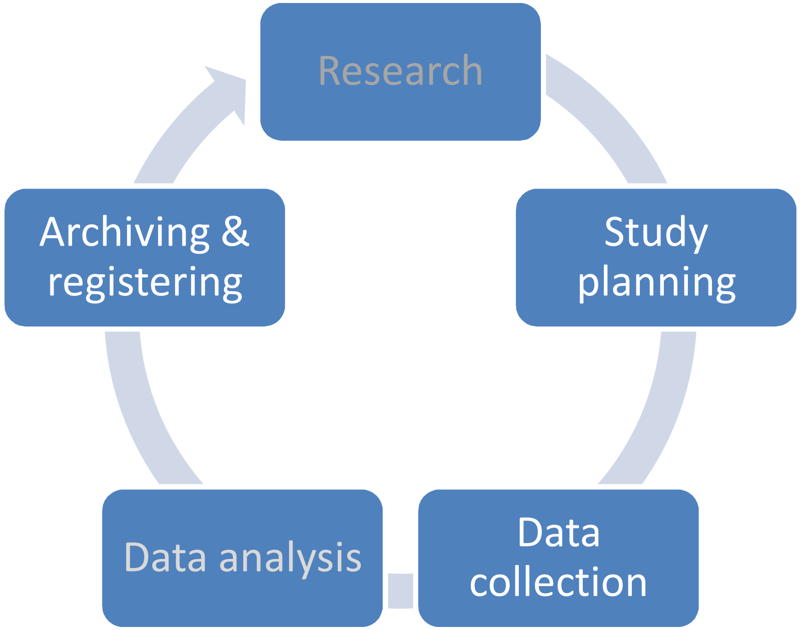
Data Handling:

