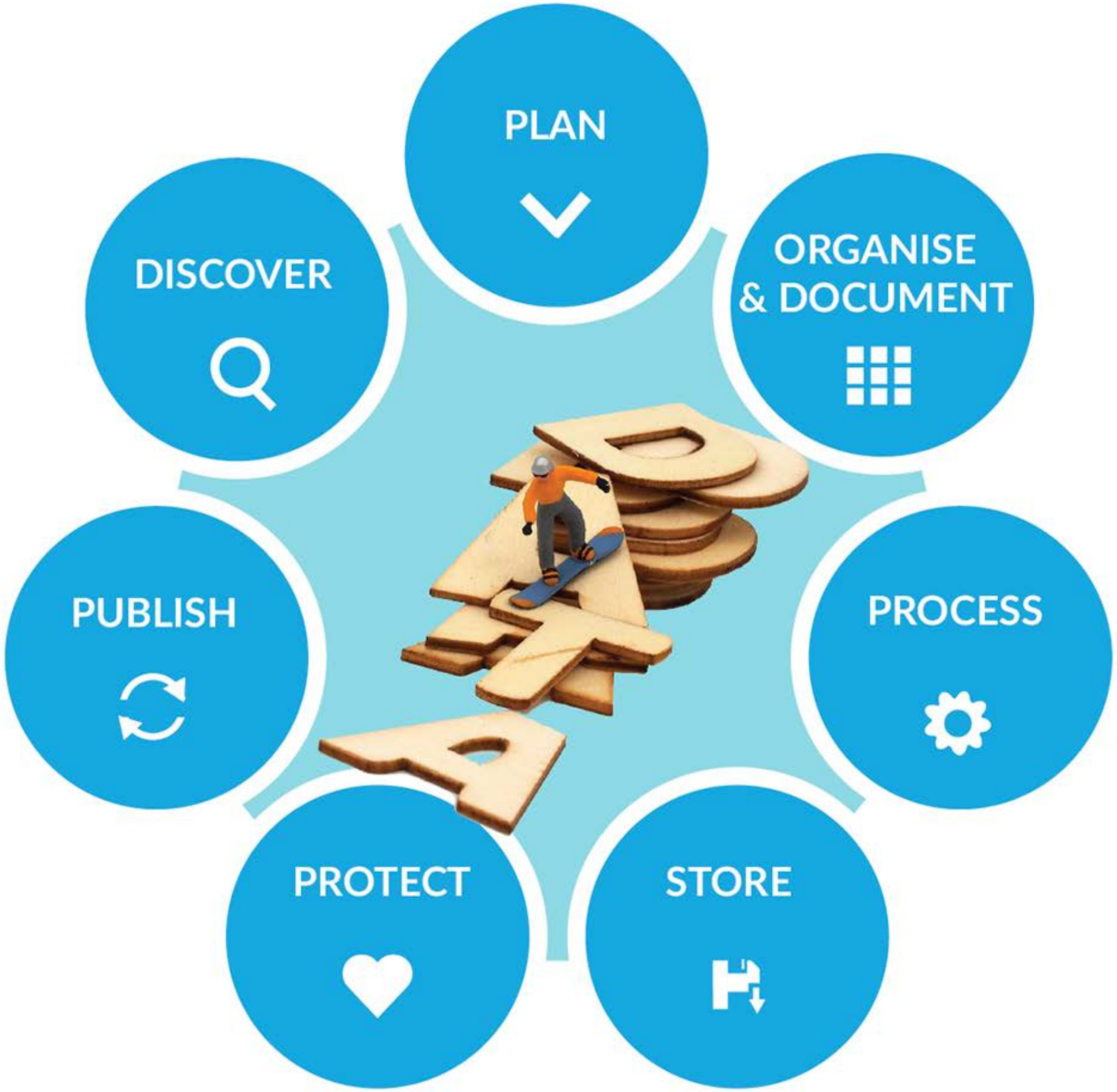
Data Management Expert Guide

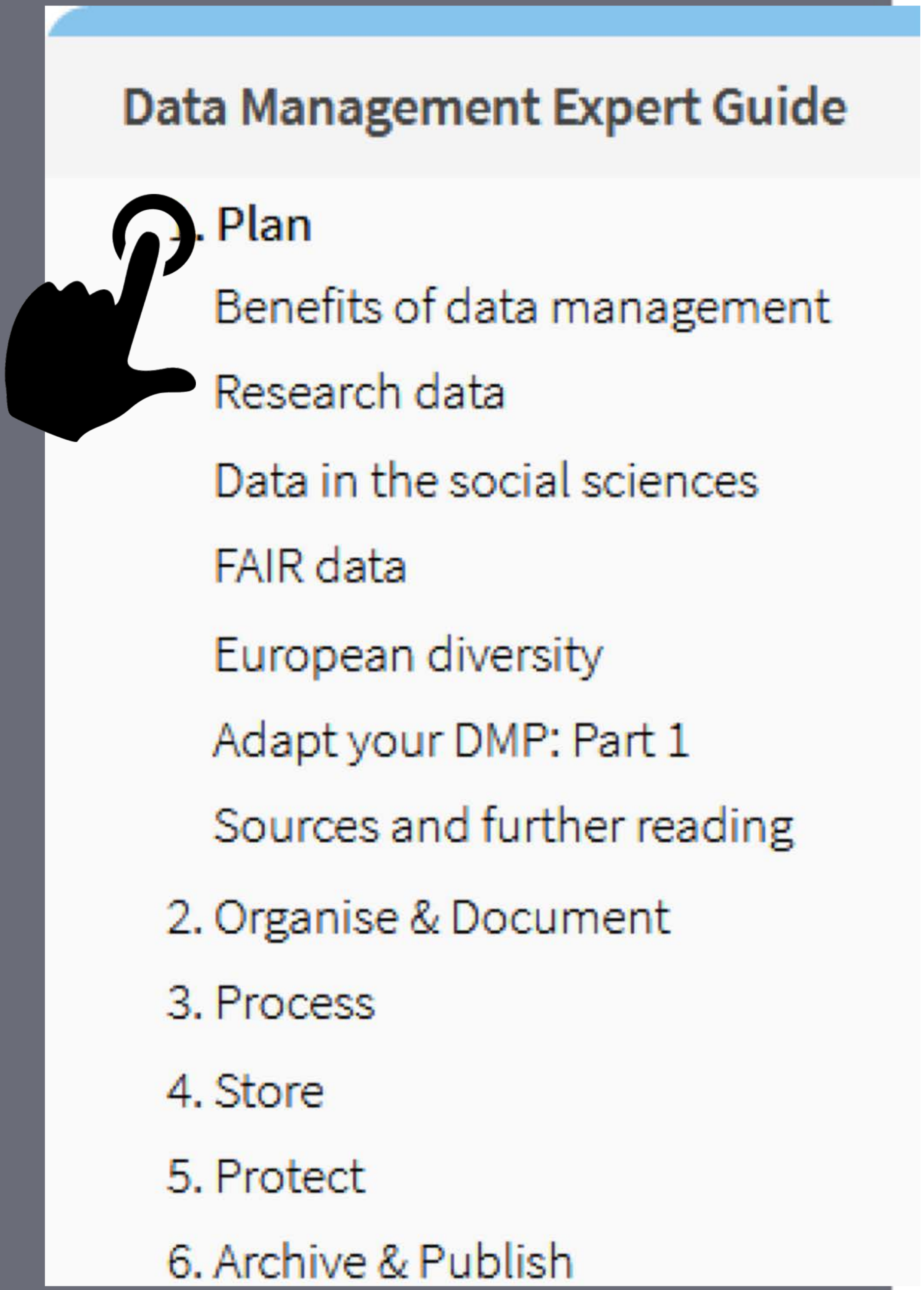
This guide is designed by European experts to help social science researchers make their research data findable, accessible, interoperable and reusable.

Training

The CESSDA Training website provides a collection of resources and events for learning about the management, preservation and distribution of research data.

CESSDA Training Working Group (2017). *CESSDA Data Management Expert Guide.* Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMEG





Data Management Expert Guide

1. Plan
   - Benefits of data management
   - Research data
   - Data in the social sciences
   - FAIR data
   - European diversity
   - Adapt your DMP: Part 1
   - Sources and further reading
2. Organise & Document
3. Process
4. Store
5. Protect
6. Archive & Publish

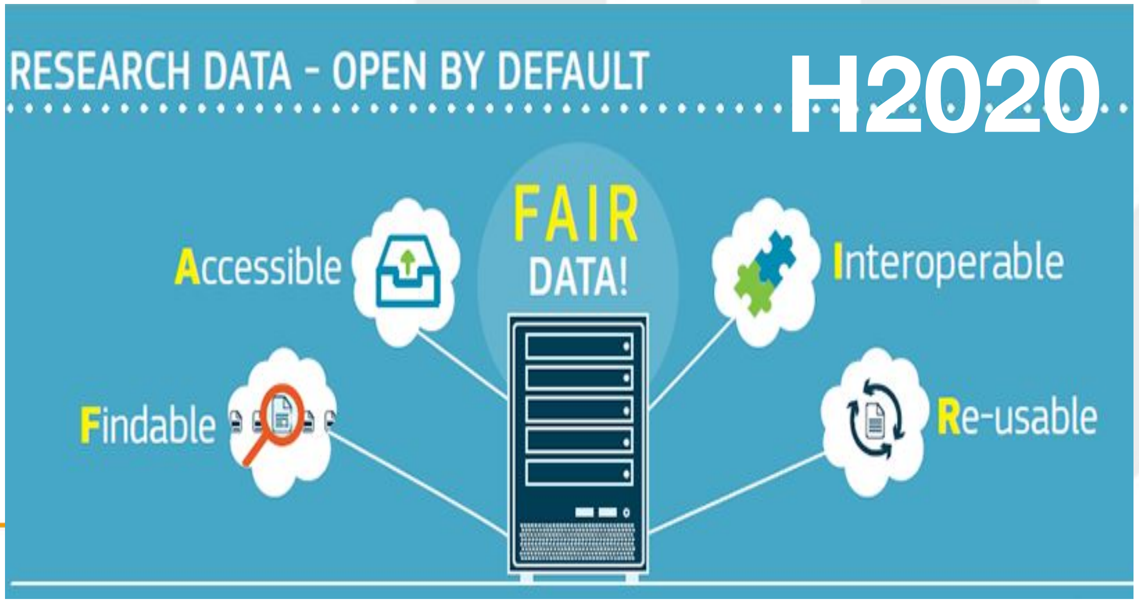
1. PLAN

- Personal data
- FAIR data principles
- Data management plan (DMP)
- DMP content elements
- Answer DMP questions and develope your own DMP

FAIR Data

FAIR Data Action Plan

Interim recommendations and actions from the European Commission



Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accesible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

