



# Data Handling: Documentation, Organization and Storage

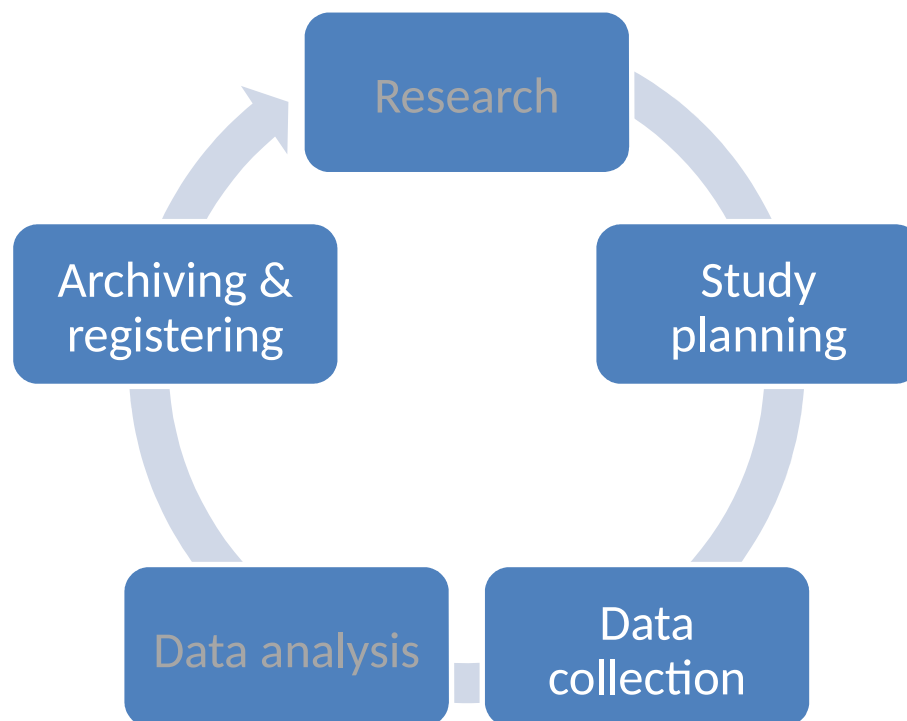
Sebastian Netscher

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	1.040e+14	4	26	104012	99	1	1973	2	1642	10	1040	1	6	1	5	5
2	1.040e+14			104008	99	2	1943	1	99	99	99		1	6	4	5
3	1.040e+14			104007	99	1	1990	1	99	99			1	3	99	11
4	1.040e+14			104007	99	1	1983	2	1756	1			1	3	99	11
5	1.040e+14			104007	99	2	1927	1	99				1	6	4	1
6	1.040e+14			104005	99	1	1983	1	99				1	2	5	3
7	1.040e+14			104005	99	2	1970	1	99				1	2	4	6
8	1.040e+14			104005	99	1	1942	1	99				1	6	5	88
9	1.040e+14			104005	99	1	1965	2	1040				1	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99				1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99			1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	1	104007	99	1	1949	1	99	99	99	1	1	6	5	5
15	1.040e+14	2	15	104007	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4		104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14			104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14			104007	99	1	1932	1	99	99	99	1	6	1	4	3
21	1.040e+14			104011	99	2	1965	1	99	99	99	1	1	5	5	5
22	1.040e+14			104011	99	1	1956	1	99	99	99	1	1	5	2	2
23	1.040e+14			104004	99	2	1923	1	99	99	99	1	1	3	3	3
24	1.040e+14			104006	99	2	1952	1	99	99	99	1	1	5	3	3
25	1.040e+14			104004	99	2	1947	2	1276	1040	1040	1	6	3	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 study understandable

Who has been observed etc.?

What do the codes mean?

Comparative Study of Electoral Systems (CSES), [www.cses.org](http://www.cses.org).