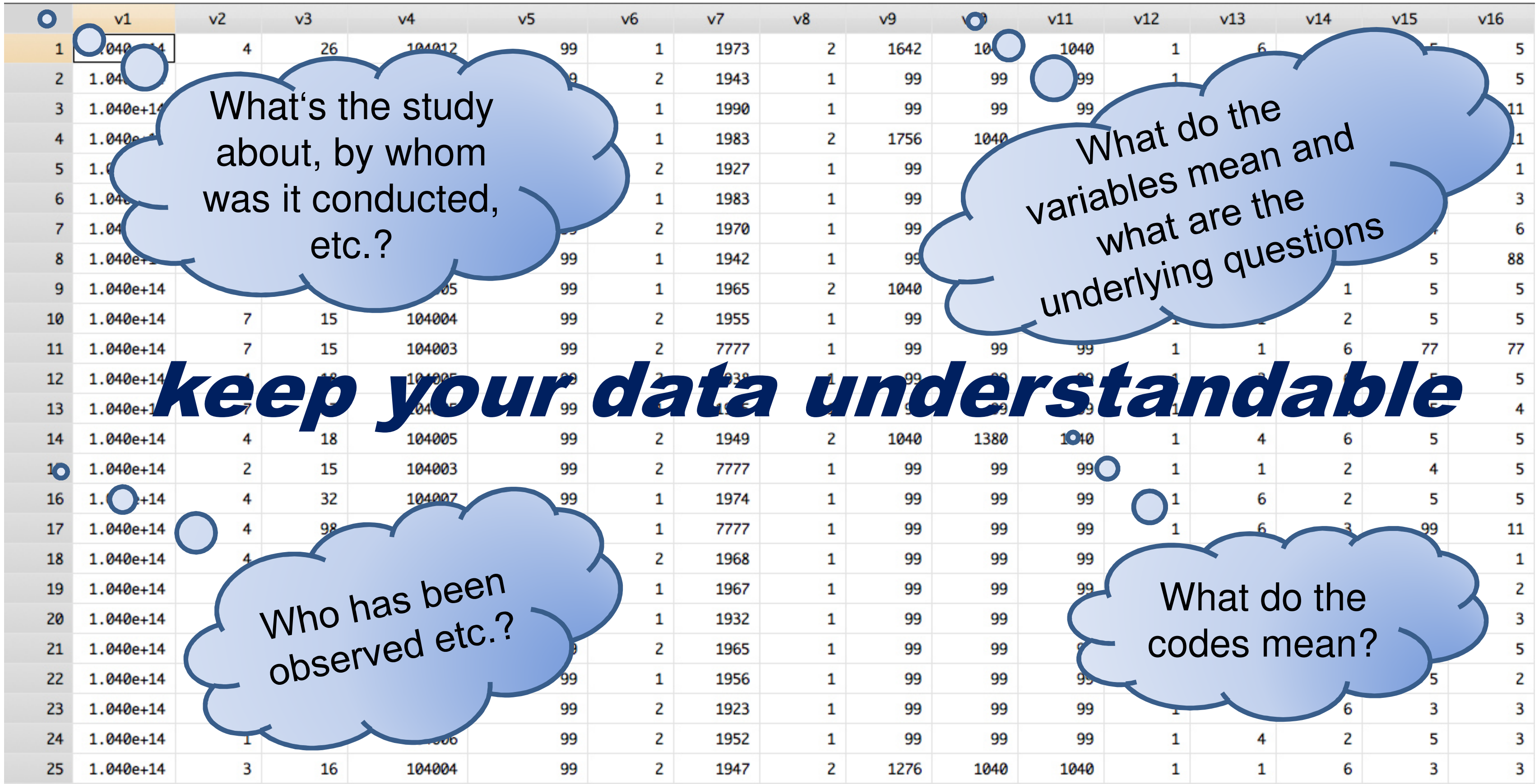
Documentation, Organization and Storage

Sebastian Netscher
CESSDA Training at the Data Archive for the Social Sciences
GESIS - Leibniz Institute for the Social Sciences
@CESSDA_Data

Data Documentation



Why Data Documentation?



	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	[illegible]	4	26	104012	99	1	1973	2	1642	[illegible]	1040	1	6	[illegible]	[illegible]	5
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	1943	1	99	99	99	1	[illegible]	[illegible]	[illegible]	5
3	1.040e+14	[illegible]	[illegible]	[illegible]	[illegible]	1	1990	1	99	99	99	[illegible]	[illegible]	[illegible]	[illegible]	11
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1983	2	1756	1040	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	1927	1	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1983	1	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2	1970	1	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6
8	[illegible]	[illegible]	[illegible]	[illegible]	99	1	1942	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5	88
9	1.040e+14	[illegible]	[illegible]	[illegible]	99	1	1965	2	1040	[illegible]	[illegible]	[illegible]	[illegible]	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	[illegible]	[illegible]	[illegible]	[illegible]	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5
13	[illegible]	[illegible]	[illegible]	[illegible]	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1380	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	[illegible]	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	[illegible]	[illegible]	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	[illegible]	[illegible]	[illegible]	2	1968	1	99	99	99	[illegible]	[illegible]	[illegible]	[illegible]	1
19	1.040e+14	[illegible]	[illegible]	[illegible]	[illegible]	1	1967	1	99	99	99	[illegible]	[illegible]	[illegible]	[illegible]	2
20	1.040e+14	[illegible]	[illegible]	[illegible]	[illegible]	1	1932	1	99	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3
21	1.040e+14	[illegible]	[illegible]	[illegible]	[illegible]	2	1965	1	99	99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5
22	1.040e+14	[illegible]	[illegible]	[illegible]	99	1	1956	1	99	99	[illegible]	[illegible]	[illegible]	[illegible]	5	2
23	1.040e+14	[illegible]	[illegible]	[illegible]	99	2	1923	1	99	99	99	[illegible]	[illegible]	6	3	3
24	1.040e+14	1	[illegible]	[illegible]	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3

What's the study about, by whom was it conducted, etc.?

What do the variables mean and what are the underlying questions

keep your data understandable

Who has been observed etc.?

What do the codes mean?

Comparative Study of Electoral Systems (CSES), www.cses.org

Levels of Data Documentation

- Study level
  - study description
  - study design
  - data processing
- Variable level
  - questionnaire
  - variables and codes

Image by A. Herrema & H. Bouwteam (CC-by)

Structured versus Unstructured Metadata

Unstructured documentation

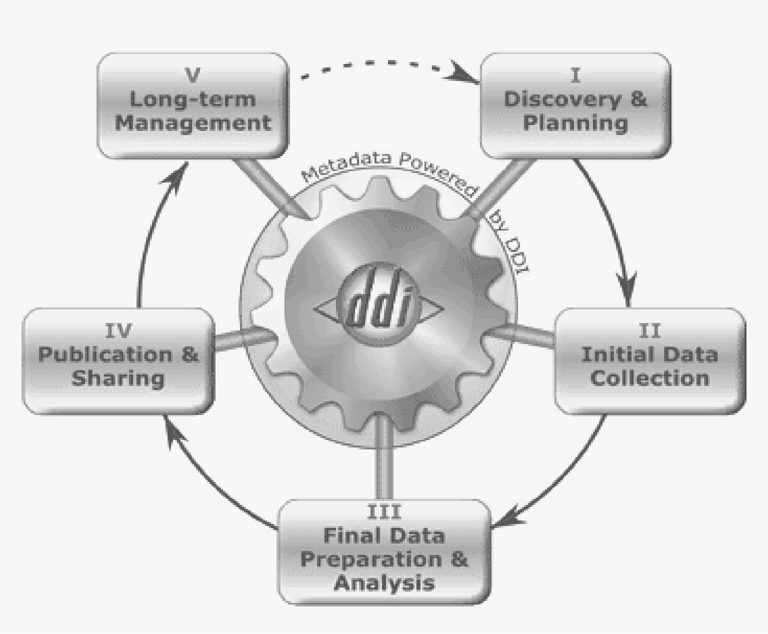
- technical reports etc.
- questionnaire, show cards, interviewer instructions etc.
- codebook etc.