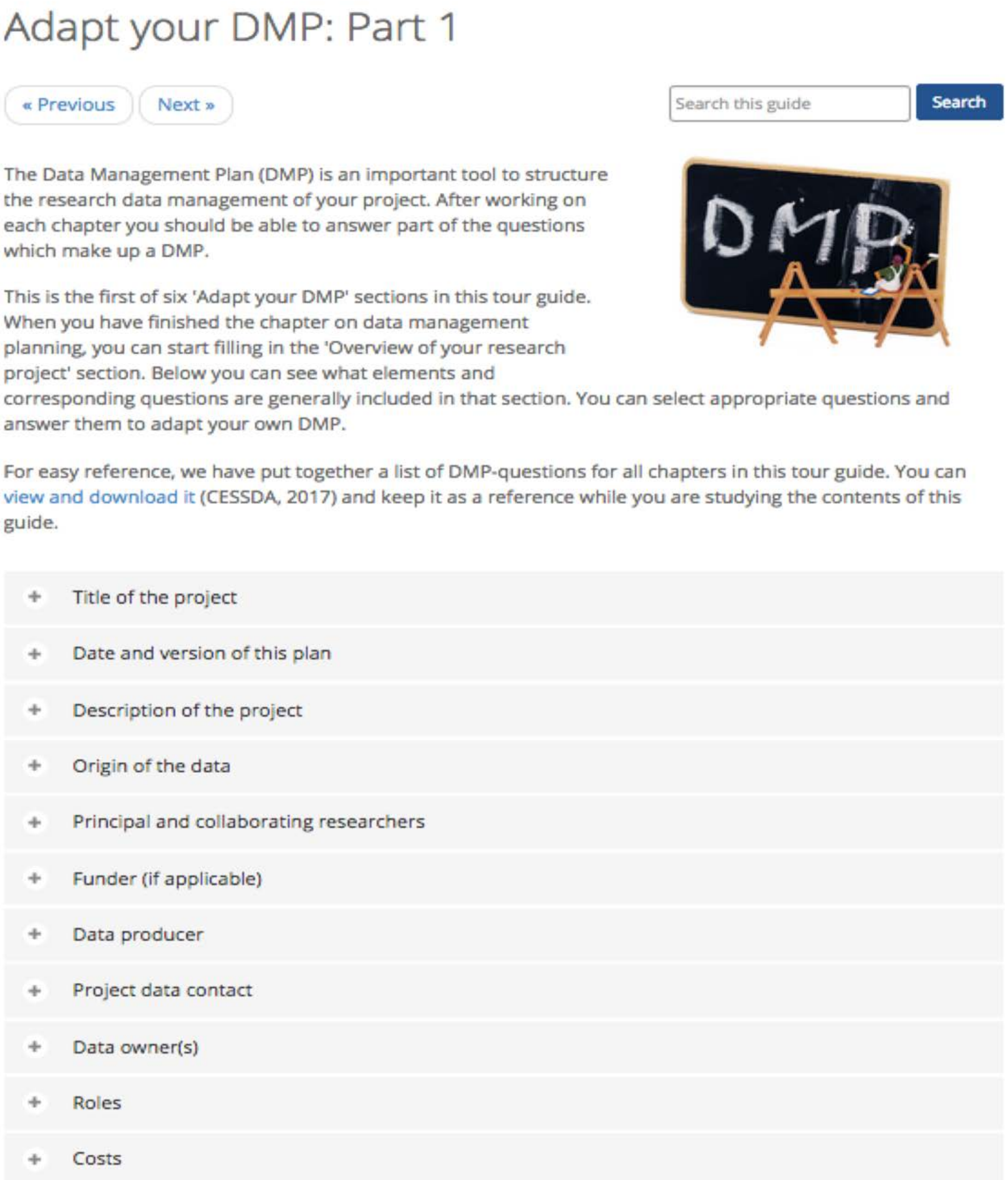
Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.

Data Management Plan (DMP)



Adapt your DMP: Part 1

The Data Management Plan (DMP) is an important tool to structure the research data management of your project. After working on each chapter you should be able to answer part of the questions which make up a DMP.

This is the first of six 'Adapt your DMP' sections in this tour guide. When you have finished the chapter on data management planning, you can start filling in the 'Overview of your research project' section. Below you can see what elements and corresponding questions are generally included in that section. You can select appropriate questions and answer them to adapt your own DMP.

For easy reference, we have put together a list of DMP-questions for all chapters in this tour guide. You can view and download it (CESSDA, 2017) and keep it as a reference while you are studying the contents of this guide.

- + Title of the project
- + Date and version of this plan
- + Description of the project
- + Origin of the data
- + Principal and collaborating researchers
- + Funder (if applicable)
- + Data producer
- + Project data contact
- + Data owner(s)
- + Roles
- + Costs

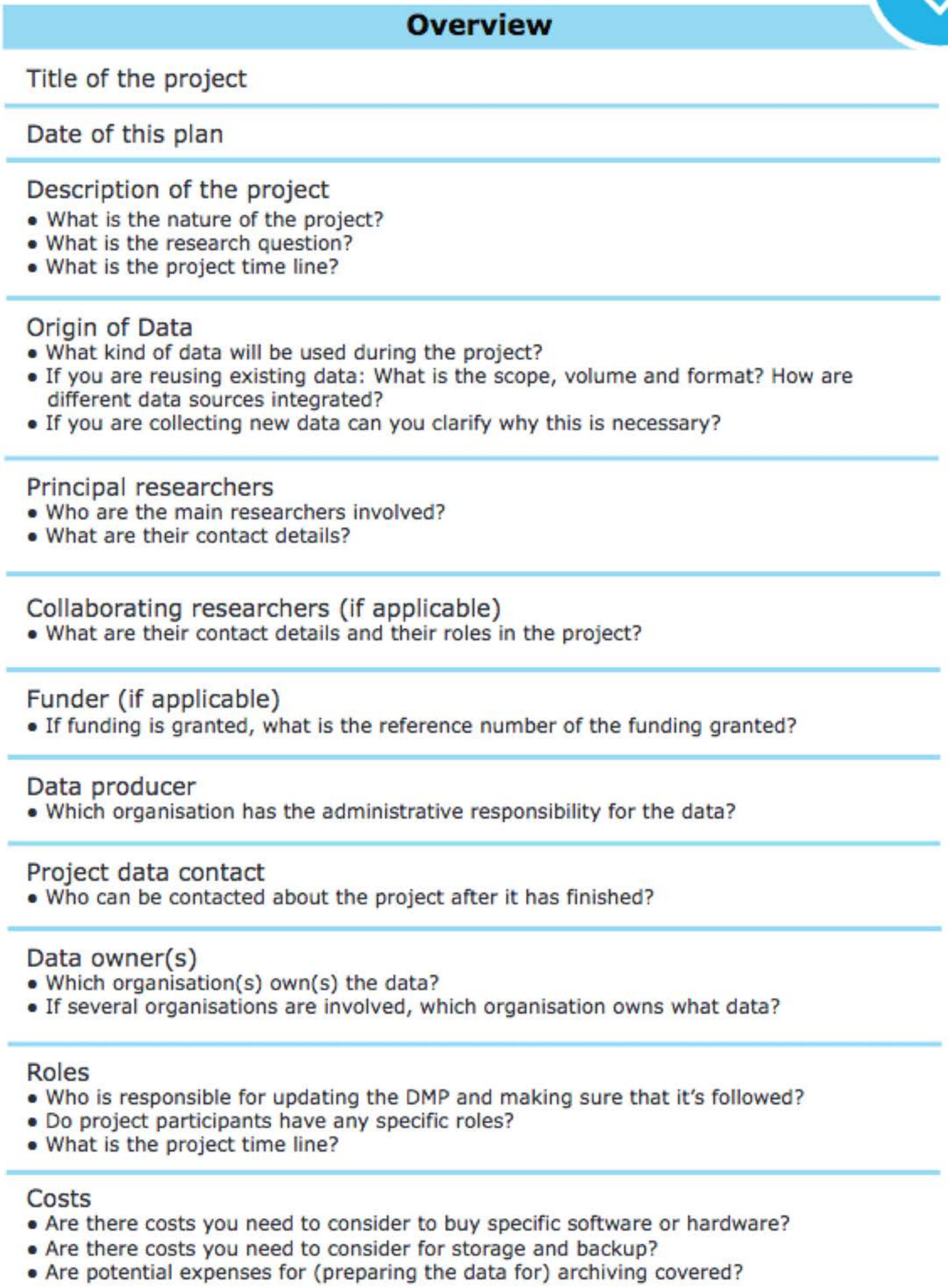
- Adapt your DMP section at the end of every chapter
- Correspronding questions to each chapter

PLAN

Downloadable DMP checklist

Adapt your Data Management Plan

A list of Data Management Questions based on the Expert Tour Guide on Data Management



Overview

Title of the project

Date of this plan

Description of the project
- What is the nature of the project?
- What is the research question?
- What is the project time line?

Origin of Data
- What kind of data will be used during the project?
- If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated?
- If you are collecting new data can you clarify why this is necessary?

Principal researchers
- Who are the main researchers involved?
- What are their contact details?

Collaborating researchers (if applicable)
- What are their contact details and their roles in the project?

Funder (if applicable)
- If funding is granted, what is the reference number of the funding granted?

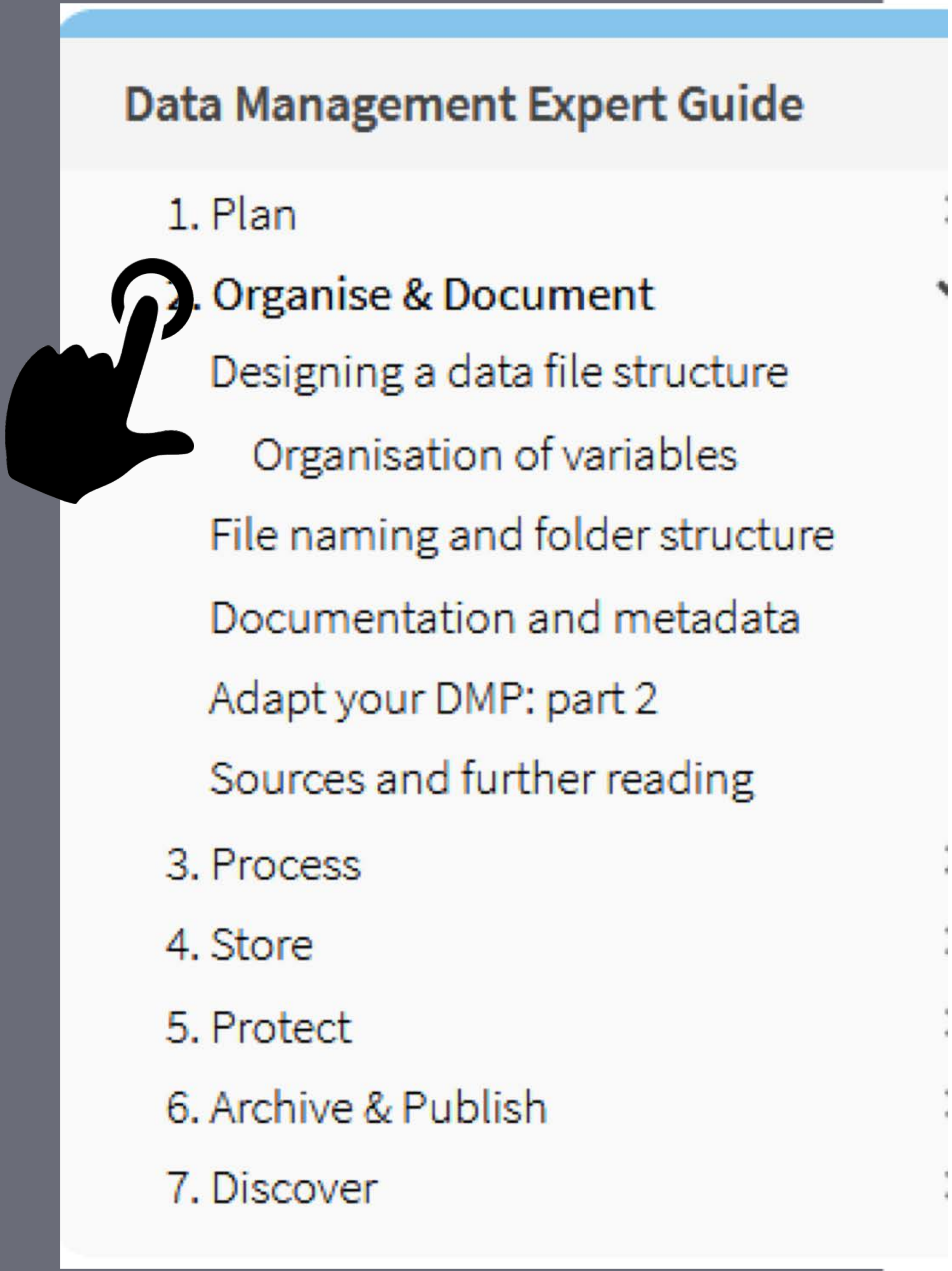
Data producer
- Which organisation has the administrative responsibility for the data?

Project data contact
- Who can be contacted about the project after it has finished?

Data owner(s)
- Which organisation(s) own(s) the data?
- If several organisations are involved, which organisation owns what data?

