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

- Research
- Study planning
- Data collection
- Data analysis
- Archiving & registering
Why Data Documentation?

What do the codes mean?

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

Who has been observed etc.?

What do the variables mean and what are the underlying questions

What do the codes mean?

keep your data understandable
Levels of Data Documentation

- **Study level**
  - study description
  - study design
  - data processing

- **Variable level**
  - questionnaire
  - variables and codes
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
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

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

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

- Research
- Archiving & registering
- Study planning
- Data analysis
- Data collection
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
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
  – separate personal data from other data

• Use passwords and encryption
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.
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 Readings

- Starr, J., 2011, DataCite Metadata Schema for the Publication and Citation of Research Data (p. 29). doi:10.5438/0005
work in 3 groups,
⏰ time: about 30 minutes
☑ choose one of the following topics
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