Standardized forms
for standardized information

- coding schemas, e.g. ISCO, ISCED etc.
- international metadata standards, e.g. DDI

The Data Documentation Initiative (DDI)

- International standard for the description of data
  - DDI-Codebook (DDI2)
    ⇒ based on the codebook
  - DDI-Lifecycle (DDI3)
    ⇒ based on the (DDI) data lifecycle



Source: http://www.ddialliance.org/

Persistent Identifiers (PIDs)

- Persistent identifiers
  - provide permanency
  - assure unique retrieval of data
  - assign citation for reuse
- The DOI system
  - controlled by IDF (*International DOI Foundation*)
  - DOI Resolver, e.g. http://www.doi.org/index.html

DOI: 10.ORGANISATION/ID
Example: 10.4232/1.11159
prefix/suffix

da|ra Schema: Main Categories

Publication Date

Sampled Universe

Other Titles

Title

Creator

Description

Version

Temporal Coverage

Language

Classification Internal

Geographic Coverage

Collection Mode (free)

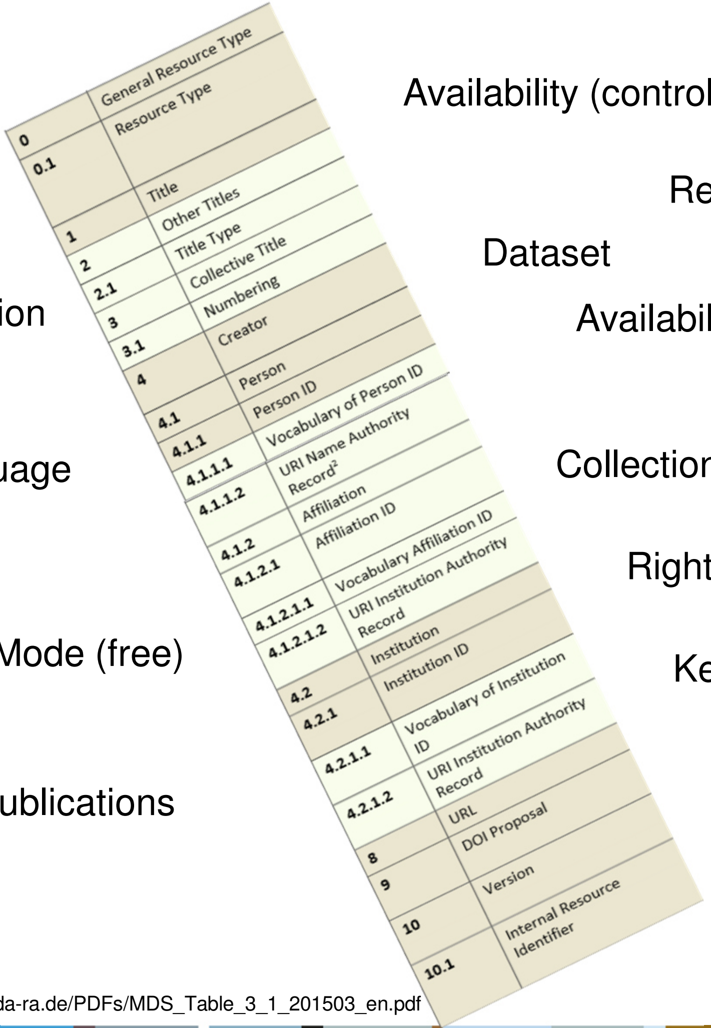
Collective Title

Time Dimension

Publications

DOI Proposal

General Resource Type



0	General Resource Type
0.1	Resource Type
1	Title
2	Other Titles
2.1	Title Type
3	Collective Title
3.1	Numbering
4	Creator
4.1	Person
4.1.1	Person ID
4.1.1.1	Vocabulary of Person ID
4.1.1.2	URI Name Authority Record[2]
4.1.2	Affiliation
4.1.2.1	Affiliation ID
4.1.2.1.1	Vocabulary Affiliation ID
4.1.2.1.2	URI Institution Authority Record
4.2	Institution
4.2.1	Institution ID
4.2.1.1	Vocabulary of Institution ID
4.2.1.2	URI Institution Authority Record
8	URL
9	DOI Proposal
10	Version
10.1	Internal Resource Identifier

