GEN-8001: Take control of your PhD journey

Research data management
Part 2: (Qualitative) data containing personal/sensitive information

Helene N. Andreassen, PhD
Erik Axel Vollan, senior engineer

University Library, October 11, 2018
To address this afternoon

- The planning
- The collection
- The archiving
- The citing
To address this afternoon

The planning

The collection

The archiving

The citing
The Data Management Plan: content

- Project information
- Responsibilities and rights

1. Collecting/generating data
2. Documentation and metadata
3. Storage and backup in project period
4. Archiving and sharing

- Ethics and data protection
  - Consent form*
  - Judgment of the likelihood of collecting sensitive data
  - RoS analysis, notification of project to NSD (Norwegian Data Protection services)


Have all future stages in mind when working on this.

The library not only offers courses on DMP, but can also give tips and advice in the DMP writing process.
The Data Management Plan: template

Project subject to notification to NSD:
NSD template

Project funded by EU, Horizon 2020:
DMPonline

All other projects:
UIT template
Searching for other people’s research data

Get a deeper understanding of and an increased trust in the text you’re reading.

Avoid starting your data collection from scratch.

Adjust method and focus of your planned data collection.

THE SEARCH ENGINE
DataCite
BASE – Bielefeld Academic Search Engine
Google Dataset Search

THE REPOSITORY
Figshare
Dryad
Zenodo

THE REPOSITORY REGISTRY
Re3data
The repository registry

https://www.re3data.org/
The Fragile Families & Child Wellbeing Study is following a cohort of nearly 5,000 children born in large U.S. cities between 1998 and 2000 (roughly three-quarters of whom were born to unmarried parents). We refer to unmarried parents and their children as 'fragile families' to underscore that they are families and that they are at greater risk of breaking up and living in poverty than more traditional families. The core Study was originally designed to primarily address four questions of great interest to researchers and policy makers: (1) What are the conditions and capabilities of unmarried parents, especially fathers? (2) What is the nature of the relationships between unmarried parents? (3) How do children born into these families fare? and (4) How do policies and environmental conditions affect families and children?
CESSDA: Consortium of European Social Science Data Archives

https://www.cessda.eu/Consortium
Some other repositories to check out

• ICPSR (Inter-University Consortium for Political and Social Research) database): [https://www.icpsr.umich.edu/icpsrweb/](https://www.icpsr.umich.edu/icpsrweb/)

• NSD: [https://search.nsd.no/vi-er-paa-vei](https://search.nsd.no/vi-er-paa-vei)

• QDR (Qualitative Data Repository): [https://qdr.syr.edu/deposit](https://qdr.syr.edu/deposit)

• SHARE (Survey of Health, Aging and Retirement in Europe): [http://www.share-project.org/](http://www.share-project.org/)
To address this afternoon

The planning

The collection

The archiving

The citing
Storing data

Collection
• Use a laptop, preferably UiT supplied.
• Store recordings etc. DIRECTLY to hardware encrypted USB Device. ITA has guides.

Treatment
• For sensitive data (health info etc.), use TSD – Services for Sensitive Data.
• For “yellow” data we do not have a good solution PT. Use encrypted devices.
  • ITA is working on this now. Risk Assessments plus adherence to Rules and Regulations are paramount!
• For Green data general storage is available. Terminal server access to statistics software.
  • Researchers quota: 5 TB of storage

Support
• For all questions regarding type of equipment and storing, contact us via research-support@uit.no.
• *We aim to have better solutions ready for you in 2019!*
Sensitive data

• UiT buys service from the University of Oslo

• Closed environment for storage and analysis
  • Statistics
  • NVIVO

• Very high level of security
  • Strict controls of data import and export
  • All projects totally isolated from other projects

• Nettskjema can send encrypted forms directly to TSD

For more info: TSD homepage
Digital Research Services (UIT) can give local support. Contact us via research-support@uit.no.
GDPR (the EU General Data Protection Regulation)

• “The GDPR applies to personal data, meaning any information relating to an identifiable person who can be directly or indirectly identified” (https://eugdpr.org)

• GDPR @ UiT: several working groups already in place, to improve routines and help the UiT staff fulfilling the requirements. For more info, see https://uit.no/om/art?p_document_id=554272&dim=179105

• Data protection @ UiT
  • Speak with your supervisor!
  • NSD webpages: http://www.nsd.uib.no/
  • Contact point between NSD and UiT: Sølvi B. Andersen.
  • From Nov 1: A common contact point at UiT for all issues related to privacy protection (Joachim Bakkevold). Until Nov 1, contact Sølvi if you have questions.
Treatment of files with personal data

• De-identification should be carried out simultaneously with treatment (e.g. transcription). Waiting until the end might end up with you not having time to do it, and in the worst case, the data must be deleted.

• Store the scrambling key separately from the data material.

• Anonymise when planned, incl. e.g. destroying the scrambling key linking data and person.