# Levels of Data Documentation

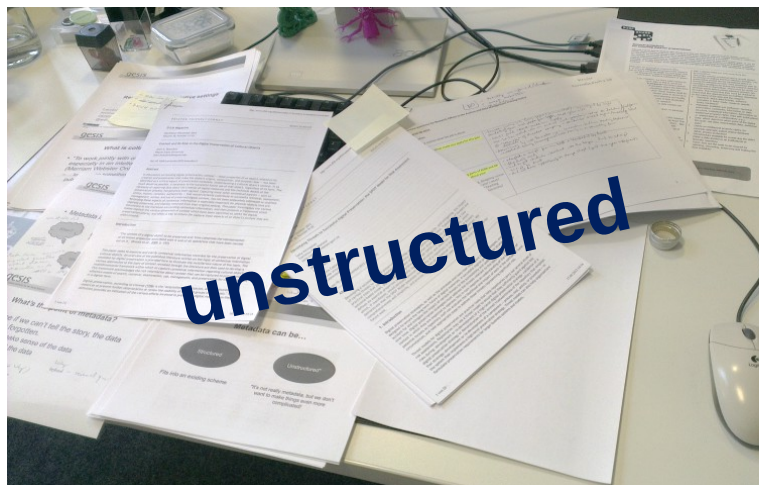
- Study level
  - study description
  - study design
  - data processing
- Variable level
  - questionnaire
  - variables and codes



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 documentation

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



# The Data Documentation Initiative (DDI)

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



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





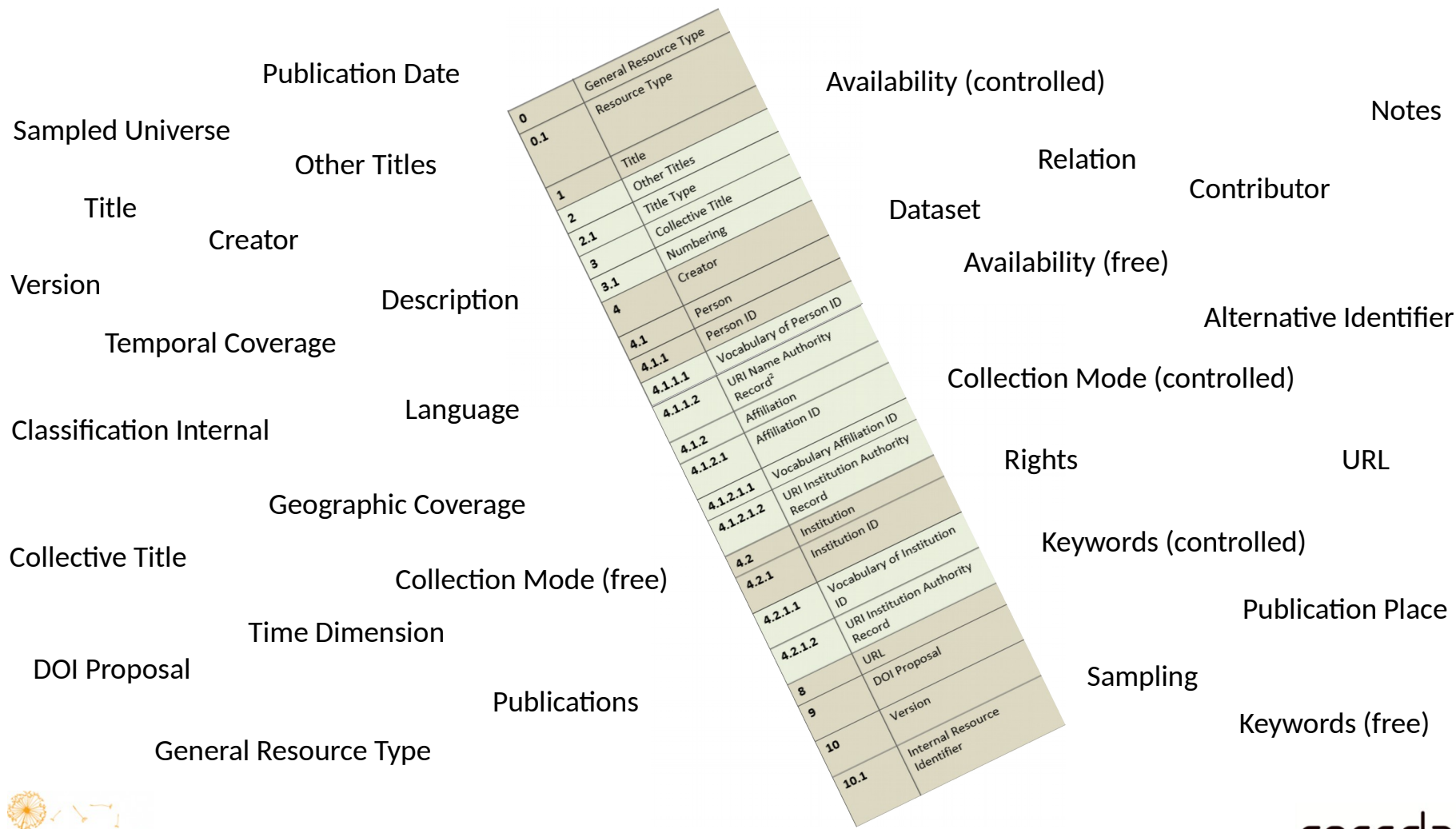
# Persistent Identifiers (PIDs)