Availability (controlled)

Notes

Relation

Contributor

Dataset

Availability (free)

Alternative Identifier

Collection Mode (controlled)

Rights

URL

Keywords (controlled)

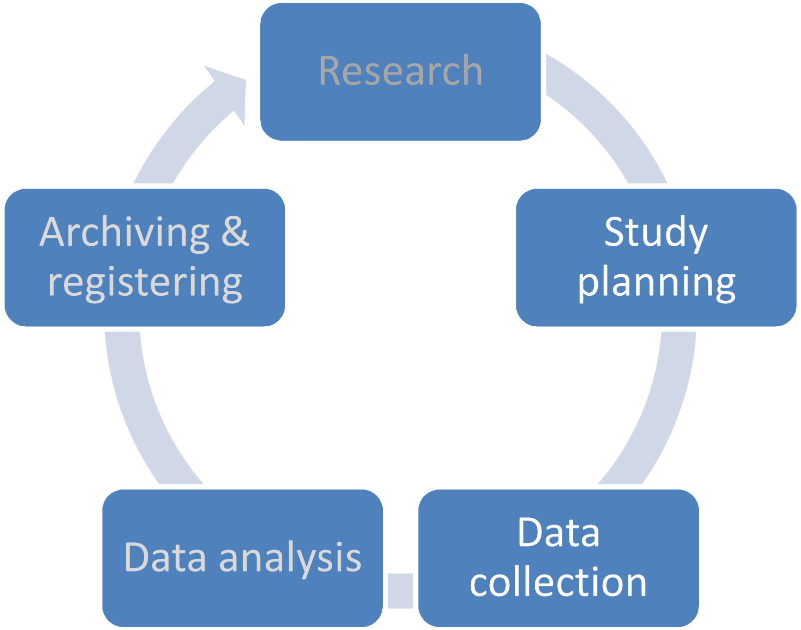
Publication Place

Sampling

Keywords (free)

Source: http://www.da-ra.de/fileadmin/media/da-ra.de/PDFs/MDS_Table_3_1_201503_en.pdf

Organizing Folders and Files



Structuring Folders

- Systematically managing folders
  - saves time and effort
  - simplifies the use (collaborative projects)
  - protects your folders and files from accidental clean-up
- Hierarchical structure of folders
  - structure by topic, data type etc.
- Develop standards early in the project

⇒ use these standards consistently within a project

File Names and Versions

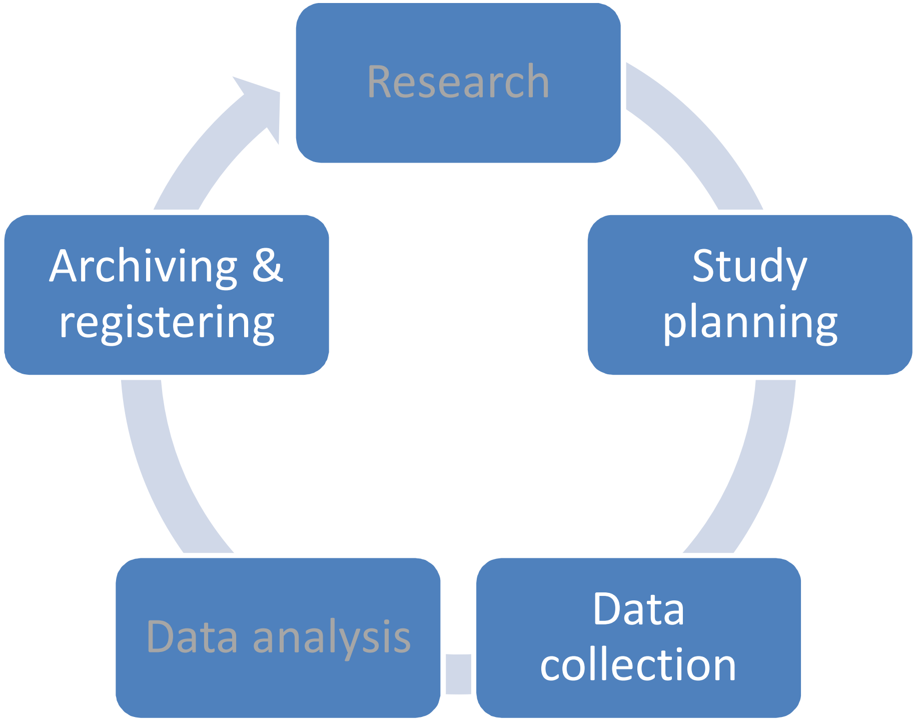
- File names
  - can contain various information, e.g. title of project, editor‘s name, date of creation, version etc.
  - neither include punctuation characters or blanks nor be too long
- File versioning
  - as a part of the file names, e.g. including the date or numbering the files
  - included in the header of the file
  - in a separate log-file