• Vocabulary demystifier:
Structuring and documentation

Key issues: file format – file naming – file description – readme file
Persistent file formats ensure that people in the future may open (and reuse) your files

- Non-proprietary
- Open, with documented international standards
- In common usage by the research community
- Using standard character encodings (e.g. UTF-8)
- Uncompressed (space permitting)
Persistent file formats (examples)

<table>
<thead>
<tr>
<th><strong>Document type</strong></th>
<th><strong>Persistent format (examples)</strong></th>
<th><strong>Non-persistent format (examples)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Plain text (.txt), PDF/A</td>
<td>MS Word (.docx)</td>
</tr>
<tr>
<td>Spreadsheet</td>
<td>Tabulator-separated Unicode UTF-8 text (.txt)</td>
<td>MS Excel (.xlsx)</td>
</tr>
<tr>
<td>Image</td>
<td>Uncompressed TIFF</td>
<td>Windows Bitmap (.bmp)</td>
</tr>
<tr>
<td>Sound</td>
<td>WAV</td>
<td>AAC (.m4a)</td>
</tr>
<tr>
<td>Video</td>
<td>MPEG-4</td>
<td>Quicktime (.mov)</td>
</tr>
</tbody>
</table>

See the UiT Open Research Data Deposit Guide for more information: [https://site.uit.no/dataverseno/deposit/prepare/](https://site.uit.no/dataverseno/deposit/prepare/)
File naming

CONSISTENT
observation_notes_2017-11-01.txt
interview_notes_2017-11-01.txt

DESCRIPTIVE
interview_notes_Kirkenes.txt
interview_notes_Altar.txt

SIMPLE (why is this bad file naming?)
Field notes: May new 171017 Åse
“..a systematic human error in coding the name of the files had been made during the extraction of the EEG template topographic maps best differentiating the two experimental conditions at the single subject level.”
# File naming & organising

<table>
<thead>
<tr>
<th>BY DATE</th>
<th>BY TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1955-04-12_notes_potatofarms</td>
<td>Notes_potatofarms_1955-04-12</td>
</tr>
<tr>
<td>1955-04-12_questionnaire_potatofarms</td>
<td>Notes_cornfarms_1956-05-11</td>
</tr>
<tr>
<td>1956-05-11_notes_cornfarms</td>
<td>Questionnaire_potatofarms_1955-04-12</td>
</tr>
<tr>
<td>1956-05-11_questionnaire_cornfarms</td>
<td>Questionnaire_cornfarms_1956-05-11</td>
</tr>
<tr>
<td></td>
<td>FORCED ORDER</td>
</tr>
<tr>
<td></td>
<td>01_Potatofarms_questionnaire_1955-04-12</td>
</tr>
<tr>
<td></td>
<td>02_Potatofarms_notes_1955-04-12</td>
</tr>
<tr>
<td></td>
<td>03_Cornfarms_questionnaire_1956-05-11</td>
</tr>
<tr>
<td></td>
<td>04_Cornfarms_notes_1956-05-11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BY SUBJECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornfarms_notes_1956-05-11</td>
</tr>
<tr>
<td>Cornfarms_questionnaire_1956-05-11</td>
</tr>
<tr>
<td>Potatofarms_notes_1955-04-12</td>
</tr>
<tr>
<td>Potatofarms_questionnaire_1955-04-12</td>
</tr>
</tbody>
</table>
The ReadMe file: the guide to your data

What is the dataset about?
Overview of the files
Methods (conditions for data collection and treatment)
File structure and naming conventions
Column headings in tabular data
Abbreviations

This post contains five csv datasets of Russian nouns, plus an R script for their analysis.

The five csv datasets are the following:
- procent-I-m.aa.csv: This is data on masculine animate I-declension nouns.
- procent-I-m.nn.csv: This is data on masculine inanimate I-declension nouns.
- procent-I-nr.nn.csv: This is data on neuter I-declension nouns.
- procent-II-f.nn.csv: This is data on feminine inanimate II-declension nouns.
- procent-III-f.nn.csv: This is data on feminine inanimate III-declension nouns.

This data comes from the SynTagRus corpus (https://github.com/UniversalDependencies/UD_Russian-SynTagRus).
Each dataset has the same structure.
Column 1 “freq” lists the frequency of the lemma in the corpus.
Column 2 “lemma” lists the lemma in question (in Cyrillic).
Column 3 “gramm” lists the type of noun and is the same throughout each file. For the file procent-I-m.aa.csv, for example, all items are marked “Masc.Anim.”
Column 4 “total” is the total frequency and is identical to column 2.
Columns 5 through 16 give the relative frequency (percent) of attestations for each case/number combination for this lemma. sg=singular, pl=plural, nom=nominative, gen=genitive, dat=dative, acc=accusative, ins=instrumental, loc=locative.