- 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



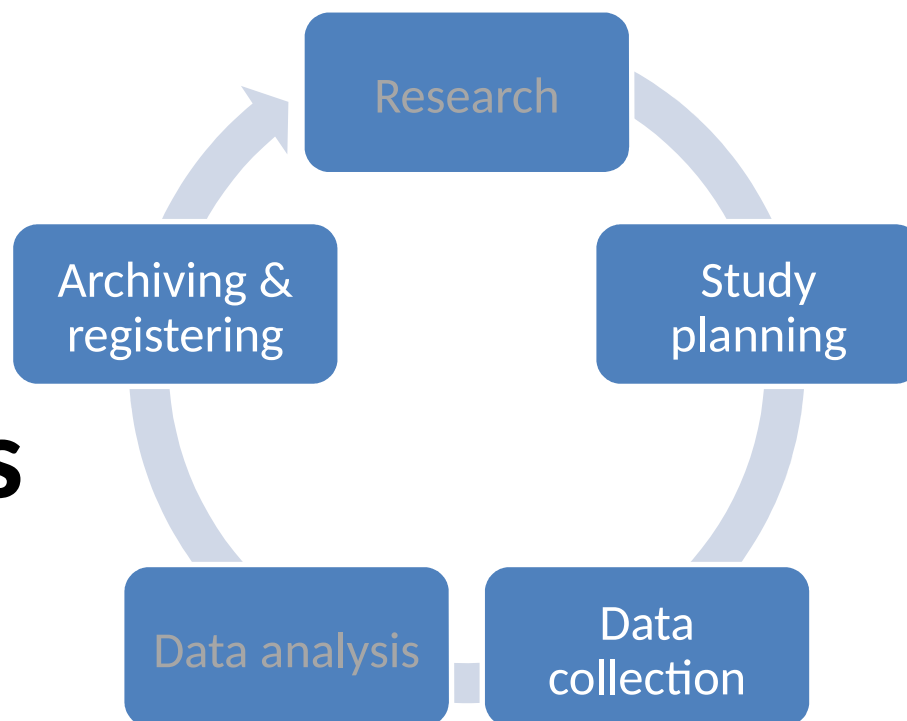
# da|ra Schema: Main Categories







# Organizing Folders and Files





# Structuring Folders

- Systematically managing folders and files
  - 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



