GEN-8001: Take control of your PhD journey

Research data management

Part 2: Best practices, data with sensitive information

Helene N. Andreassen, PhD
Erik Axel Vollan, Senior Engineer

University Library, 27 February 2020
To address this afternoon

The planning

The collection and the treatment

The archiving

The citing (if we have time)
The searching (if we have time)
To address this afternoon

- The planning
- The collection and the treatment
- The archiving
- The citing
$5,000 REWARD
Backpack with PhD Data
No Questions Asked
call +1 (315) 657-5266 or +1 (315) 546-6298
Google Voice: 352-561-8317
Whats app: +90 543 493 97 30
cloismeg@gmail.com

When: Nov 25 (Saturday) around noon

What: backpack with PhD data on thumb drive, laptop, handwritten notebook, leather bag Lenox

Where: My car at Drummond and Sherbrooke streets

https://montreal.ctvnews.ca/phd-student-offering-5-000-reward-after-car-thief-steals-all-his-research-1.3700484
Information security

• UiT Information Security Governance

• 4 classes of confidentiality:

  - Grønn: Open
  - Gul: Intern
  - Rød: Fortrolig
  - Svart: Streng fortrolig

• You are responsible for the correct classification of your information

• https://uit.no/sikkerhet
Classification dictates your storage options

<table>
<thead>
<tr>
<th>System / tjeneste</th>
<th>Åpen/Grønn</th>
<th>Intern/Gul</th>
<th>Fortrolig/Rød</th>
<th>Strengt fortrolig/Svart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canvas</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Ephorte</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>E-post (office 365)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>EURO</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Fellesområder</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Felles Studensystem (FS)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Hjemmeområdet (H1)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Medisiste</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Nettiskjema / Sikkert netskjema</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>OneDrive for Business (office 365)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Request Tracker (RT)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Sharepoint (office 365)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Skype for Business</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Sway (office 365)</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Teams (office 365)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>TopDesk</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Tjeneste for sensitive data (TSD)</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>Yammer (office 365)</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
<tr>
<td>WiseFlow</td>
<td>OK</td>
<td>OK</td>
<td>ikke godkjent</td>
<td>ikke godkjent</td>
</tr>
</tbody>
</table>
Office 365 for research data

• Office 365 has tools for storage and collaboration
  • SharePoint - groups and projects
  • OneDrive - Personal storage  NB!! OneDrive is deleted when user leaves UiT!

• Azure Information Protection and 2-factor auth is mandatory for data in Red Category!
  • E-learning in the UiT e-learning portal Xtramile

• More info in Orakel (helpdesk) pages on Office 365
Services for Sensitive data

• UiT buys this service from the University of Oslo

• Closed environment for storage and analysis
  • Statistics software
  • NVIVO

• Very high level of security
  • Strict controls of data import and export
  • All projects totally isolated from other projects
  • Can have users that are external to UiT

• Nettskjema can send encrypted forms directly to TSD
  • Smartphone Dictaphone app can store recordings 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](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](https://uit.no/om/art?p_document_id=554272&dim=179105)

• Data protection @ UiT
  • Speak with your supervisor!
  • Data protection officer @ UiT: Joachim Bakkevold ([personvernombud@uit.no](mailto:personvernombud@uit.no))
  • NSD webpages: [http://www.nsd.uib.no/](http://www.nsd.uib.no/)
  • Contact point between NSD and UiT: Sølvi B. Anderssen
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.

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>Document type</th>
<th>Persistent format (examples)</th>
<th>Non-persistent format (examples)</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 DataverseNO Deposit Guide for more information:
https://site.uit.no/dataverseno/deposit/prepare/
Structuring and documentation

Key issues: file format – file naming – file description – readme file
File naming

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

& DESCRIPTIVE
interview_notes_Kirkenes.txt
interview_notes_Alta.txt

(why 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

### BY DATE
- 1955-04-12_notes_potatofarms
- 1955-04-12_questionnaire_potatofarms
- 1956-05-11_notes_cornfarms
- 1956-05-11_questionnaire_cornfarms

### BY TYPE
- Notes_potatofarms_1955-04-12
- Notes_cornfarms_1956-05-11
- Questionnaire_potatofarms_1955-04-12
- Questionnaire_cornfarms_1956-05-11