(Recommended) File Formats

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.)	SPSS (.sav); Stata (.dta); MS Access (.mdb/.accdb)
Tabular data with minimal metadata	comma-separated (.csv); tab-delimited file (.tab)	MS Excel (.xls/.xlsx); MS Access (.mdb/.accdb), dBase (.dbf); OpenDocument (.ods)
Textual data	Rich Text Format (.rtf); plain text, ASCII (.txt)	HTML(.html); MS Word (.doc/.docx); software-specific formats, e.g. NUD*IST or NVivo
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg); RAW image format (.raw), Photoshop (.psd); PDF/A or PDF (.pdf)
Audio data	Free Lossless Audio Codec (.flac)	MPEG-1 (.mp3); Waveform (.wav)
Video data	MPEG-4 (.mp4); JPEG 2000 (.mj2)	
Documentation and scripts	Rich Text Format (.rtf); PDF/A or PDF (.pdf); HTML (.htm); OpenDocument (.odt)	plain text (.txt); MS Word (.doc/.docx), MS Excel (.xls/.xlsx); XML (.xml)

Source: UK DATA Service, http://ukdataservice.ac.uk/manage-data/format/recommended-formats

Data Storage and Security



Back-Up

- Digital media are fallible
- A back-up is an additional copy that can be used to restore originals
- Backing-up implies having a back-up strategy

Image by A. Herrema & H. Bouwteam (CC-by)

Towards a Back-up Strategy

- A systematic back-up strategy defines
  a) what ⇒ all, some, just changes …
  b) where ⇒ external, local, remote copies …
  c) how often ⇒ at least in triplicates
  d) for how long ⇒ how long are things needed
  e) responsibility ⇒ automate the back-up process
- Verify and recover your back-ups
  ⇒ never assume, regularly test a restore
- Treat back-ups the same as the original files

Data Protection

- Protect your data from unauthorized access, use, change, disclosure, destruction etc.
- Take care of personal data
  - data protection legislation (EU Directive 1995/46/EC)
  - separate personal data from other data
- Use passwords and encryption

Image by P. Hochstenbach (CC-by)

Passwords

- A strong password has
  - eight to fifteen characters or even more
  - a random distribution of characters
- Combine…

…	upper case letters:	A - Z
…	lower case letters:	a - z
…	numerals:	0 - 9
…	special characters:	! " # $ % & ' ( ) * + , - . / : *etc.*

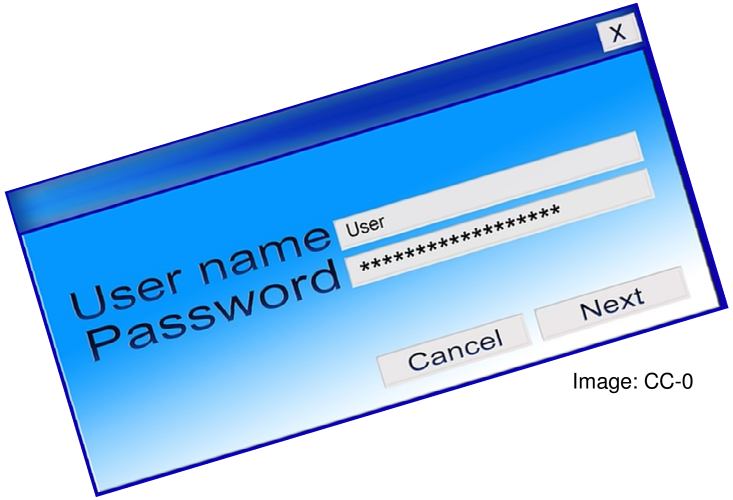
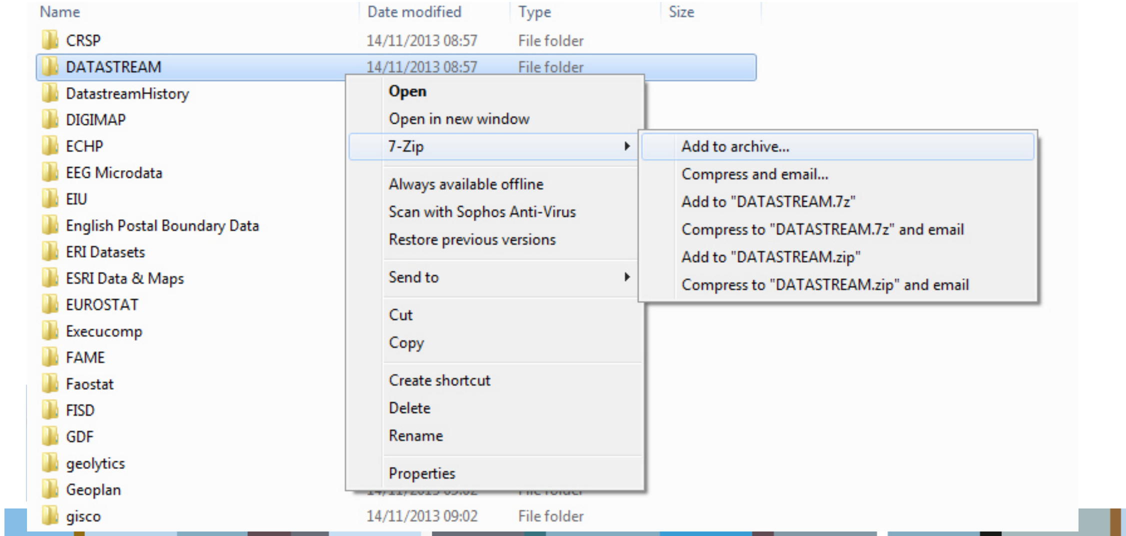