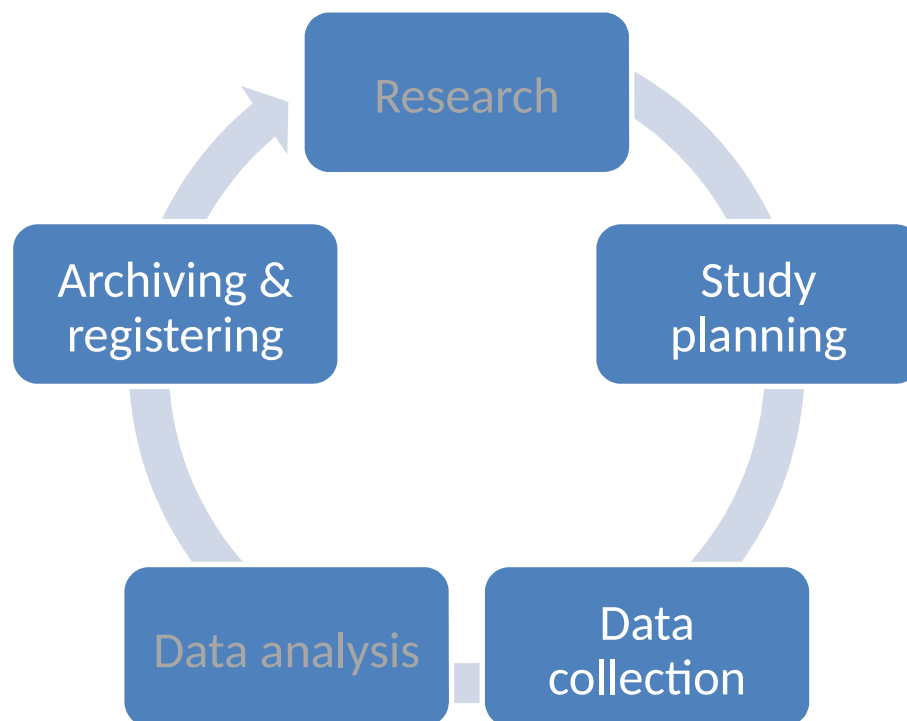
# (Recommended) File Formats

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
  - what  $\Rightarrow$  all, some, just changes ...
  - where  $\Rightarrow$  external, local, remote copies ...
  - how often  $\Rightarrow$  at least in triplicates
  - for how long  $\Rightarrow$  how long are things needed
  - who is in charge  $\Rightarrow$  automate the back-up process
- Verify and recover your back-ups
  - $\Rightarrow$  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)

cc-by <http://hochstenbach.wordpress.com>



# 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.

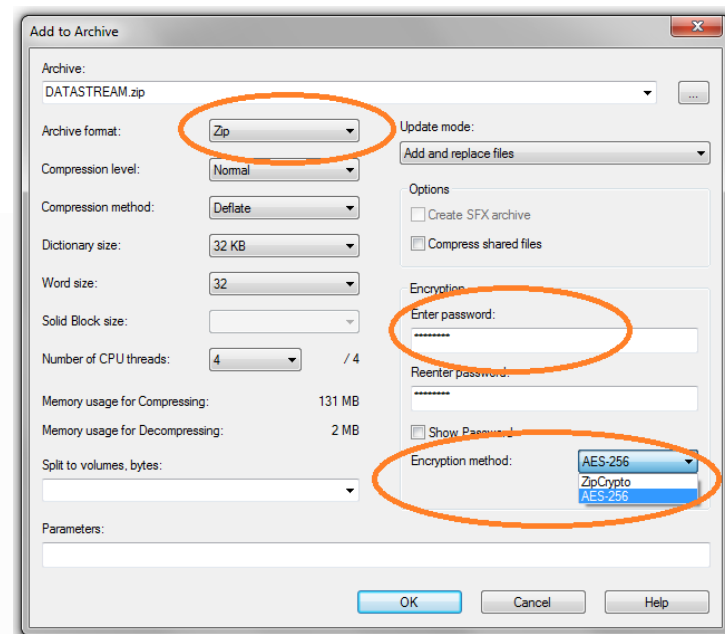
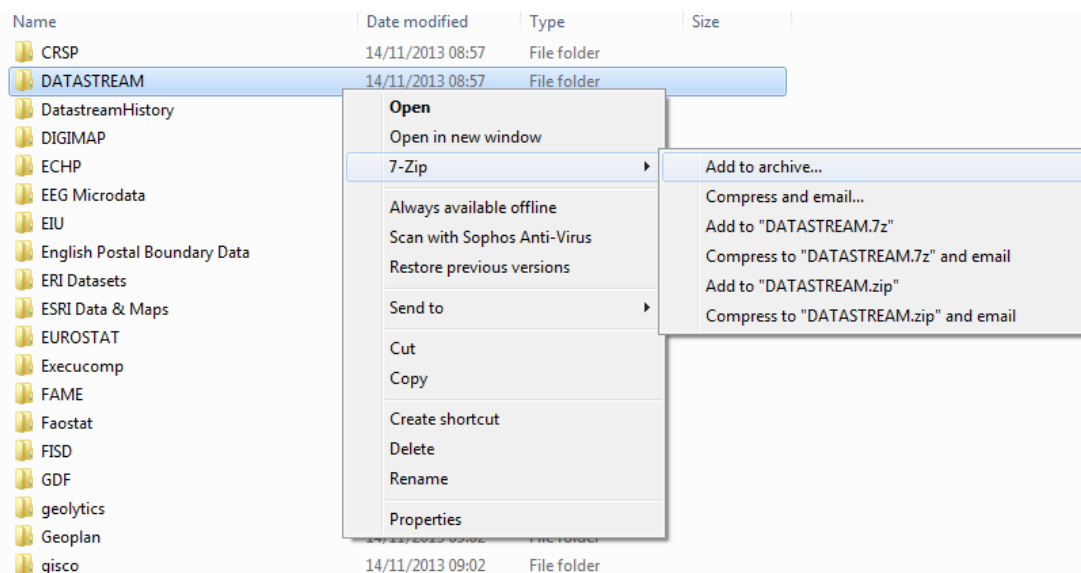