### BY SUBJECT
- Cornfarms_notes_1956-05-11
- Cornfarms_questionnaire_1956-05-11
- Potatofarms_notes_1955-04-12
- Potatofarms_questionnaire_1955-04-12

### FORCED ORDER
- 01_Potatofarms_questionnaire_1955-04-12
- 02_Potatofarms_notes_1955-04-12
- 03_Cornfarms_questionnaire_1956-05-11
- 04_Cornfarms_notes_1956-05-11
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

Check out https://data.research.cornell.edu/content/readme

README file for replication data for “Less is More: Why All Paradigms are Defective, and Why that Is a Good Thing” by Laura A. Janda and Francis M. Tyers

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

The five csv datasets are the following:
- percent-l-m.aa.csv This is data on masculine animate I-declension nouns.
- percent-l-m.nn.csv This is data on masculine inanimate I-declension nouns.
- percent-l-nt.nn.csv This is data on neuter I-declension nouns.
- percent-ll-f.nn.csv This is data on feminine inanimate II-declension nouns.
- percent-lll-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 percent-l-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 percent-l-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

...
Activity: Making our data reusable

(se handout)
Activity: Making our data reusable

Folder: Tromsø class F

Files in the folder:
- Notes Dec2018.docx
- Notes Nov2018_new.docx
- 2018 oktober med notater.docx
- Original interview guide.doc
- Data-class6.xlsx
- Data class F treated.nvp
- Data class F treated V2.nvp
- Data class E treated together with class F_ÅseØstmo.nvp
Activity: Making our data reusable

Part of the content in the file Data-class6.xlsx

<table>
<thead>
<tr>
<th>Group</th>
<th>Lng</th>
<th>date</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Troms-1</td>
<td>1</td>
<td>Eng</td>
<td>0.25</td>
</tr>
<tr>
<td>Troms-2</td>
<td>1</td>
<td>Eng</td>
<td>0.75</td>
</tr>
<tr>
<td>Troms-3</td>
<td>1</td>
<td>Eng</td>
<td>1.25</td>
</tr>
<tr>
<td>Troms-1</td>
<td>1</td>
<td>Fra</td>
<td>1.75</td>
</tr>
<tr>
<td>Troms-1</td>
<td>2</td>
<td>Spa</td>
<td>2.25</td>
</tr>
<tr>
<td>Troms-2</td>
<td>2</td>
<td>Spa</td>
<td>2.75</td>
</tr>
</tbody>
</table>
To address this afternoon

- The planning
- The collection and the treatment
- The archiving
- The citing (if we have time)
- The searching (if we have time)
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 allow analysis of reuse?

(Whyte, 2015)
Archiving sensitive data with restricted access

- If NSD is to archive personal data, the data owner must document that storage of the data is permitted pursuant to the applicable regulations.

- Before NSD can store personal data, it is required by law that a data processor agreement be entered into between NSD and the data owner.

- Personal data are not published on NSD’s website, but it is possible to publish information about the study. If the data are to be freely accessible, the data must be anonymised.

- Personal data can only be distributed in accordance with the data processor agreement entered into between the data owner and NSD. An alternative is to anonymise the survey before distribution, but this must always be approved by the data owner.

(information taken from https://nsd.no/)
UiT Open Research Data

A data archiving service for archiving, sharing, reusing, and citing open research data.

Available for upload to all employees and students at UiT via Feide login.

Available to all for download and reuse.

Visible to main search engines.

Possible to create private URL for reviewers prior to publication.

opendata.uit.no
To address this afternoon

- The planning
- The collection and the treatment
- The archiving
- The citing (if we have time)
- The searching (if we have time)
A data management plan (DMP) is a plan that documents how you are going to manage your research data during and after the project period.

Covers all phases in the research data management lifecycle
- Search
- Collection
- Processing: analysis and storage
- Archiving and sharing

Created before/not long after project start, but usually revised during the project.
What is the purpose of a data management plan?