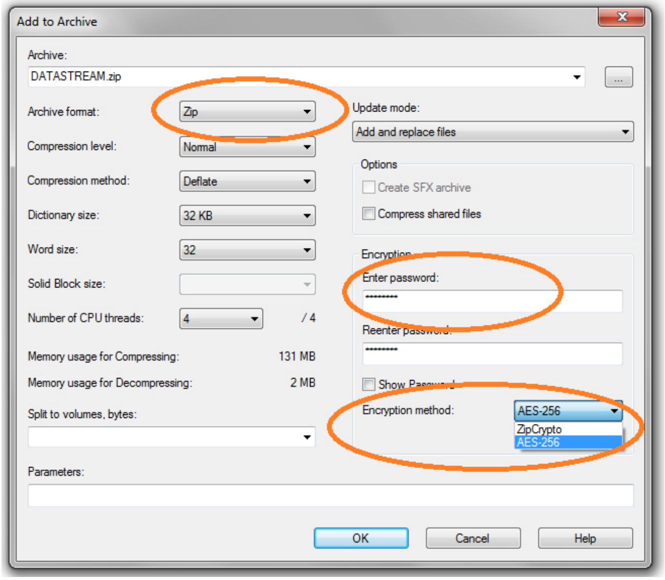
Image: CC-0

Encryption

- Helps maintain the security of data and documentation
  - uses an algorithm to transform information
  - requires a "key" to decrypt
- For example, encrypt ZIP files securely using 7Zip



Name	Date modified	Type	Size
CRSP	14/11/2013 08:57	File folder	
DATASTREAM	14/11/2013 08:57	File folder	
DatastreamHistory			
DIGIMAP			
ECHP			
EEG Microdata			
EIU			
English Postal Boundary Data			
ERI Datasets			
ESRI Data & Maps			
EUROSTAT			
Execucomp			
FAME			
Faostat			
FISD			
GDF			
geolytics			
Geoplan			
gisco	14/11/2013 09:02	File folder	



Add to Archive

Archive: DATASTREAM.zip

Archive format: Zip

Compression level: Normal

Compression method: Deflate

Dictionary size: 32 KB

Word size: 32

Solid Block size:

Number of CPU threads: 4 / 4

Memory usage for Compressing: 131 MB

Memory usage for Decompressing: 2 MB

Split to volumes, bytes:

Parameters:

Update mode: Add and replace files

Options

☐ Create SFX archive

☐ Compress shared files

Encryption

Enter password: *********

Reenter password: *********

☐ Show Password

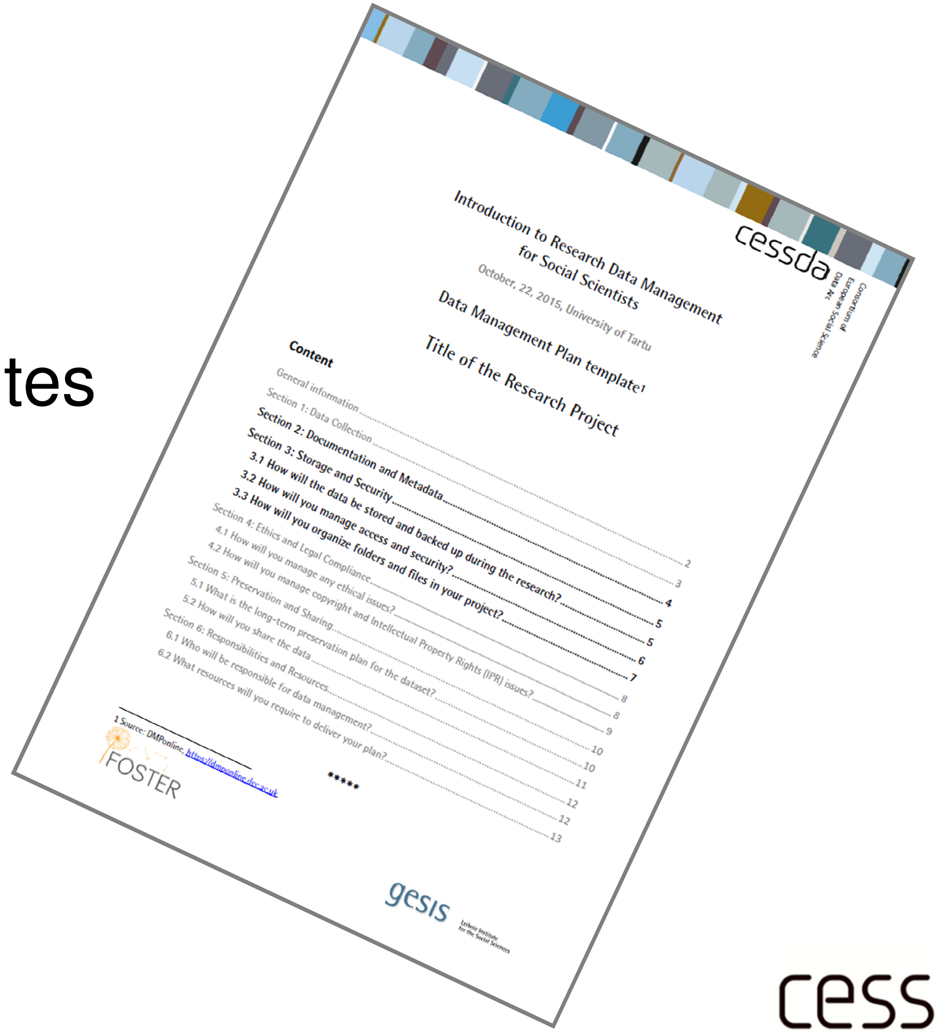
Encryption method: AES-256

Further Readings

- Aryal, M. (ed.) (2012): Speak Safe. Media Workers' Toolkit for Safer Online and Mobile Practices. https://www.internews.org/sites/default/files/ resources/Internews_SpeakSafeToolkit.pdf
- Borgmann, M., Hahn, T., Herfert, M., Kunz, T., Richter, M., Viebeg, U., and Vowé, S. (2012): On the Security of Cloud Storage Services. Frauenhofer Institut, SIT Technical Report. https://www.sit.fraunhofer.de/fileadmin/dokumente/studien_und_technical_reports/Cloud-Storage-Security_a4.pdf.
- Directive 95/46/EC of the European Parliament and of the Council, 24 October 1995. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do? uri=CELEX:31995L0046:EN:NOT
- Gregory, A., Heus, P., & Ryssevik, J., 2009, Metadata. Berlin. http://www.ratswd.de/download/workingpapers2009/57_09.pdf.
- Miller, K., & Vardigan, M., 2005, How Initiative Benefits the Research Community - the Data Documentation Initiative. In First International Conference on e-Social Science, Manchester, UK, June 2005. http://www.ddialliance.org/sites/default/files/miller.pdf.
- National Information Standards Organization, 2004, Understanding Metadata (p. 17). Bethesa, MD: NISO Press. www.niso.org/standards/resources/UnderstandingMetadata.pdf.
- Plant, R. R., 2012, How to add metadata to your data so that you and others can make sense of it. Retrieved from http://www.shef.ac.uk/polopoly_fs/1.158828!/file/Metadatav6.pdf.
- Starr, J., 2011, DataCite Metadata Schema for the Publication and Citation of Research Data (p. 29). doi:10.5438/0005
- Vardigan, M., Heus, P., & Thomas, W., 2008, Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation, 3(1), 107–113. doi:10.2218/ijdc.v3i1.45.