Image: pixabay (CC-0)



# 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





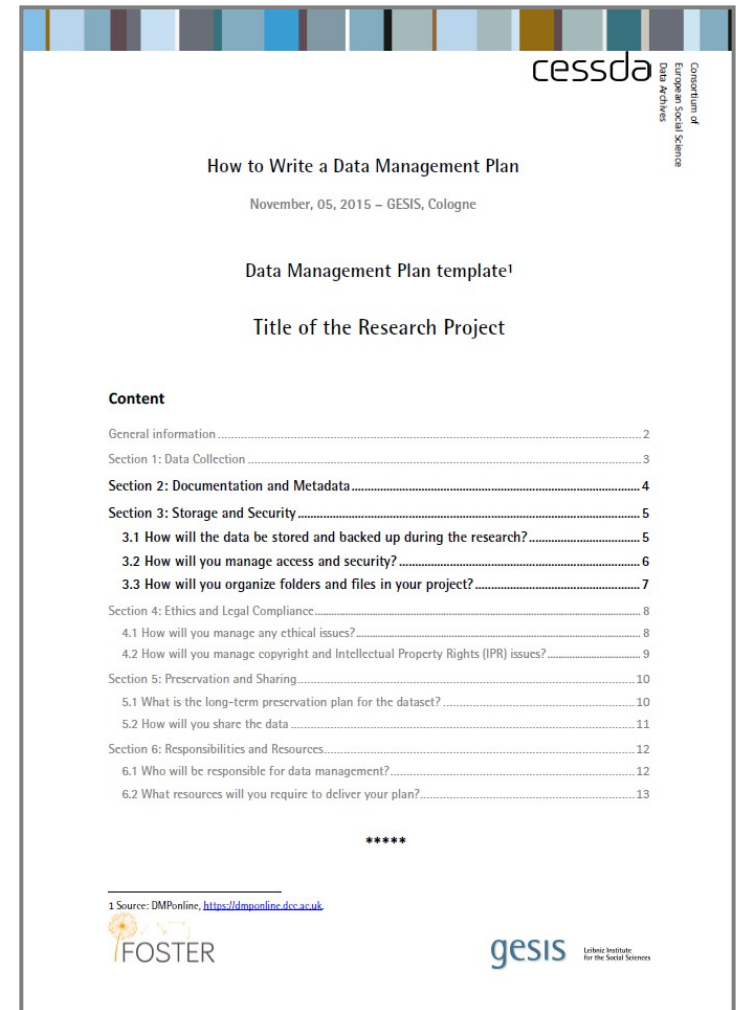
# Further Reading

- 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](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. Fraunhofer Institut, SIT Technical Report. [https://www.sit.fraunhofer.de/fileadmin/dokumente/studien\\_und\\_technical\\_reports/Cloud-Storage-Security\\_a4.pdf](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](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). Bethesda, MD: NISO Press. [www.niso.org/standards/resources/UnderstandingMetadata.pdf](http://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](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 3-6 groups,
- ☐ Time: about 30 minutes
- ☒ Choose one of the following topics
- ☐ For more detailed information have a look at the exercise sheet in your folder
- ☐ At the end, we will ask participants to briefly present their results







# DMP Sections 2 and 3

a) Documentation (Section 2), focusing on the documentation of inconsistencies in the data

- how to deal with such inconsistencies
- how to document decision and steps undertaken to deal with inconsistencies

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

- what, where and how often / long are files 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

- Several dissatisfying solutions

- delete observations or single responses
- recode information
- ignore inconsistency

- A more satisfying solution for re-users

⇒ keep inconsistency but highlight it for reuse of data

| There are some instances in which the number of persons  
| in household is equal to or less than the number of persons  
| under age 18. These data remained unchanged.

...

	EQUAL	LESS
AUSTRIA (2008)	17	1
BELARUS (2008)	2	0
CANADA (2008)	2	0
CZECH REPUBLIC (2010)	0	1



# 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

- to name files  
⇒ e.g. [type\_name\_version]
- to version files  
⇒ e.g. by the date and editor's acronyms

*data\_RDMDData\_20150822sn*

