GEN-8001: Take control of your PhD journey

Research data management
Part 2: Data without sensitive information

Glennda Villaflor, MSc.

University Library, March 14, 2019
Lifecycle of Research Data Management

Phases:
- Searching / reusing
- Planning
- Collecting
- Processing
- Archiving / sharing

Adapted original source: The University of California, Santa Cruz, Data Management LibGuide, Research Data Management Lifecycle, diagram, viewed May 2, 2016 at <http://guides.library.ucsc.edu/datamanagement>
Outline

PART 2 – BEST PRACTICE FOR RESEARCH DATA MANAGEMENT

• Search and cite
  • re3data, DataCite, BASE, Google Dataset Search

• Structure and document

• Archive and share

• Data management plan
  • Distribute UiT’s DMP template
  • ACTIVITY: Write down and discuss in groups any questions you might have regarding your DMP
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research data</td>
<td><em>is understood in this policy to mean the registration/recording/reporting of numerical scores, textual records, images and sounds that are generated by or arise during research projects.</em></td>
</tr>
<tr>
<td>Secondary data</td>
<td><em>are data that already exist, regardless of the research to be conducted.</em></td>
</tr>
<tr>
<td>Publicly funded</td>
<td><em>refers to all projects and activities that are wholly or partly funded with government allocations.</em></td>
</tr>
<tr>
<td>Open access</td>
<td><em>is understood as the principle that research data should be accessible to relevant users, on equal terms, and at the lowest possible cost.</em></td>
</tr>
<tr>
<td>Data management plan</td>
<td><em>is a document describing how research data from a project are to be managed, from project start to finish.</em></td>
</tr>
<tr>
<td>Metadata</td>
<td><em>are data used to define or describe other data.</em></td>
</tr>
<tr>
<td>Digital Object Identifier (DOI)</td>
<td><em>is a persistent, unique identifier for research data and publications.</em></td>
</tr>
</tbody>
</table>

https://www.forskningsradet.no/servlet/Satellite?cid=1254032622112&pagename=VedleggPointer&target=_blank
Searching / reusing
Searching data: The archive registry

Registry of Research Data Repositories (re3data)

- Searchable registry for research data archives with open metadata
- www.re3data.org
Searching data:
The search engine

DataCite

• Multi-disciplinary search engine
• Gathers metadata for each DOI assigned to an object
• Metadata free to access
• Combinatory searches possible
• https://search.datacite.org/
Searching data: The search engine

**Bielefeld Academic Search Engine (BASE)**

- Multi-disciplinary search engine
- Indexes all resources providing an Open Archives Initiative interface
- Metadata free to access
- Refined search options
- [https://www.base-search.net/](https://www.base-search.net/)
Searching data: The search engine (NEW)

Google Dataset Search (Beta)

- Launched 5. september, 2018
- Interdisciplinary search engine
- Index data from DataCite
- Metadata freely available
- [https://toolbox.google.com/datasetsearch](https://toolbox.google.com/datasetsearch)
Searching data: The archive

- Zenodo, Dryad, Figshare: Internasjonalt mye brukte arkivtjenester
- UiT Open Research Data: UiTs eget institusjonsarkiv
- Tverrfaglige
- Gratis
- Gode hjelpemenyer

UiT Open Research Data
support transparent and reproducible research
Searching data: The archive

UiT Open Research Data

- Multidisciplinary repository
- Metadata and data freely accessible
- Refined search options
- [https://opendata.uit.no](https://opendata.uit.no)
Description

The files contain data for reproducing all the results in the article "Benchmarking density functional methods for harmonic vibrational frequencies" (in REVIEW). The file frequency_data_for_statistical_analysis.xlsx is an Excel file containing 11 differently named worksheets. Each worksheet contains the name of the XC functional used. All the quantities are calculated using the standard mathematical formula of Excel. The distribution_of_signed_error_plot.pdf is a PDF file containing the distribution of signed error obtained for each molecule using 17 different XC functionals. The distribution plots are obtained using the distribution formula given in the upcoming article. All the plots have been created using GNUPLOT software. The text files are tab delimited text files obtained from the Excel worksheets. (2018-07-16)

Subject

Chemistry

Keyword

Harmonic vibrational frequency, Benchmarking DFT functionals, Computational Chemistry, Density functional theory, Quantum Chemistry
Alam, Md Mehboob, 2018, "Replication Data for: Benchmarking density functional methods for harmonic vibrational frequencies. IN REVIEW", https://doi.org/10.18710/2DQK6Z, DataverseNO, V1

- Automatically generated reference
- Version control
- DOI (permanent link)
- Temporary lock on files (ex. until the publication date of article)
- Possibility to share access and editing permission before publication (ex. for peer review collaboration project)
Replication data for: Are universal linguistic hierarchies innately wired?
Oct 21, 2018 - TROLLing (The Tromsø Repository of Language and Linguistics)

Lerdon, Evelina, 2018, "Replication data for: Are universal linguistic hierarchies innately wired?", https://doi.org/10.18710/9UYJNB, DataverseNO, V1

This research put the nature and rigidity of linguistic hierarchies to test, taking multiple adjective placement as a case study. We developed an online forced choice experiment that measured (i) acceptability judgment ratings and (ii) reaction times, in a big sample of uncolling.