For example, the first row of the dataset procent-I-m.aa.csv begins like this:

<table>
<thead>
<tr>
<th>freq</th>
<th>lemma</th>
<th>gramm</th>
<th>total</th>
<th>sg.nom</th>
<th>sg.gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>2651</td>
<td>человек</td>
<td>Masc.Anim</td>
<td>2651</td>
<td>14.79</td>
<td>14.22</td>
</tr>
</tbody>
</table>

This means that 2651 forms of the word человек ‘person’ appear in the corpus, and that 14.79% of them are Nominative Singular forms, 14.22% are Genitive Singular forms, etc.

The R script shows the code needed to read these files into R and perform the correspondence analysis.

(Janda & Tyers, 2018)
The ReadMe file: the guide to your data

What is the dataset about?
Overview of the files
Methods (conditions for data collection and treatment)
File structure and naming conventions
Column headings in tabular data
Abbreviations

<table>
<thead>
<tr>
<th>File name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview Transcripts</td>
<td>Folder containing interview transcripts, in six sub folders by category of participant: BRITISH PAK MARRIED INTRANATIONALLY BRITISH PAK MARRIED TRANSNATIONALLY PAKISTANI MIGRANT SPOUSES BRITISH SIKHS MARRIED INTRANATIONALLY BRITISH SIKHS MARRIED TRANSNATIONALLY SIKH MIGRANT SPOUSES</td>
</tr>
<tr>
<td>Guide to the data archived</td>
<td>This file provides background on the project, an overview of the data archived, including a list of interview transcripts with a key for transcript titles</td>
</tr>
<tr>
<td>Interview Schedule</td>
<td>This file contains the interview schedule used for the semi-structured interviews</td>
</tr>
<tr>
<td>Consent Form</td>
<td>Consent form for interview participants</td>
</tr>
<tr>
<td>Marriage Migration and Integration Project Report</td>
<td>Project Report providing overview of findings from the project as a whole, and more information on the project including methods and sampling,</td>
</tr>
</tbody>
</table>

The project report can also be found at: http://www.bristol.ac.uk/ethnicity/projects/mmi/
To address this afternoon

- The planning
- The collection
- The archiving
- The citing
Which data to archive (openly)?

Archiving with
- open access to primary data, structured data, and metadata

Archiving with
- closed/restricted access to primary data
- open access to structured data, and metadata

Archiving with
- closed/restricted access to primary data and structured data
- open access to metadata

Archiving with
- closed/restricted access to primary data, structured data, and metadata
Exercise: Where to put your (meta)data?
Turn to your neighbor, and help each other identify in which box(es) your primary/structured data and metadata are to be placed.
Repository information to pay attention to

Questions to ask, regardless of whether you plan to publish data openly, or whether only metadata will be made (fully/partly) available.

1. Is the data repository reputable?
2. Will the repository take the data you want to deposit?
3. Will the data be safe in legal terms?
4. Will the repository sustain the data value?
5. Will the repository support analysis and track data usage?

(Whyte, 2015)
To address this afternoon

The planning

The collection

The archiving

The citing
Elements in a data reference

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier*</td>
<td>Unique string that identifies the dataset (doi, handle)</td>
</tr>
<tr>
<td>Author*</td>
<td>The researcher(s) having produced the data and are authors of the corresponding journal article</td>
</tr>
<tr>
<td>Title*</td>
<td>Name of the dataset</td>
</tr>
<tr>
<td>Publisher*</td>
<td>Name of the archive</td>
</tr>
<tr>
<td>Year of publication*</td>
<td>Moment when the data are made available</td>
</tr>
<tr>
<td>Version</td>
<td>If dataset changes, the version number changes</td>
</tr>
<tr>
<td>Type of data</td>
<td>e.g. dataset, corpus, picture archive</td>
</tr>
<tr>
<td>Related identifier</td>
<td>Full dataset in the case of subset use</td>
</tr>
</tbody>
</table>

*Elements that are obligatory in the reference

The citation for this study is:

Contribute to the discussion!

Become a member of the Research Data Alliance

- Linguistics Data Interest Group: https://www.rd-alliance.org/groups/linguistics-data-ig
An inspirational paper:

Show me the data: Research reproducibility in qualitative research

by

Neil Dymond-Green

2018, September 18

http://blog.ukdataservice.ac.uk/show-me-the-data/
GEN-8001: Take control of your PhD journey

Research data management
Part 2: (Qualitative) data containing personal/sensitive data

UiT Research Data Portal: https://uit.no/researchdata
Email: research-data@support.uit.no
References


Janda, L. A. & Tyers, F. M. (2018). *Replication Data for: Less is More: Why All Paradigms are Defective, and Why that is a Good Thing*, [https://doi.org/10.18710/VDWPZS](https://doi.org/10.18710/VDWPZS), DataverseNO, V1, UNF:6:hiDJKeFSs1ZOccEs0dU0gw== [fileUNF]


All pictures are taken from Colourbox.com, if not otherwise stated.