DMP Sections 2 and 3

- work in 3 groups,
- time: about 30 minutes
- ☑ choose one of the following topics



Introduction to Research Data Management
for Social Scientists

October, 22, 2015, University of Tartu

Data Management Plan template[1]

Title of the Research Project

Content

*****

[1] Source: DMPonline, https://dmponline.dcc.ac.uk

cessda
Consortium of European Social Science Data Arc
FOSTER
gesis Leibniz Institute for the Social Sciences

DMP Sections 2 & 3

a) documentation (*Section 2*), considering …

… what information is needed

… how you (will) capture this information

b) data storage and back-ups (*Section 3.1*), developing a back-up strategy, i.e. …

… what, where and how often / long it is backed-up

… how are back-ups verified

c) managing folders and files (*Section 3.3*), considering …

… how you will organize your folders

… how you will name and version your files

DMP Section 2: Documentation

- study description
  - study‘s aim, primary researcher (and funders), population and sampling procedures, method of data collection, data cleaning and anonymization etc.
    - ⇒ technical and methodological report
- variable description
  - questionnaire: original question wording and provided answer categories, explanations, interviewer instructions;
  - variables: labels and meanings of variables and codes, variable notes, scales etc.
    - ⇒ codebook, questionnaire and labels in the dataset

DMP Section 3.1: Back-up

- developing a back-up strategy

  ⇒ defining clear and consistent guidelines

  - what: all, something, only changed files
  - where: at least in triplicates and different locations
  - how long are different files (and versions) needed
    never destruct or overwrite original data
  - who: name researcher(s) and assign responsibilities

  ⇒ verify back-ups frequently (e.g. once a week),
  e.g. restoring the files (name researcher(s) and responsibilities)

DMP Sections 3.3: Organizing …

- developing guidelines to organize …

  … folders

  – define a consistent structure of folders

  ⇒ e.g. by topic

  … and files, i.e. define a consistent strategy

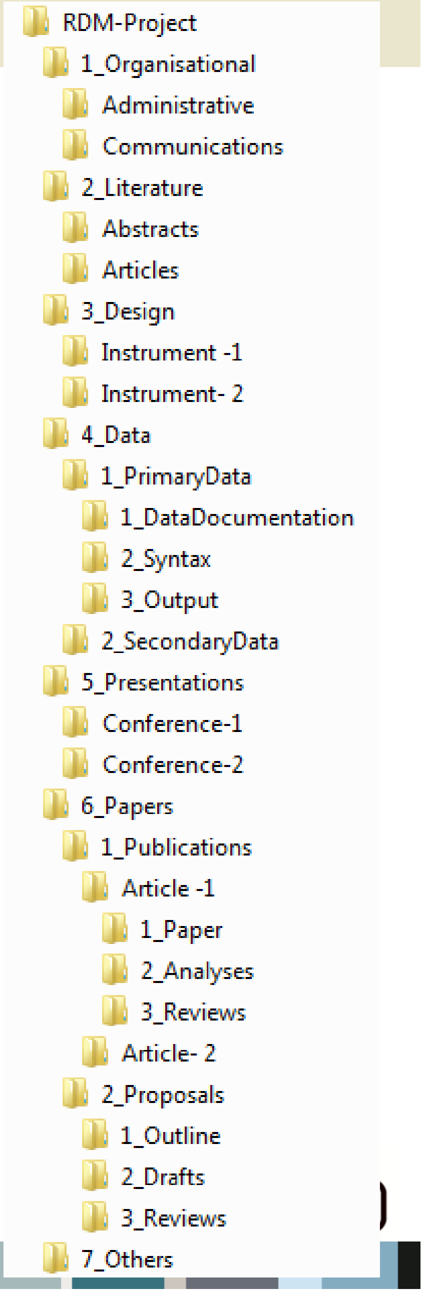
  – to name files

  ⇒ e.g. [type_name_version]

  – to version files

  ⇒ e.g. by the date and editor’s acronyms

  – *data_RDMData_20150822sn*



- RDM-Project
  - 1_Organisational
    - Administrative
    - Communications
  - 2_Literature
    - Abstracts
    - Articles
  - 3_Design
    - Instrument -1
    - Instrument- 2
  - 4_Data
    - 1_PrimaryData
      - 1_DataDocumentation
      - 2_Syntax
      - 3_Output
    - 2_SecondaryData
  - 5_Presentations
    - Conference-1
    - Conference-2
  - 6_Papers
    - 1_Publications
      - Article -1
        - 1_Paper
        - 2_Analyses
        - 3_Reviews
      - Article- 2
    - 2_Proposals
      - 1_Outline
      - 2_Drafts
      - 3_Reviews
  - 7_Others