TGO Ramfjordmoen Ionosonde Data January 1985
Oct 15, 2018 - Tromsø Geophysical Observatory

Tromsø Geophysical Observatory, 2018, "TGO Ramfjordmoen Ionosonde Data January 1985", https://doi.org/10.18710/7BFJQQ, DataverseNO, V1

About this dataset. This dataset contains ionograms in PNG format and covers data from January 01.31.1985. About the Tromsø ionosonde (Sit. 1932). During the period 1904-1992 the ionosonde was situated at 69° 35' N, 19° 13' E at Ramfjord near Tromsø, Norway and operated...
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/JEWIVW, DataverseNO, V1

The citation for this study is:


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>
Coretta, Stefano, 2016, «Data on Icelandic pre-aspiration», doi:10.18710/7NLJSG, DataverseNO, V1
Example

Bibliographical reference:
Coretta, Stefano, 2016, "Data on Icelandic pre-aspiration", doi:10.18710/7NLJSG, UiT Open Research Data Dataverse, V1

In-text citation
This is just a nonsense text that is actually about consonants (Coretta, 2016). Now I will refer to a particular file in the dataset (Coretta, 2016, file name: brs02_20160531.wav), which I have looked at more closely because it contains data on the consonants [t,k], which is the subject of my own study.
Collecting

Processing

Adapted original source:
The University of California, Santa Cruz, Data Management LibGuide, Research Data Management Lifecycle, diagram, viewed May 2, 2016 at <http://guides.library.ucsc.edu/datamangement>
Structuring and documentation – Why?

Key issues:
- Data storage
- File naming
- Documentation
- File format

How?

Morgan Edwards
@mangoedwards

I can't send you the original data because I don't remember what my excel file names mean anymore. #overlyhonestmethods

Heather Piwowar
@researchremix

Data are available upon request, but we really hope no one will ask. #overlyhonestmethods
3:06 PM · Jan 8, 2013

Brent Neal
@BrentN

Outliers have been removed from the dataset because the second author was drunk when he took the data. #overlyhonestmethods
8:42 PM · Jan 8, 2013
Good storage routines – avoid data loss!

- Versioning
- Regular backups
- Multiple backups

Home directory at UiT\homer.uit.no

Automatic backup & versioning

O365 OneDrive (← myDoc)
O365 SharePoint (← uDoc)
Box

Info & help (for more advanced storage service):
- UiT Research Data Portal: https://uit.no/researchdata
  >> Working with your research data
- Separate course module: How to store research data

Illustration: www.colourbox.no
Why YOU need a data management plan, CC-BY-2.0, by Peter Murray-Rust
File and folder naming

• Use consistent file and folder names

• Use descriptive, but short names (< 25 characters)

• Avoid space: Project notes.pdf

• Avoid special characters and non-roman letters: “/\:*?‘<[]()&æÆøØåÅ

• International date format: YYYY-MM-DD
“.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.”
Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander
BBC News

20 April 2013

This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

We are grateful to Herndon et al. for the careful attention to our original Growth in a Time of Debt AER paper and for pointing out an important correction to Figure 2 of that paper. It is sobering that such an error slipped into one of our papers despite our best efforts to be consistently careful. We will redouble our efforts to avoid such errors in the future. We do not, however, believe this regrettable slip affects in any significant way the central message of the paper or that in our subsequent work.
Cornell nutrition scientist resigns after retractions and research misconduct finding

By Kelly Servick | Sep. 21, 2018, 11:25 AM

Brian Wansink, the Cornell University nutrition researcher known for probing the psychology behind human eating habits, has resigned after a university misconduct investigation, and following the retraction this week of six of his papers.

In a statement issued yesterday, Cornell's provost, Michael Kotlikoff, said the investigation had revealed “misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship.”

Wansink contested the university's conclusion in a statement shared with Science, saying, "The interpretation of these four acts of misconduct can be debated, and I did so for a year without the success I expected." He admitted to mistaken reporting, poor documentation, and "some statistical mistakes," but maintains that there was "no fraud, no intentional misreporting, no plagiarism, or no misappropriation" in his work. "I believe all of my findings will be either supported, extended, or modified by other research groups," he added.