Roles
- Who is responsible for updating the DMP and making sure that it's followed?
- Do project participants have any specific roles?
- What is the project time line?

Costs
- Are there costs you need to consider to buy specific software or hardware?
- Are there costs you need to consider for storage and backup?
- Are potential expenses for (preparing the data for) archiving covered?

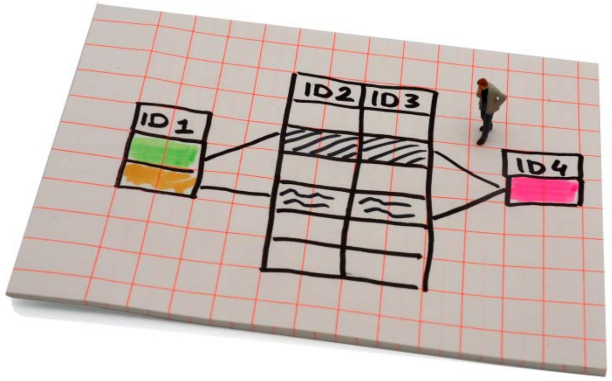


Data Management Expert Guide

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
7. Discover

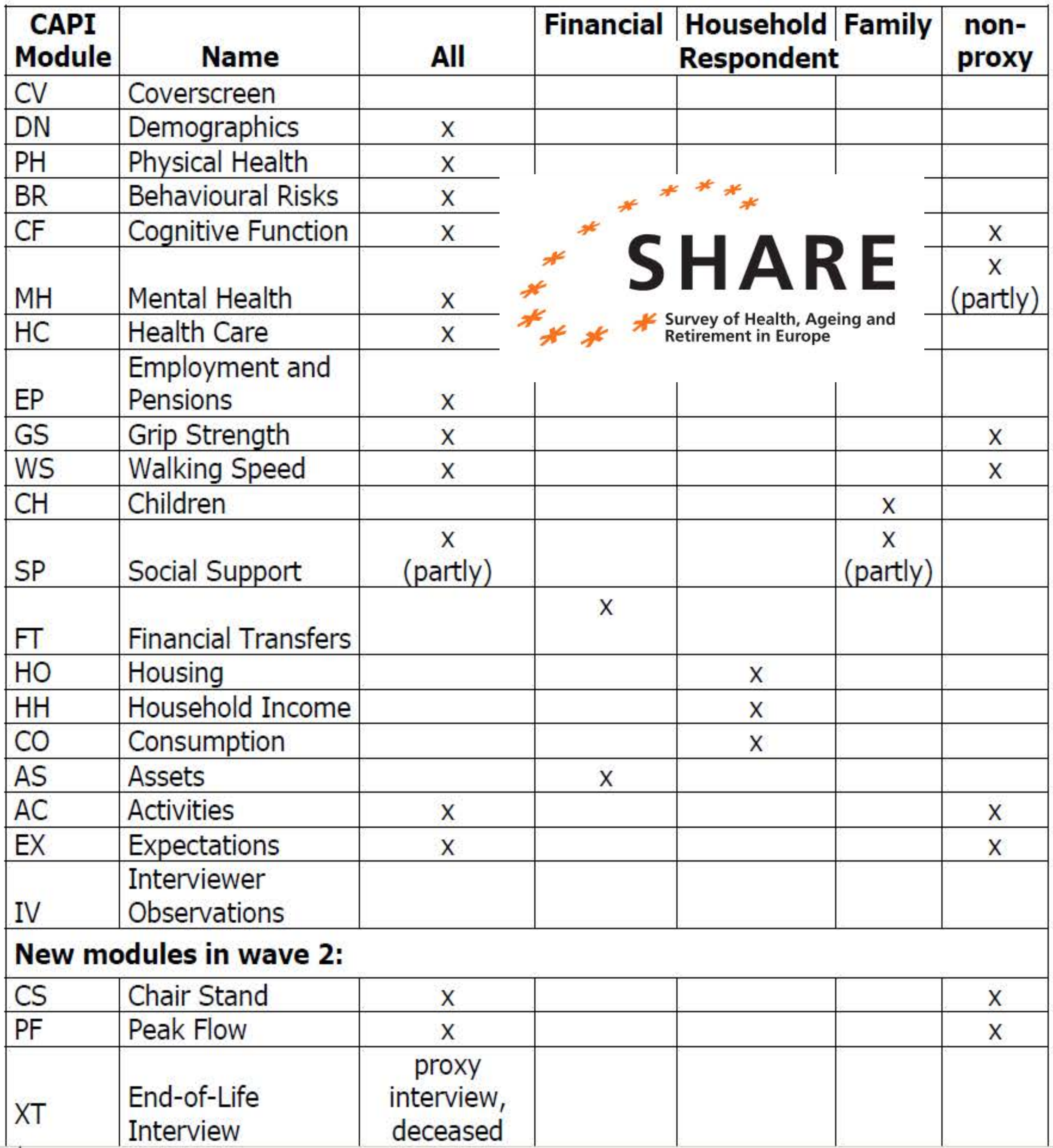
2. ORGANISE & DOCUMENT

- Organising data for research and data sharing
- Elements of data structure
- File naming, folder structure
- Data documentation
- Metadata standards



Data file structure

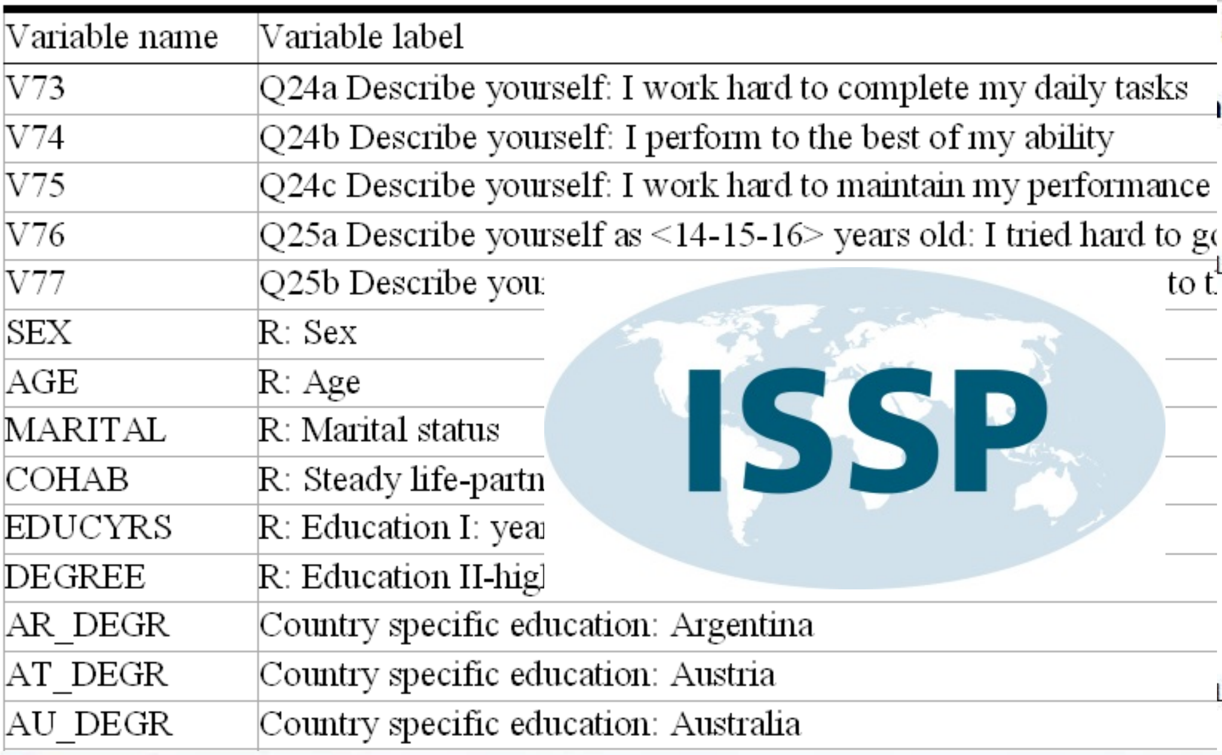
- Units of analysis / analytical objectives / methods of analysis
- Relations: different content items / sources of data/ other relevant external information
- Connections to other existing or future data
- Strategies for version control
- Technical limitations (e.g. the size, software)
- Software



CAPI Module	Name	All	Financial	Household Respondent	Family	non-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				

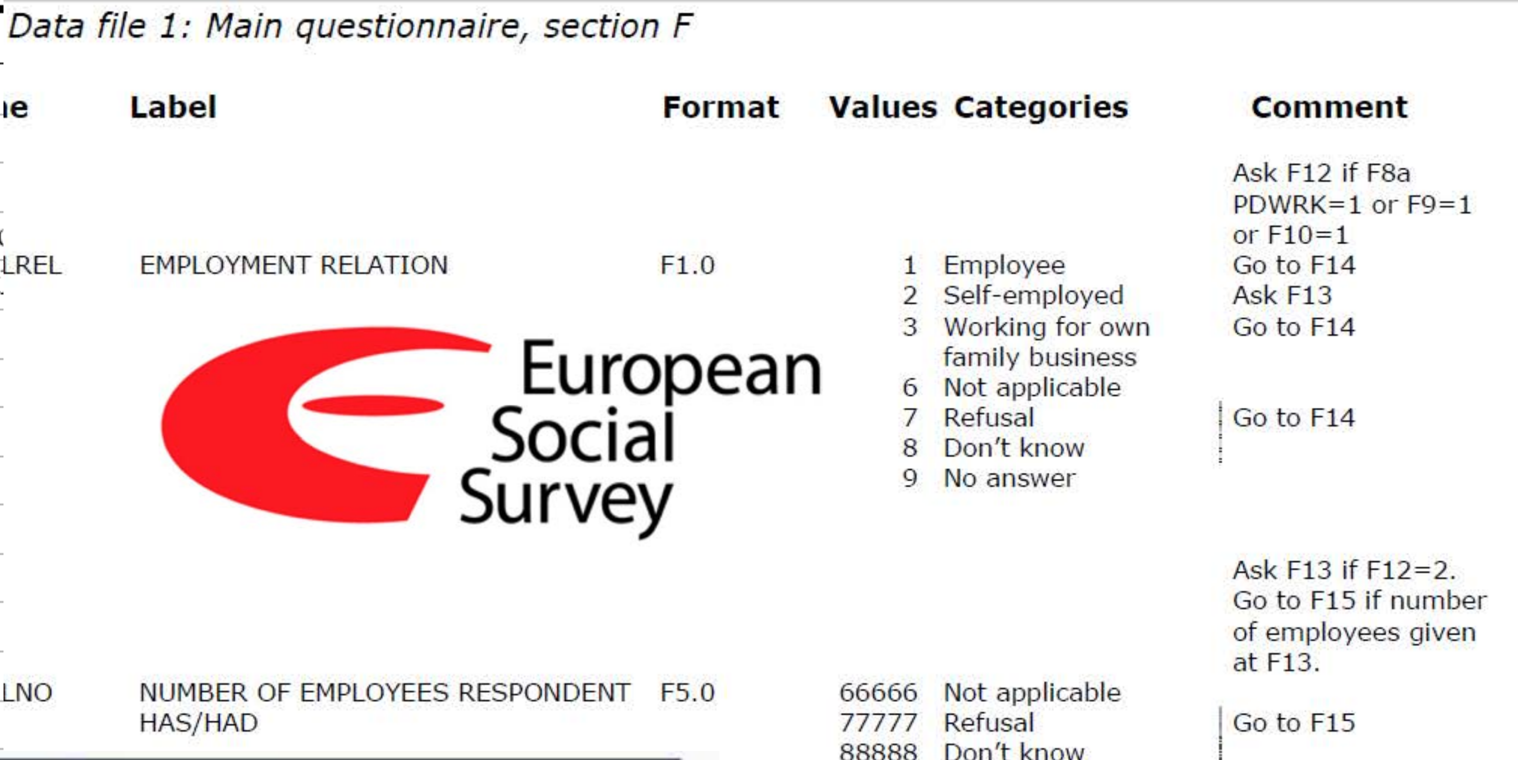
SHARE

Survey of Health, Ageing and Retirement in Europe



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 ge
V77	Q25b Describe you to t
SEX	R: Sex
AGE	R: Age
MARITAL	R: Marital status
COHAB	R: Steady life-partn
EDUCYRS	R: Education I: yea
DEGREE	R: Education II-higl
AR_DEGR	Country specific education: Argentina
AT_DEGR	Country specific education: Austria
AU_DEGR	Country specific education: Australia