- Help you to have control over your data throughout the whole project
- Help you to save time and extra work later on
- Make your data as FAIR as possible

FAIR: Findable, Accessible, Interoperable, Reusable

Image credit: Soraja Pudhi, Wikimedia Commons CC BY-SA 4.0
Typical aspects to be documented in a Data Management Plan
General information about the research project

- Project name.
- Description of the project, purpose of data collection and the utility value of data.
- Is the project part of a larger research project?
- Funding of the project.
- Project leader and participants (name and institution).
Responsibilities and rights

• Who is responsible for follow-up and revision of the DMP?
• How will the responsibilities be divided? External collaborators?
• Who has right to manage the data? Think all the steps: collection, structuring, processing, etc.
• Who can access the data during the project? (View, use)
• Who has ownership of the data? [UiT guidelines: UiT The Arctic University of Norway](http://site.uit.no/rdmtraining/research-data-and-agreements/)

**IMPORTANT:** If several partners, make sure you have a written agreement. See information about agreements on [http://site.uit.no/rdmtraining/research-data-and-agreements/](http://site.uit.no/rdmtraining/research-data-and-agreements/). Ask for advice if needed.
Collecting data

• If there are existing data, can these be integrated and reused?

• When will the data be collected?

• Which methods for data collection will be used? E.g. observations, interviews.

• What type of data will be collected? E.g. text, sound, images, numerical data.

• Does collection of the data require extra software? Need for special expertise?
Structuring and documenting

• Which file formats will be used?
• What kind of folder structure and filename conventions will be used?
• How will the data be documented? (ReadMe file)
• If metadata standards are used (to ensure standardized descriptions), which ones? (Examples are found here: https://guides.library.uwa.edu.au/c.php?g=325196&p=2178564)
• Is special software for reading/interpreting the data necessary?
Ethics and consent

• Will you collect data about people?

• If yes, are they personally identifiable?

• If yes, have you been in contact with a data protection official?

• Consent forms: see
  http://www.nsd.uib.no/personvernombud/en/help/information_consent/

• Data protection at UiT: uit.no/etikk
Storage and preservation during the project

• For collecting in the field, how will the data be safely recorded and transferred from the field to the main storage facility?
• Which procedures for storage, backup and restoration?
• Do you have sufficient storage possibilities or need for extra services? What is the expected file size for the data?
Archiving and sharing (1)

• Which data will be preserved and which will be destroyed?
• Will the data be long time preserved, and how is this decided?
• Will the data or a selection of the data be openly shared, and if so, which data?
• Do the data need processing before they can be shared? (anonymization, file conversion)
• If data will not be shared, what is the reason?
• If the data cannot be shared, what about the metadata?
• How can access be provided in case of restrictions?
Archiving and sharing (2)

- Where will data, metadata, documentation and code be archived?
- When will the data be made openly available, and how long will they be stored?
  - UiT: as early as possible, no later than the date of publication (and end of project for other data)
- How will the data be licensed for reuse?
  - UiT guidelines: as few limitations as possible
- Are there other conditions, restrictions or embargo on use?
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
To address this afternoon:

- The planning
- The collection and the treatment
- The archiving
- The citing (if we have time)
- The searching (if we have time)
Hansen, Pernille. 2016, "Replication data for: What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development", https://doi.org/10.18710/JEWIVV, DataverseNO, V1

Elements in a data reference

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author*</td>
<td>The researcher(s) having produced the data and are authors of the corresponding journal article</td>
</tr>
<tr>
<td>Year of publication*</td>
<td>Moment when the data are made available</td>
</tr>
<tr>
<td>Title*</td>
<td>Name of the dataset</td>
</tr>
<tr>
<td>Identifier*</td>
<td>Unique string that identifies the dataset (doi, handle)</td>
</tr>
<tr>
<td>Publisher*</td>
<td>Name of the archive</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:


(Altman & Crosas, 2013; Starr & Gastl, 2011)
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 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/ver-paa-vei](https://search.nsd.no/ver-paa-vei)
- QDR (Qualitative Data Repository): [https://qdr.syr.edu/deposit](https://qdr.syr.edu/deposit)
Contribute to the discussion!

Become a member of the Research Data Alliance

• Empirical Humanities Metadata Working Group: https://www.rd-alliance.org/groups/empirical-humanities-metadata-working-group.html
• 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/
Other papers on the reuse of qualitative data


Jones, K., & Alexander, S. M. (2018). Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches. [https://doi.org/10.13016/M2WH2DG59](https://doi.org/10.13016/M2WH2DG59)

GEN-8001: Take control of your PhD journey

Research data management

Part 2: Best practices, data with sensitive information

Helene N. Andreassen, PhD
Erik Axel Vollan, Senior Engineer

University Library, 27 February 2020
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, DataverseNO, V1, UNF:6:hiDJKfSs1Z0ccEs0dU0gw== [fileUNF]


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