File and folder naming

Order by date:
1955-04-12_notes_MassObs.docx
1955-04-12_questionnaire_MassObs.pdf
1963-12-15_notes_Gorer.docx
1963-12-15_questionnaire_Gorer.pdf

Order by subject:
Gorer_notes_1963-12-15.docx
Gorer_questionnaire_1963-12-15.pdf
MassObs_notes_1955-04-12.docx
MassObs_questionnaire_1955-04-12.pdf

Order by type:
Notes_Gorer_1963-12-15.docx
Notes_MassObs_1955-04-12.docx
Questionnaire_Gorer_1963-12-15.pdf
Questionnaire_MassObs_1955-04-12.pdf

Force order by numbering:
01_MassObs_questionnaire_1955-04-12.pdf
02_MassObs_notes_1955-04-12.docx
03_Gorer_questionnaire_1963-12-15.pdf
04_Gorer_notes_1963-12-15.docx

Documentation
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
Units of measure

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

The five csv datasets are the following:
- percent-I-aa.csv: This is data on masculine animate I-declension nouns.
- percent-I-aj.csv: This is data on masculine inanimate I-declension nouns.
- percent-I-at.csv: This is data on neuter I-declension nouns.
- percent-II-fn.csv: This is data on feminine inanimate II-declension nouns.
- percent-III-fn.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-I-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-I-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.
PUBLIC SERVICE ANNOUNCEMENT:

Our different ways of writing dates as numbers can lead to online confusion. That's why in 1988 ISO set a global standard numeric date format.

This is the correct way to write numeric dates:

2013-02-27

The following formats are therefore discouraged:

02/27/2013  02/27/13  27/02/2013  27/02/13
20130227   2013.02.27  27.02.13  27-02-13
27.2.13    2013. II. 27.  27½-13    2013.158904109
MMXIII-II-XXVII MMXIII MMXIII MMXIII
1330300900 1330300900 1330300900
((3+3)×((111+1)-1)×3/3-1/3³  2013  2013
10/11011/1101 02/27/20/13
File formats

Persistent file formats ensure that people in the future may open (and reuse) your files

- Non-proprietary
- Open, and based documented international standards
- In common usage by the research community
- Using standard character encodings (e.g. Unicode 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 UiT Open Research Data Deposit Guide for more information: [https://site.uit.no/dataverseno/deposit/prepare/#what-are-persistent-file-formats](https://site.uit.no/dataverseno/deposit/prepare/#what-are-persistent-file-formats)
Archiving / sharing

Adapted original source:
The University of California, Santa Cruz,
Data Management LibGuide, Research Data Management Lifecycle, diagram,
viewed May 2, 2016 at <http://guides.library.ucsc.edu/datamanagement>
Archive and share
Is the data repository reputable?

• Is it listed in re3data?

• Is it broadly recognized in your research field?

• Is it endorsed by a relevant funder or journal?

• Is it certified? (Whyte, 2015)

UiT Open Research Data has applied for CoreTrustSeal certification. Answer expected in October 2018
Will the repository sustain the data value?

- **Metadata publishing**: Data collections arecatalogued in a repository according to funder expectations so that they are discoverable by title, creator, and date of deposition.
- **Stable identifiers**: Enables a DOI or other open standard identifier to be assigned to a landing page for each ingested dataset/collection.
- **Discoverable metadata**: Provides Datacite mandatory metadata and exposes it according to open access repository protocols.
- **File checking and preservation planning**: Supports virus check and file format validation at ingest, and provides format inventory.
- **Domain metadata & context information**: Supports linking of dataset to related records according to open repository protocols and standard identifiers.
- **Version control**: Ingested objects and latest edits are time-stamped.
Planning Phases:

Adapted original source:
The University of California, Santa Cruz, Data Management LibGuide, Research Data Management Lifecycle, diagram, viewed May 2, 2016 at <http://guides.library.ucsc.edu/datamanagement>
The Data Management Plan: Why?

• Required by funders, e.g. UiT, NFR, ECR
• Not just an administrative burden, but a useful document for you to keep control of your research data throughout your research project
• Helps you make your research reproducible and transparent
• Always have the FAIR data principles in mind!

Source: https://ogsl.ca/en/fair-principles
The Data Management Plan: What?

Typical aspects to consider:

• Project information
• Responsibilities and rights

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

Always have the FAIR data principles in mind:

- Findable
- Accessible
- Interoperable
- Re-Usable

Source: https://ogsl.ca/en/fair-principles
The Data Management Plan: How?

Different templates:

• Project subject to notification to NSD: [NSD template](#)

• Project funded by EU (ERC/Horizon 2020): [DMPonline](#)

• All other projects: [UiT template](#) (unless other requirement from funder)

The library not only offers courses on DMP, but can also give tips, advice and feedback on your DMP.
Activity!

Have a look at the UiT data management plan (DMP) template, and write down any questions or comments you might have regarding your own (future) DMP! Discussion round.

Use the distributed paper version or download the template from the UiT Research Data Portal:
https://uit.no/researchdata
1. INTRODUCTORY COURSE
Research data management at UiT: An introduction

2. MODULAR COURSES
How to search and cite research data
How to select an appropriate license
How to structure and document research data
How to share research data
How to write a data management plan

3. SPECIAL ISSUES
Quantitative data (in collaboration with NSD)

Classroom / Skype
Norwegian / English
GEN-8001: Take control of your PhD journey

Research data management
Part 2: Quantitative data

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

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/VDWPZ5, DataverseNO, V1, UNF:6:hiDJKeFSs1ZOccEs0dU0gw== [fileUNF]


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