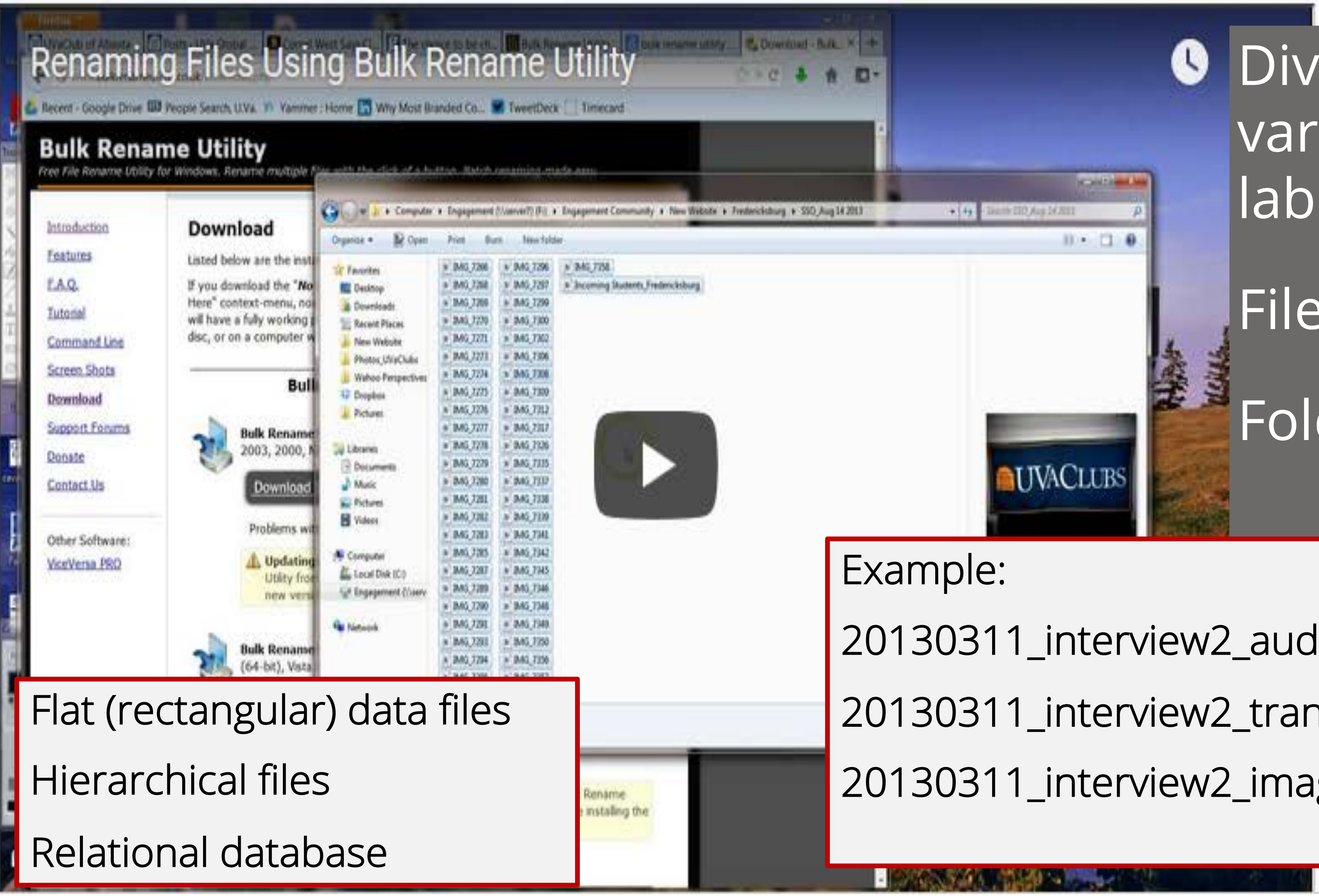
ISSP



Data file 1: Main questionnaire, section F

e	Label	Format	Values	Categories	Comment
LREL	EMPLOYMENT RELATION	F1.0	1 2 3 6 7 8 9	Employee Self-employed Working for own family business Not applicable Refusal Don't know No answer	Ask F12 if F8a PDWRK=1 or F9=1 or F10=1 Go to F14 Ask F13 Go to F14 Go to F14
LNO	NUMBER OF EMPLOYEES RESPONDENT HAS/HAD	F5.0	66666 77777 88888 99999	Not applicable Refusal Don't know No answer	Ask F13 if F12=2. Go to F15 if number of employees given at F13. Go to F15

European Social Survey



Dive in deeper - variable names and labels

File naming strategies

Folder structure

Flat (rectangular) data files

Hierarchical files

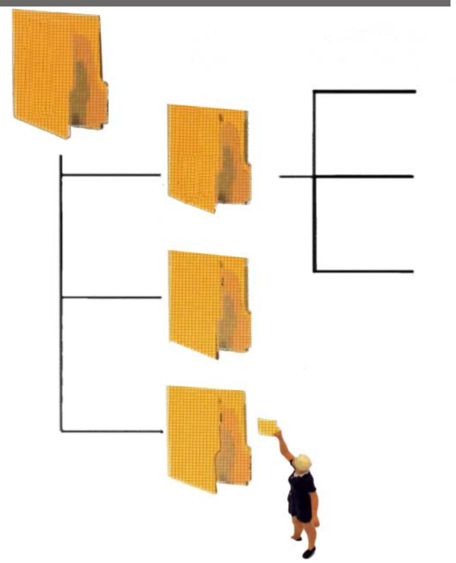
Relational database

Example:

20130311_interview2_audio.wav

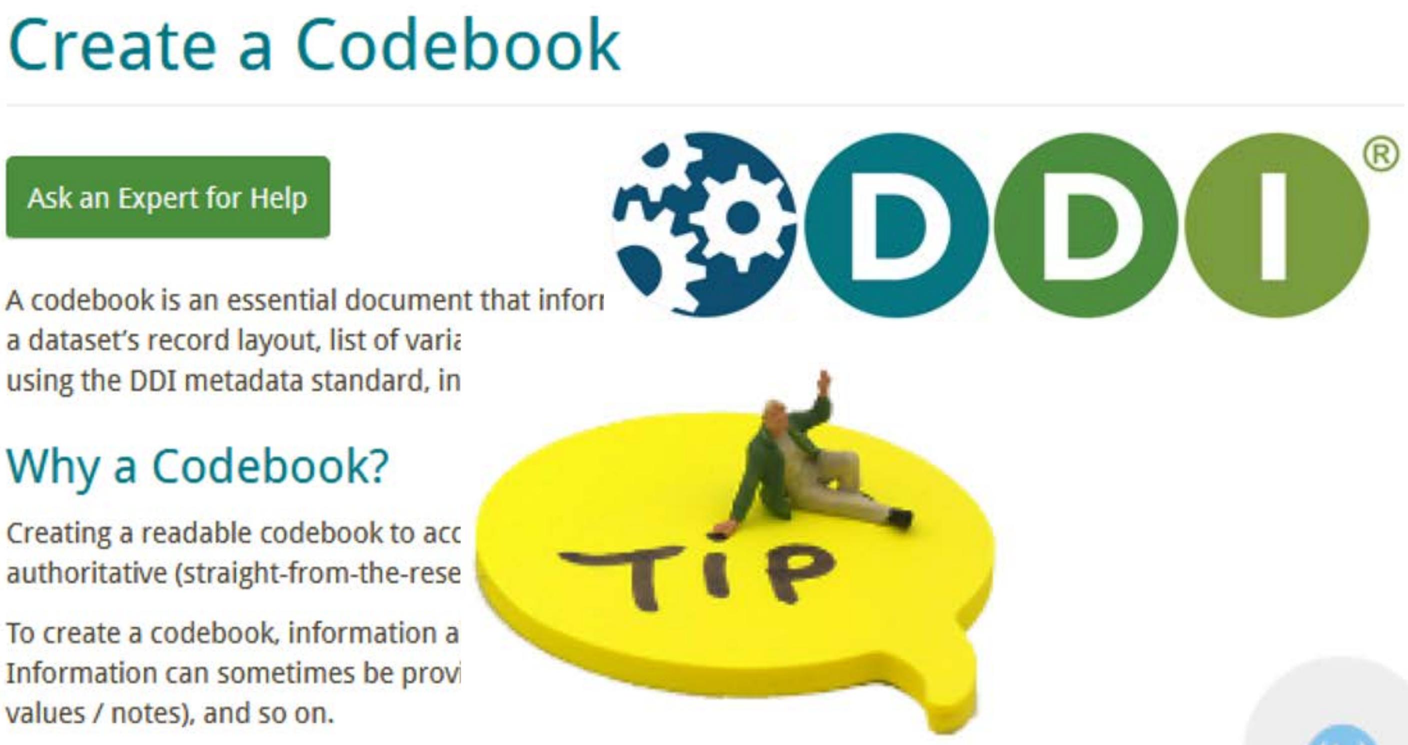
20130311_interview2_trans.rtf

20130311_interview2_image.jpg



Documentation & metadata

- Two levels of documentation: (1) project level documentation; (2) data level
- Quantitative and qualitative sections on data level



Create a Codebook

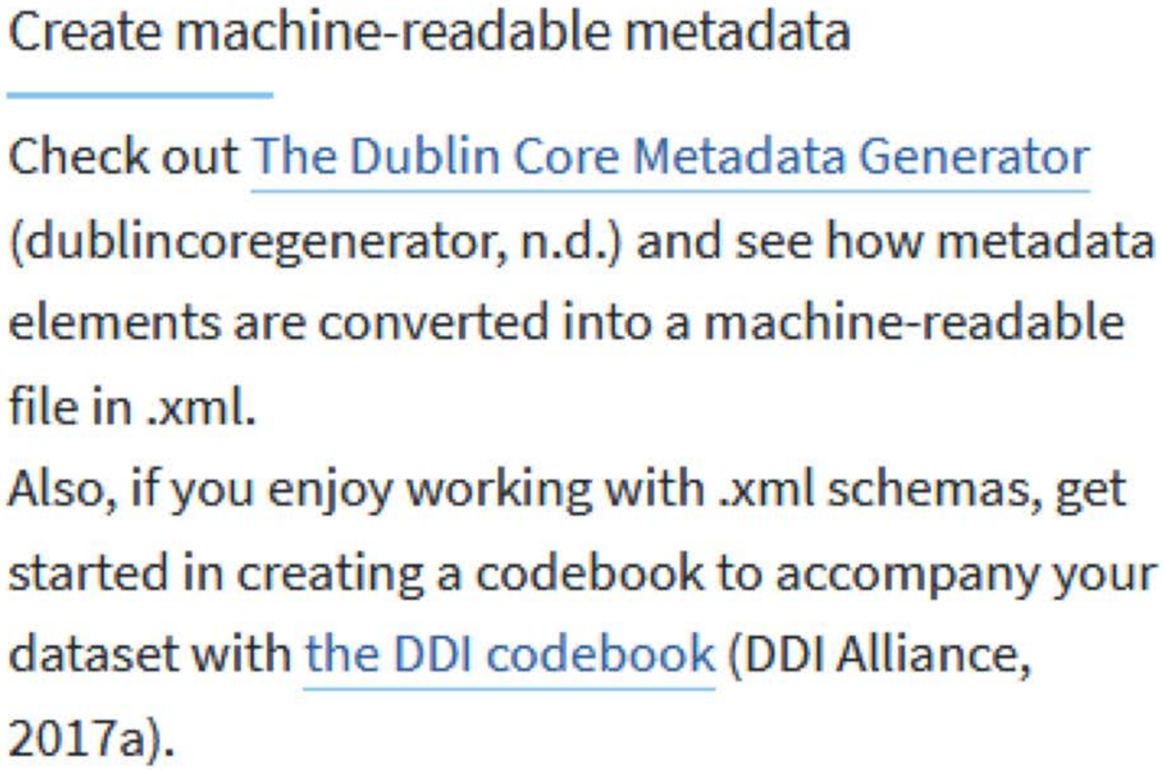
Ask an Expert for Help

A codebook is an essential document that infor
a dataset's record layout, list of varia
using the DDI metadata standard, in

Why a Codebook?

Creating a readable codebook to acc
authoritative (straight-from-the-rese

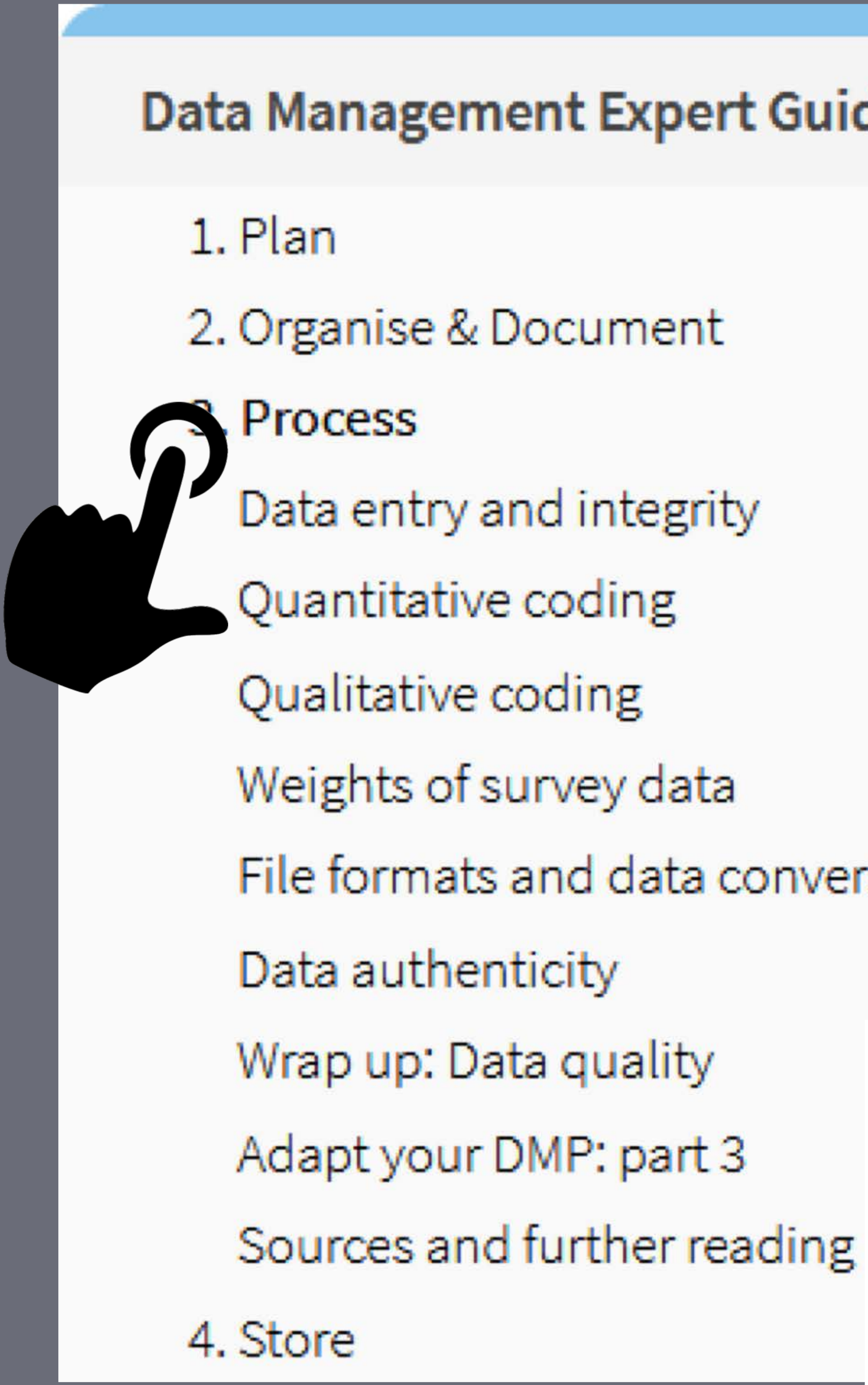
To create a codebook, information a
Information can sometimes be provi
values / notes), and so on.



Create machine-readable metadata

Check out The Dublin Core Metadata Generator (dublincoregenerator, n.d.) and see how metadata elements are converted into a machine-readable file in .xml.
Also, if you enjoy working with .xml schemas, get started in creating a codebook to accompany your dataset with the DDI codebook (DDI Alliance, 2017a).

3. PROCESS



Data Management Expert Guide

1. Plan

2. Organise & Document

3. Process

Data entry and integrity

Quantitative coding

Qualitative coding

Weights of survey data

File formats and data conversion

Data authenticity

Wrap up: Data quality

Adapt your DMP: part 3

Sources and further reading

4. Store

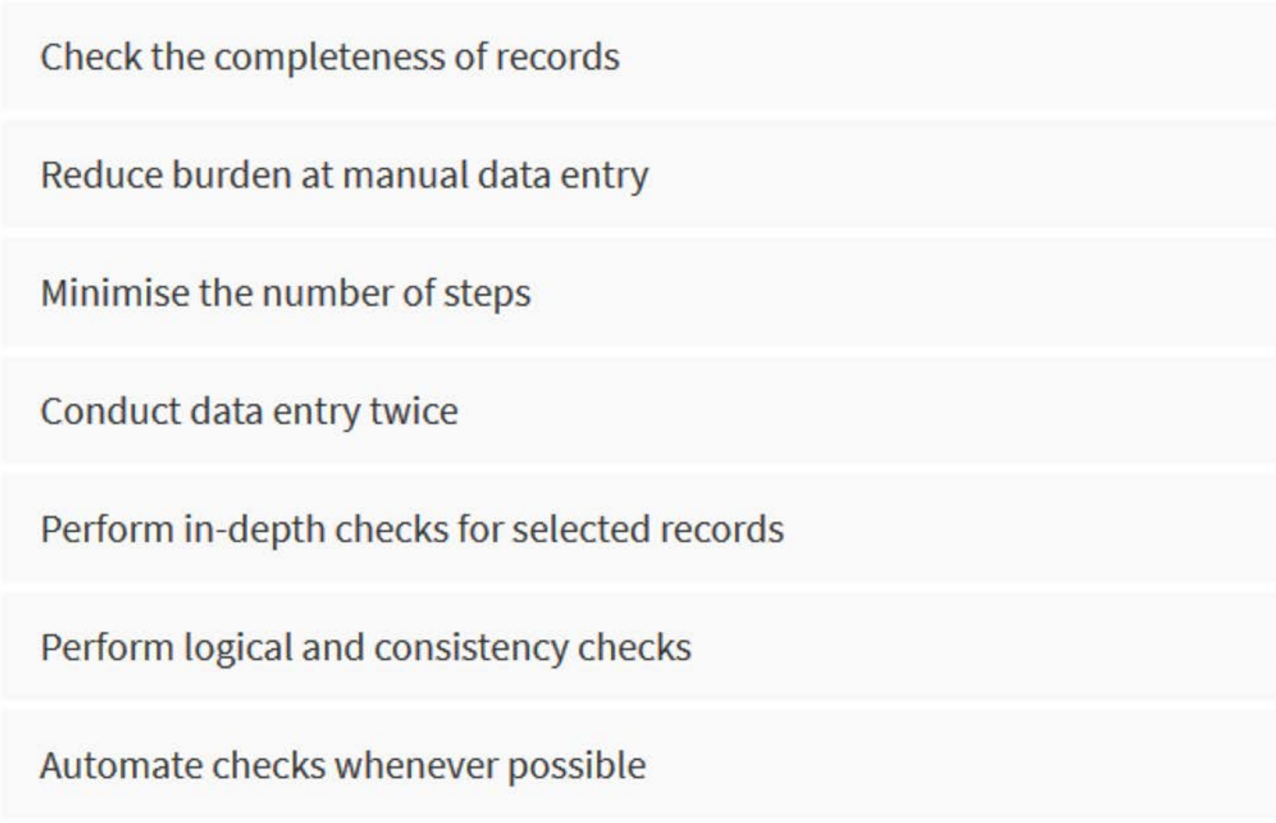
- Data entry
- Data coding (quantitative, qualitative)
- File formats
- Data integrity and authenticity
- Systematic approach to data quality

5 Service and sales workers
51 Personal service workers
511 Travel attendants, conductors and guides
512 Cooks
513 Waiters and bartenders
514 Hairdressers, beauticians and related workers
515 Building and housekeeping supervisors
516 Other personal services workers

Data entry and integrity

- Data integrity: assurance of the accuracy, consistency and completeness of original information in the data
  - based on its structure and on links between data and integrated elements of documentation

Quantitative



Check the completeness of records

Reduce burden at manual data entry

Minimise the number of steps

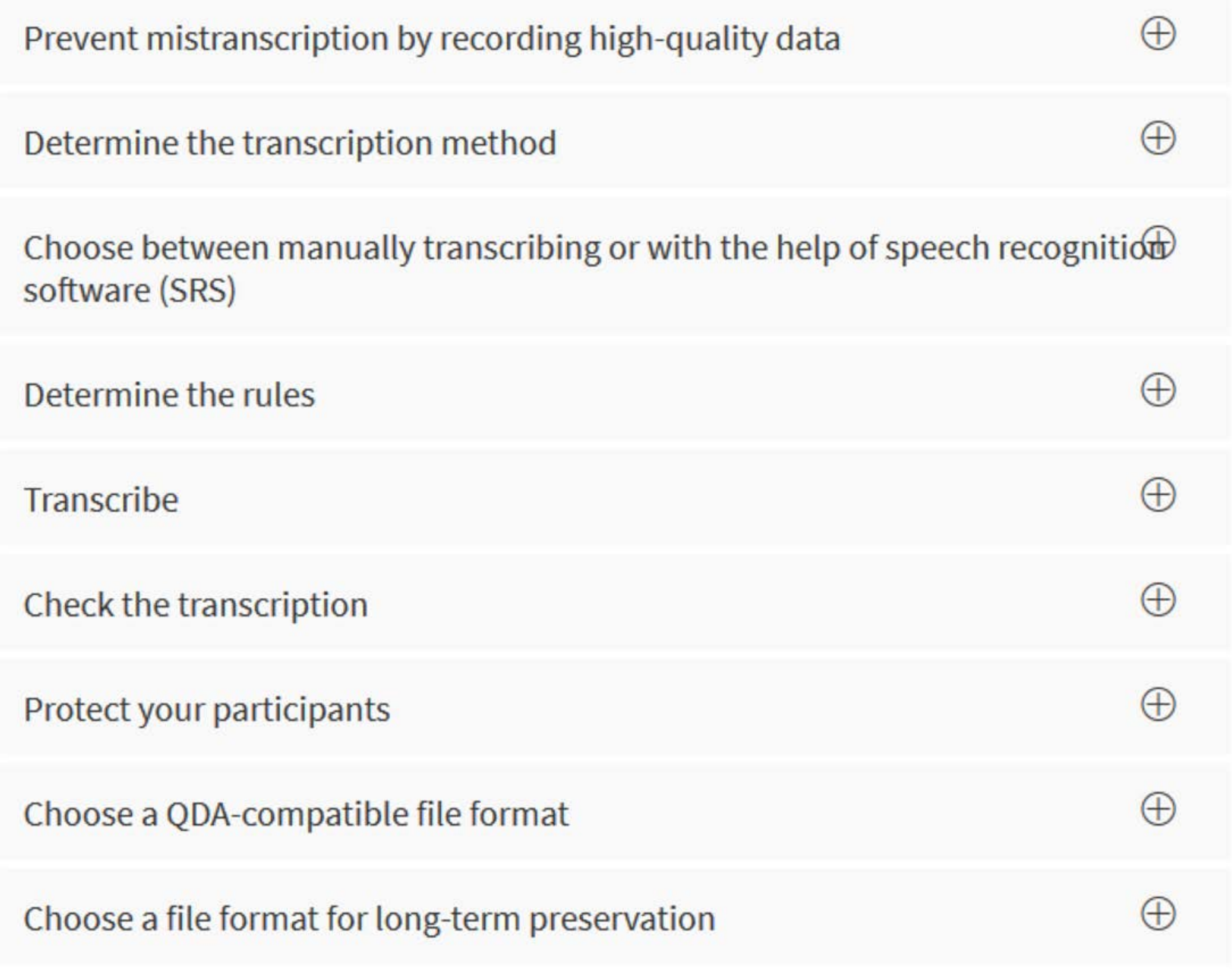
Conduct data entry twice

Perform in-depth checks for selected records

Perform logical and consistency checks

Automate checks whenever possible

Qualitative



Prevent mistranscription by recording high-quality data ⊕

Determine the transcription method ⊕

Choose between manually transcribing or with the help of speech recognition software (SRS) ⊕

Determine the rules ⊕

Transcribe ⊕

Check the transcription ⊕

Protect your participants ⊕

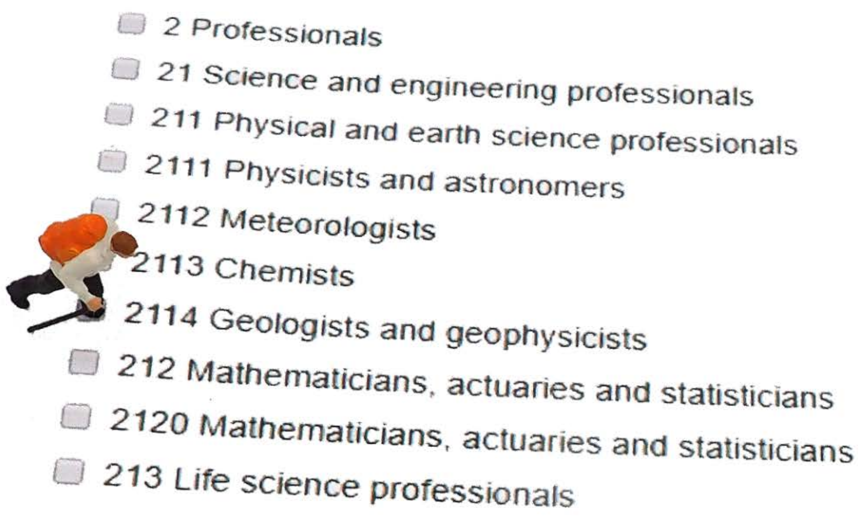
Choose a QDA-compatible file format ⊕

Choose a file format for long-term preservation ⊕

Quantitative coding/qualitative coding

- Quantitative:
  - General rules, recommendations/check lists
  - Documentation: subsection - organising variables (integrated doc./internal structure of the data file)
  - Standardised coding schemes
  - Missing values
  - Coding variance

International Labour Organization



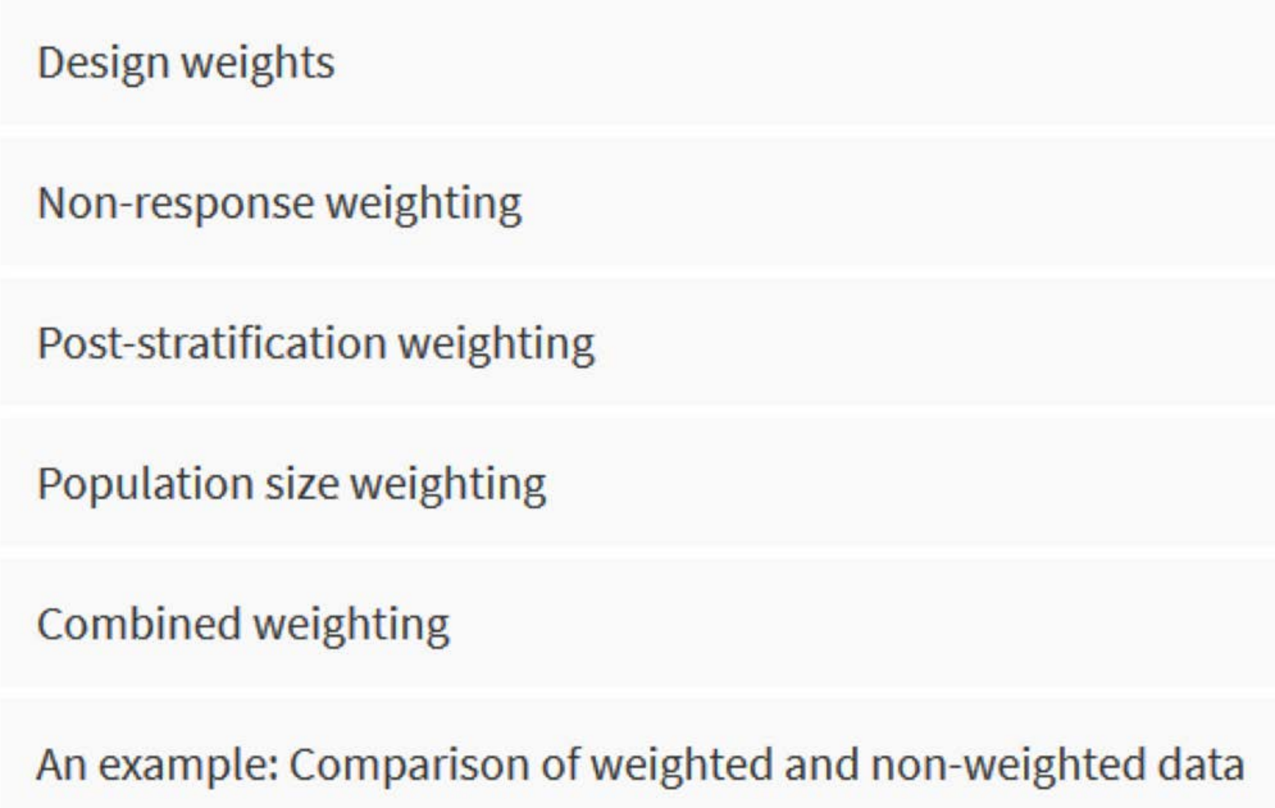
□ 2 Professionals
□ 21 Science and engineering professionals
□ 211 Physical and earth science professionals
□ 2111 Physicists and astronomers
□ 2112 Meteorologists
2113 Chemists
2114 Geologists and geophysicists
□ 212 Mathematicians, actuaries and statisticians
□ 2120 Mathematicians, actuaries and statisticians
□ 213 Life science professionals

- Qualitative:
  - *Coding is a way of indexing or categorizing the text in order to establish a framework of thematic ideas about it* | Gibbs (2007)
  - Concept driven coding versus data driven coding

Dive in deep? Weights of survey data

- Adjustment of the sample. Each individual case in the file is assigned an individual weight which is used to multiply the case in order to attain the desired characteristics of the sample.

- There are different types of weights for different purposes
- Necessary in some sitations
- Issue of quality



Design weights

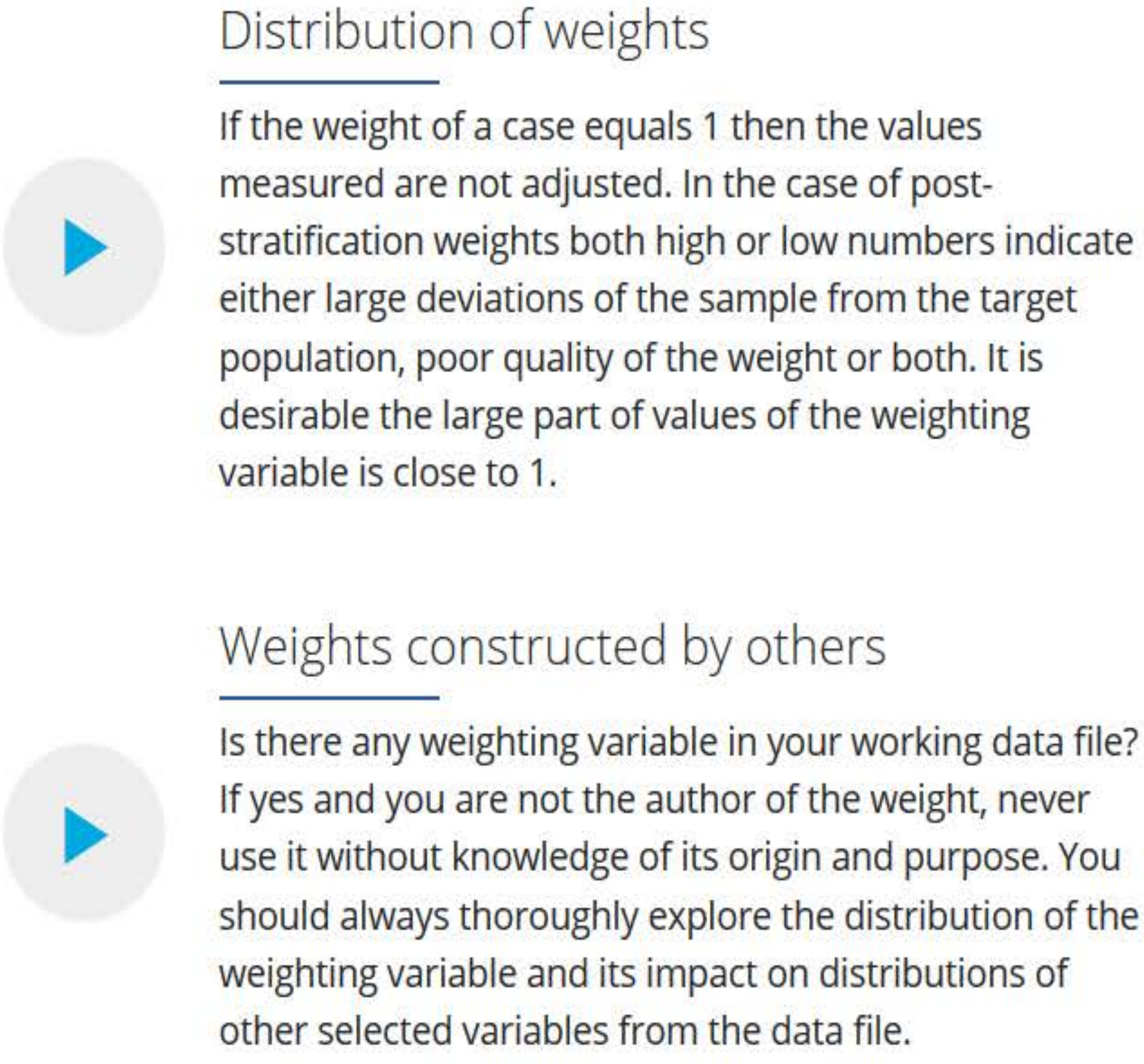
Non-response weighting

Post-stratification weighting

Population size weighting

Combined weighting

An example: Comparison of weighted and non-weighted data



Distribution of weights

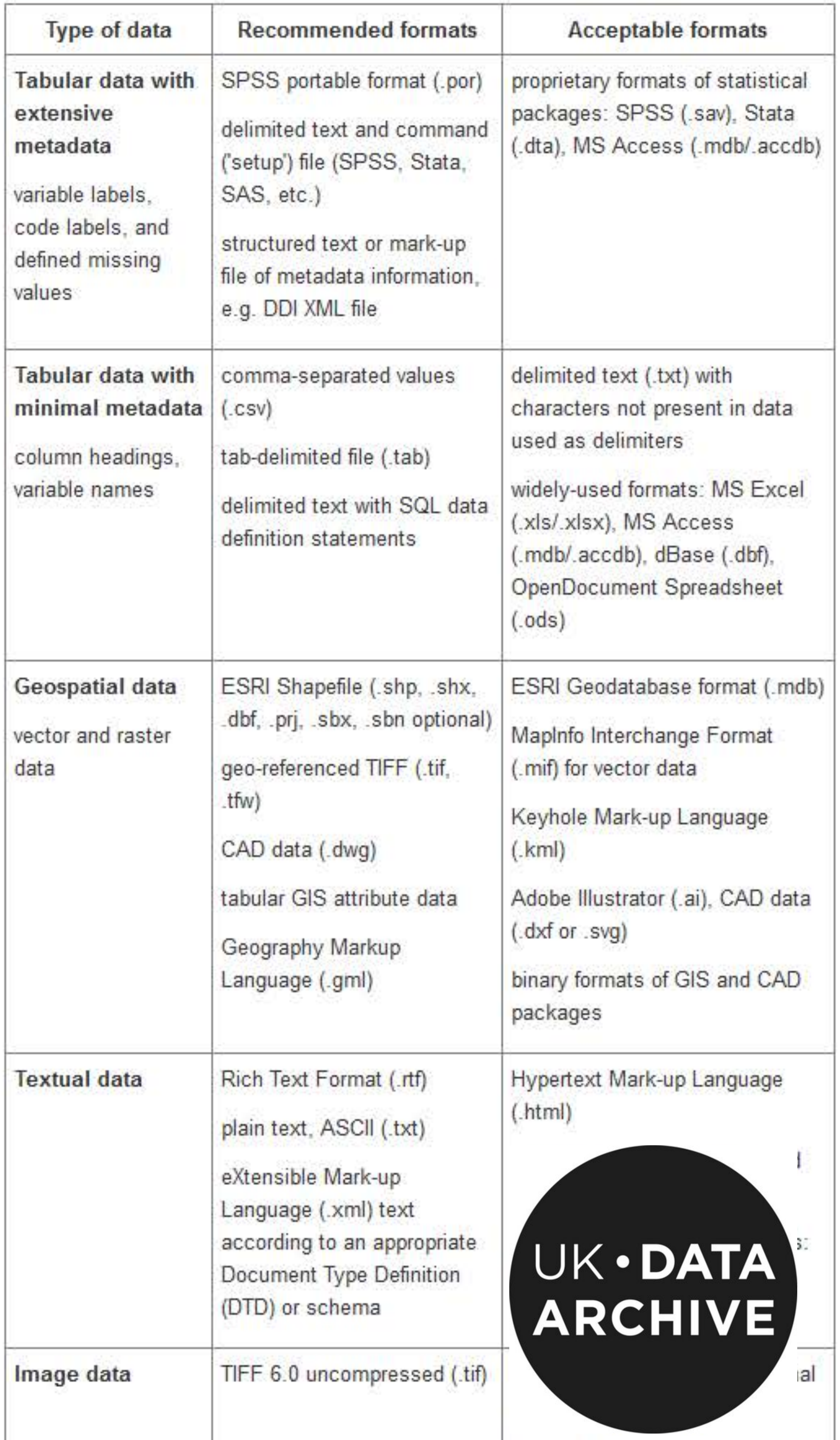
If the weight of a case equals 1 then the values measured are not adjusted. In the case of post-stratification weights both high or low numbers indicate either large deviations of the sample from the target population, poor quality of the weight or both. It is desirable the large part of values of the weighting variable is close to 1.

Weights constructed by others

Is there any weighting variable in your working data file? If yes and you are not the author of the weight, never use it without knowledge of its origin and purpose. You should always thoroughly explore the distribution of the weighting variable and its impact on distributions of other selected variables from the data file.

File formats and data conversion

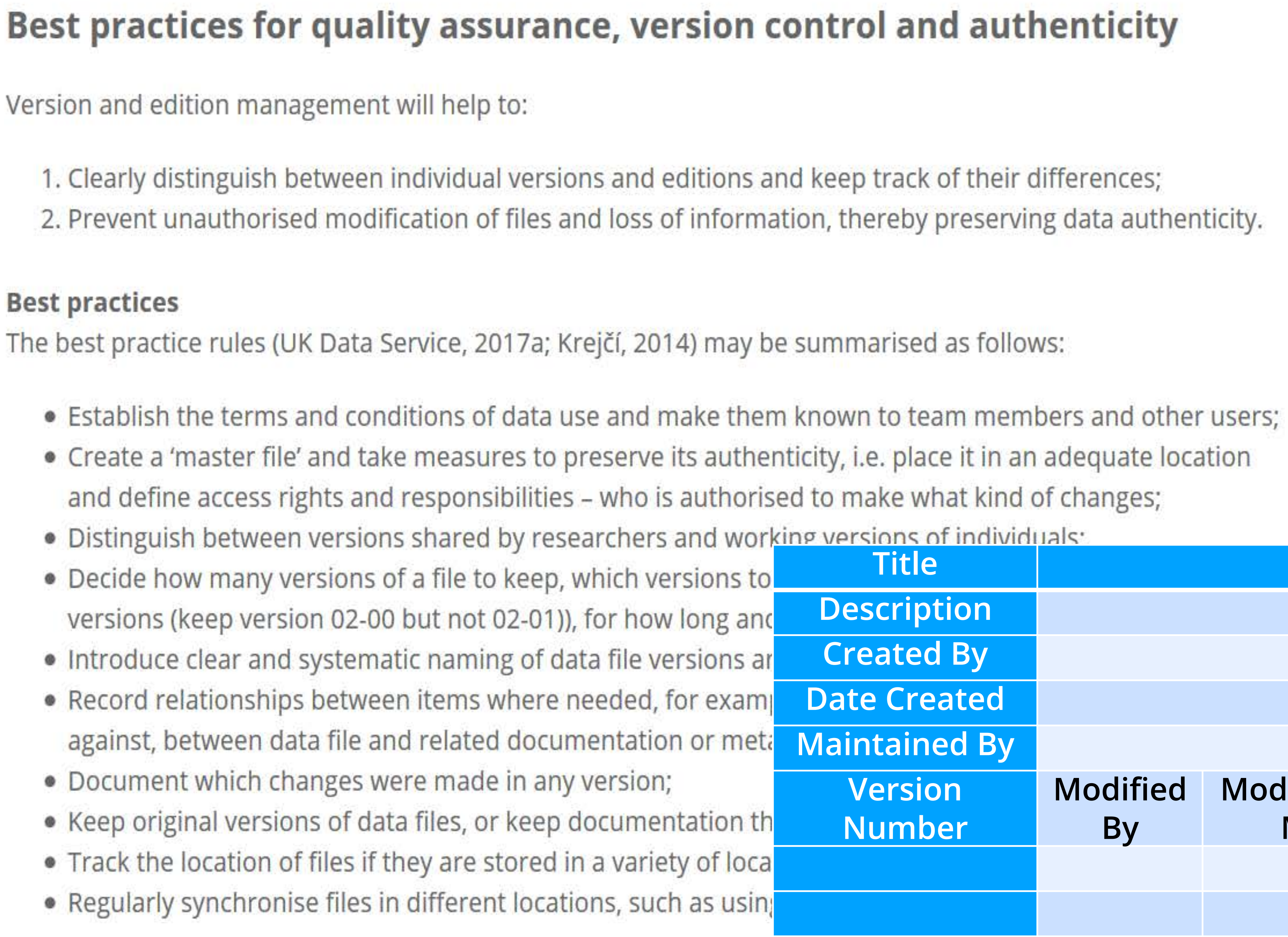
- Short-term data processing: file formats for operability
  - Proprietary vs. open formats
  - Export / portable formats
- Long-term data preservation
- Link to the table of Recommended file formats



Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) l s:
Image data	TIFF 6.0 uncompressed (.tif)	al

UK • DATA ARCHIVE

Data authenticity & version control



Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

1. Clearly distinguish between individual versions and editions and keep track of their differences;
2. Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices

The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- Establish the terms and conditions of data use and make them known to team members and other users;
- Create a 'master file' and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes;
- Distinguish between versions shared by researchers and working versions of individuals;
- Decide how many versions of a file to keep, which versions to
versions (keep version 02-00 but not 02-01)), for how long and
- Introduce clear and systematic naming of data file versions ar
- Record relationships between items where needed, for examp
against, between data file and related documentation or meta
- Document which changes were made in any version;
- Keep original versions of data files, or keep documentation th
- Track the location of files if they are stored in a variety of loca
- Regularly synchronise files in different locations, such as usin

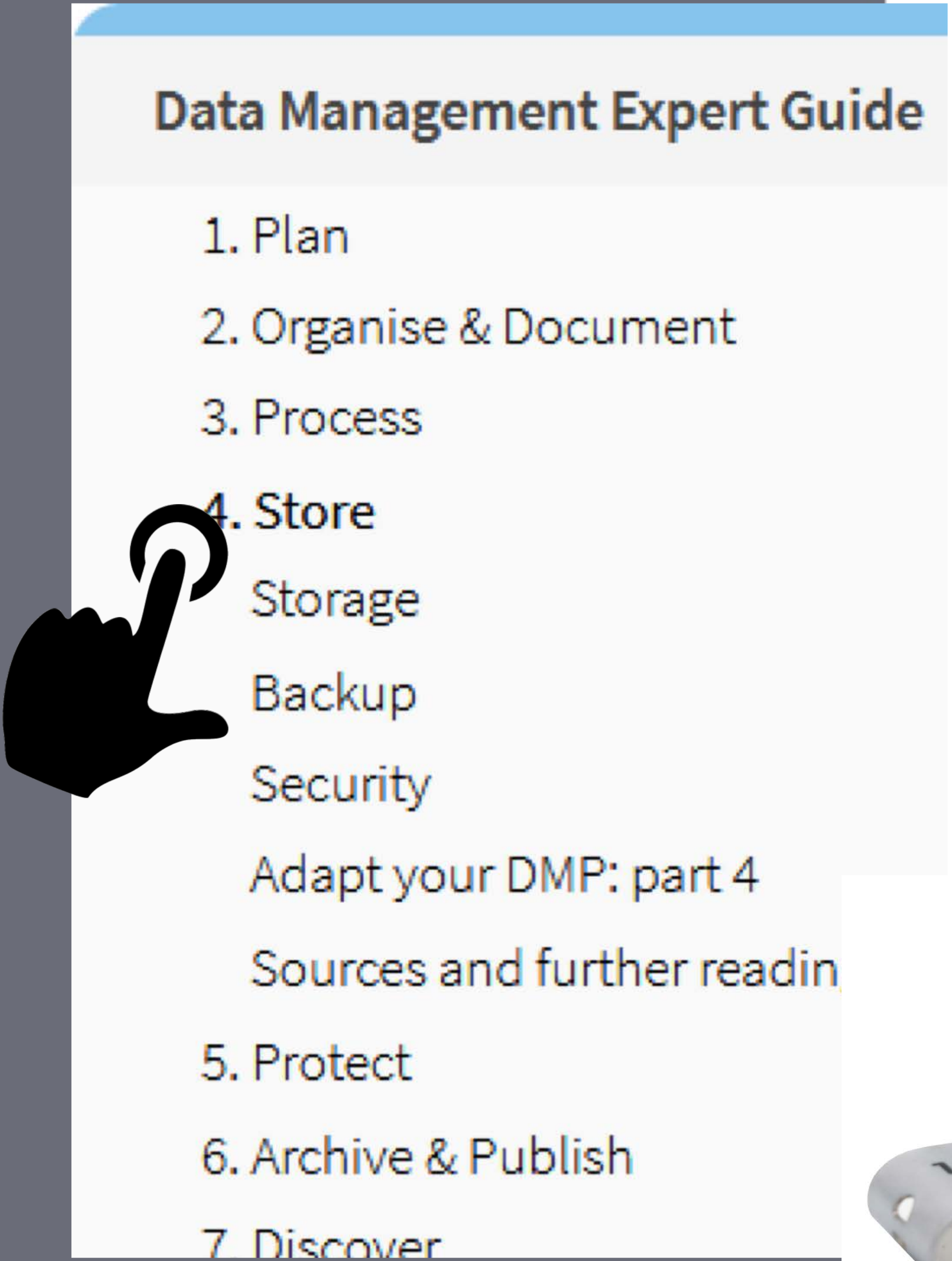
Title				
Description				
Created By				
Date Created				
Maintained By				
Version Number	Modified By	Modifications Made	Date Modified	Status

Wrap up: Data quality

- **Small things matter:** *"The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise."* American Association for Public Opinion Research (2015) (AAPOR)

- *"**In qualitative research**, discussions about quality in research are not so much based on the idea of standardization and control, as this seems incompatible with many qualitative methods. Quality is rather seen as an issue of how to manage it. Sometimes it is linked to rigour in applying a certain method, but more often to soundness of the research as a whole"* | Flick (2007).

- **A complex approach to data quality:** "*The mechanical quality control of survey operations such as coding and keying does not easily lend itself to continuous improvement. Rather, it must be complemented with feedback and learning where the survey workers themselves are part of an improvement process"* Biemer & Lyberg (2003).



Data Management Expert Guide

1. Plan

2. Organise & Document

3. Process

4. Store

Storage

Backup

Security

Adapt your DMP: part 4

Sources and further readin

5. Protect

6. Archive & Publish

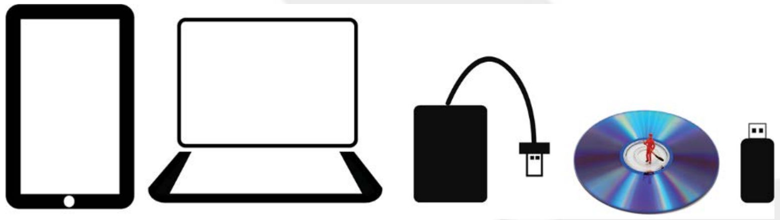
7. Discover

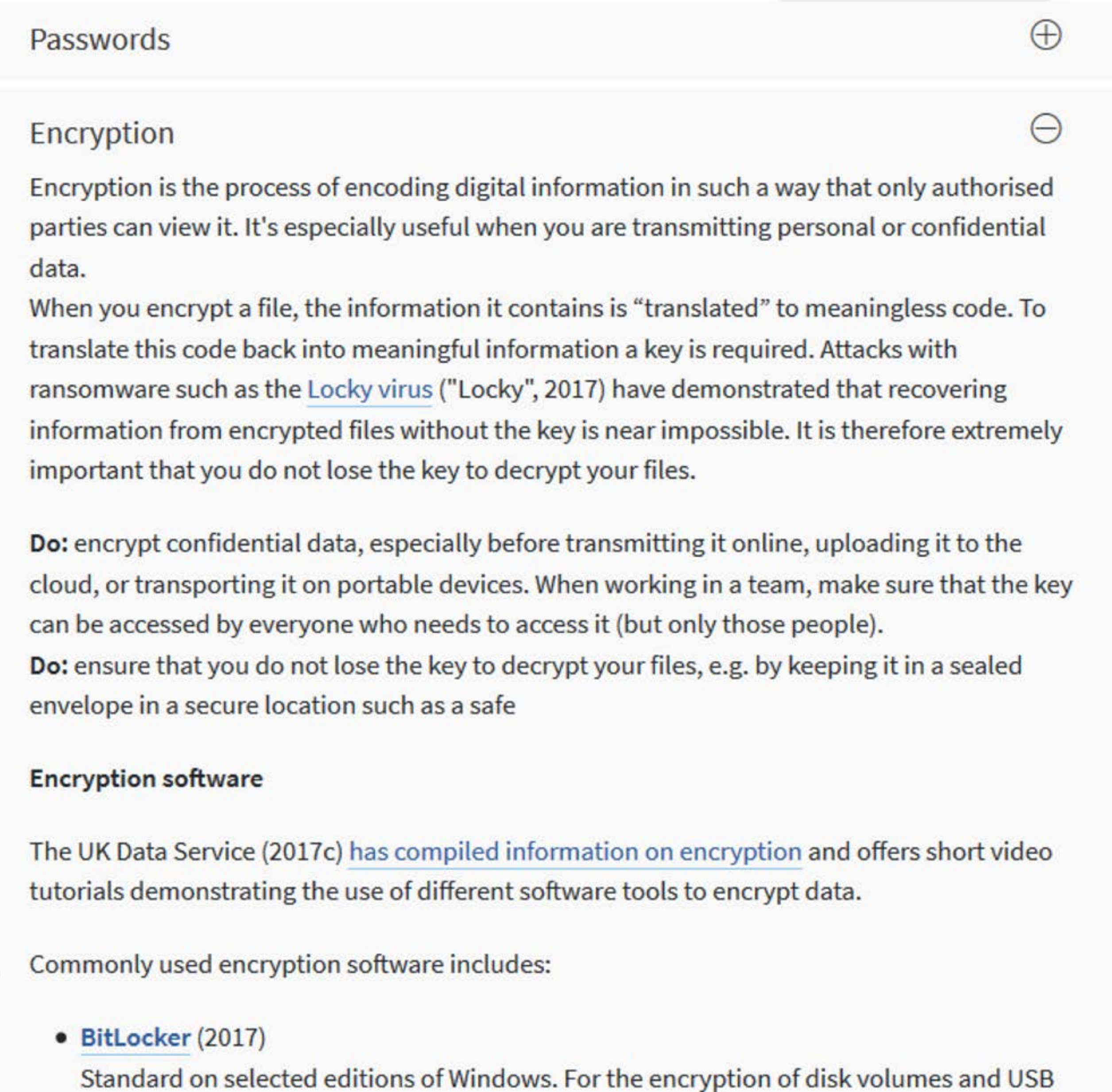
4. STORE

- Storage solutions
- Storage strategies
- Disaster recovery strategies
- Protect: passwords and encryption

Towards a Storage Strategy

- A Storage strategy contains
  - storage solutions and media
  - backup strategy and disaster recovery
  - data protection
- systematically implemented in a data management plan





Encryption is the process of encoding digital information in such a way that only authorised parties can view it. It's especially useful when you are transmitting personal or confidential data.

When you encrypt a file, the information it contains is "translated" to meaningless code. To translate this code back into meaningful information a key is required. Attacks with ransomware such as the Locky virus ("Locky", 2017) have demonstrated that recovering information from encrypted files without the key is near impossible. It is therefore extremely important that you do not lose the key to decrypt your files.

Do: encrypt confidential data, especially before transmitting it online, uploading it to the cloud, or transporting it on portable devices. When working in a team, make sure that the key can be accessed by everyone who needs to access it (but only those people).

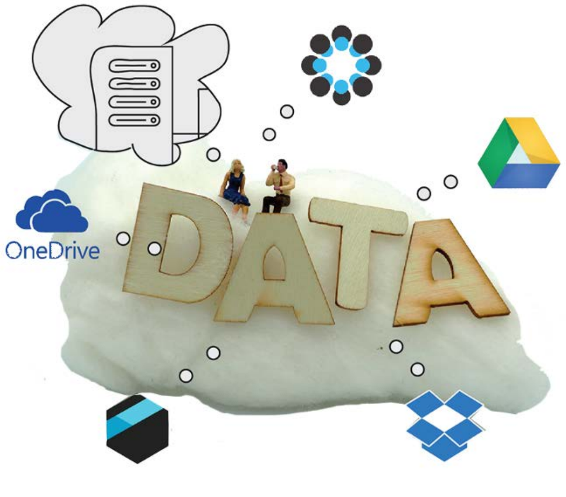
Do: ensure that you do not lose the key to decrypt your files, e.g. by keeping it in a sealed envelope in a secure location such as a safe

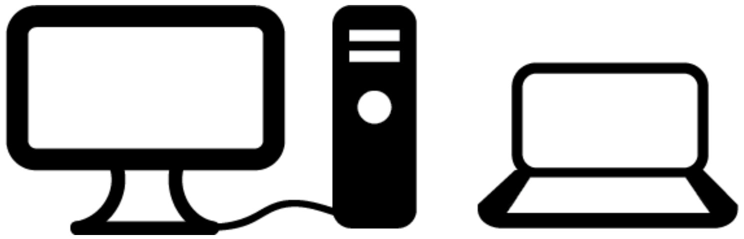
Encryption software

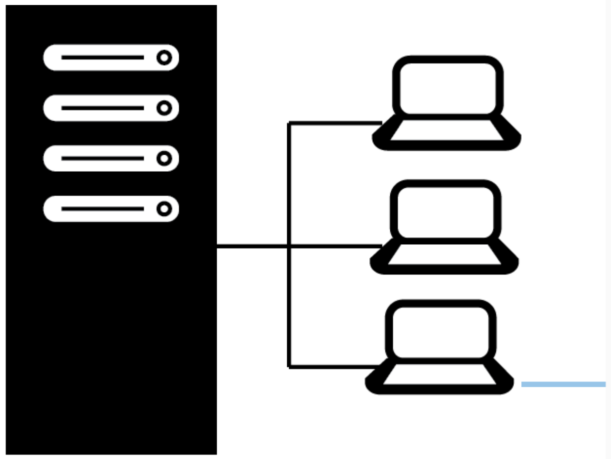
The UK Data Service (2017c) has compiled information on encryption and offers short video tutorials demonstrating the use of different software tools to encrypt data.

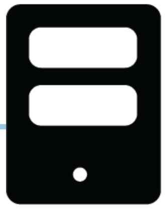
Commonly used encryption software includes:

- BitLocker (2017)
  Standard on selected editions of Windows. For the encryption of disk volumes and USB

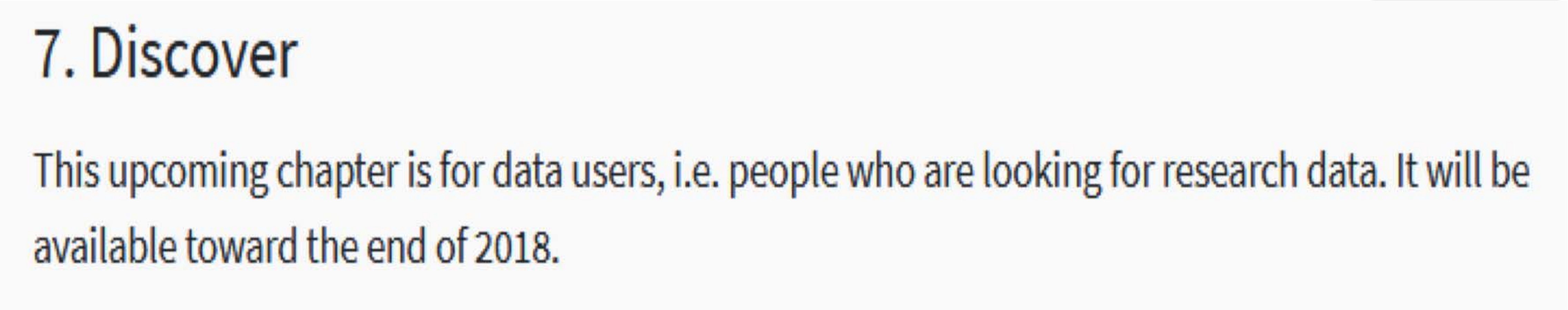








cessda



7. Discover

This upcoming chapter is for data users, i.e. people who are looking for research data. It will be available toward the end of 2018.

Main take-aways - after reading through this chapter you should:

- Be aware of different types of data resources for social sciences
- Know more about ways of searching for social science data
- Be able to use search engines in data repositories effectively
- Be aware of steps in evaluating the quality and usefulness of data for secondary analysis
- Understand different types and modes of access to data
- Be informed on research data relevant for selected research topics and recommended by experts.

Thank you

cessda

The Data Management Expert Guide has been created for CESSDA ERIC by a number of its service providers' experts at: ADP, AUSSDA, CSDA, DANS, FORS, FSD, GESIS, NSD, SND, So.Da.Net and UKDS and is illustrated and edited by Verbeeldingskr8.

Krejci, J. (2018). Data Management Expert Guide. Chapters 1, 2, 3 & 4 [presentation]. 2-nd CESSDA Widening Meeting 2018. Serbia, Belgrade, 14-15/11/2018.

cessda.eu @CESSDA_Data