TROLLing

The Tromsø Repository of Language and Linguistics

Archiving research data: Whys and hows

Meeting with AcqVA Aurora, 15 November 2021

Helene N. Andreassen & Leif Longva, UiT University Library

Outline
for this session

About TROLLing

- Key facts to repeat

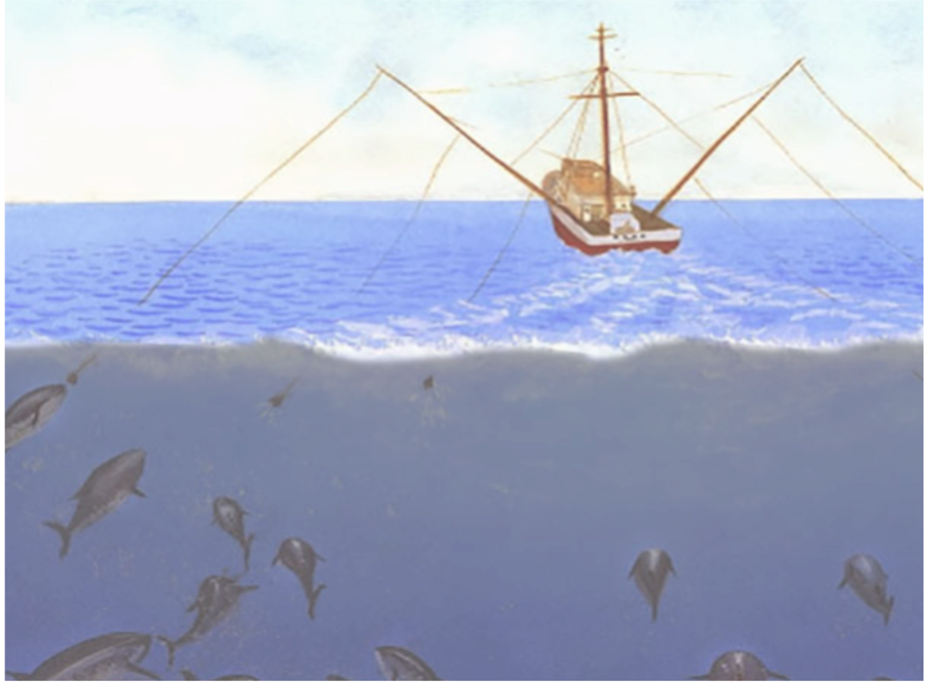
Work process

- Preparation
- Creation
- Submission
- Curation
- Revision
- Publishing

TROLLing: Scope

All subdisciplines of linguistics

International repository: Available to all for upload and download



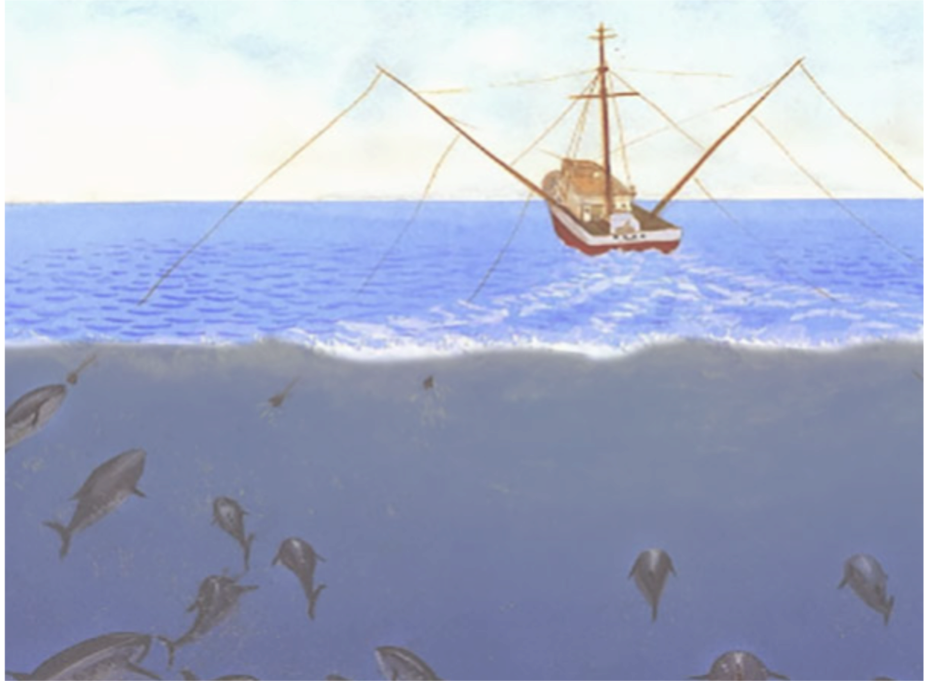
TROLLing: Scope

All types of data (but open)

- Raw data
- Processed data

All types of supplementary material

- Scripts
- Experimental protocol
- ...

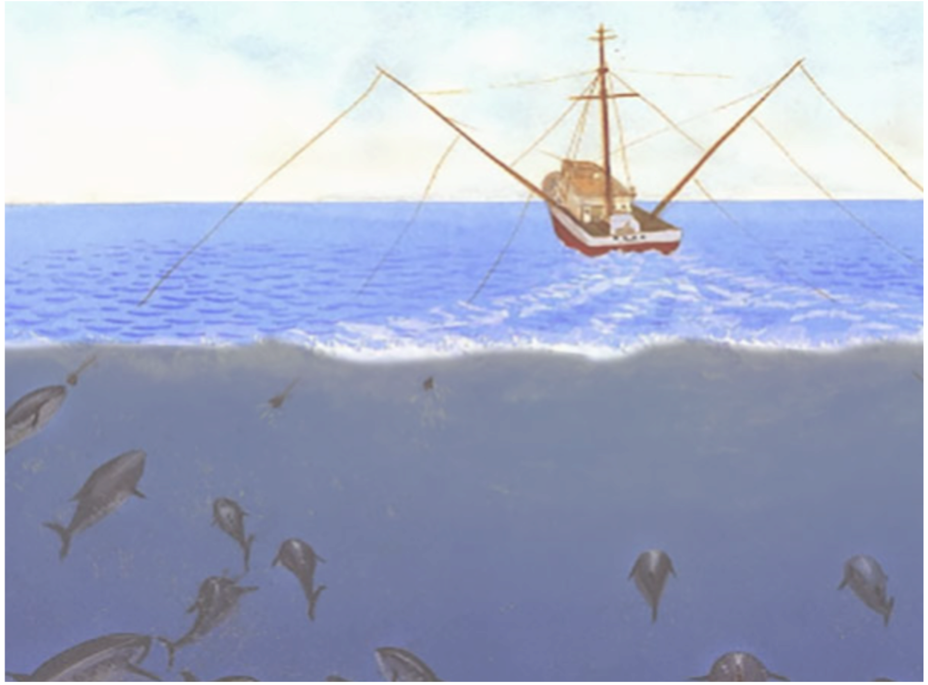


TROLLing: Infrastructure

Based on the international data platform Dataverse

Part of DataverseNO, a multi-institutional generic repository for open research data

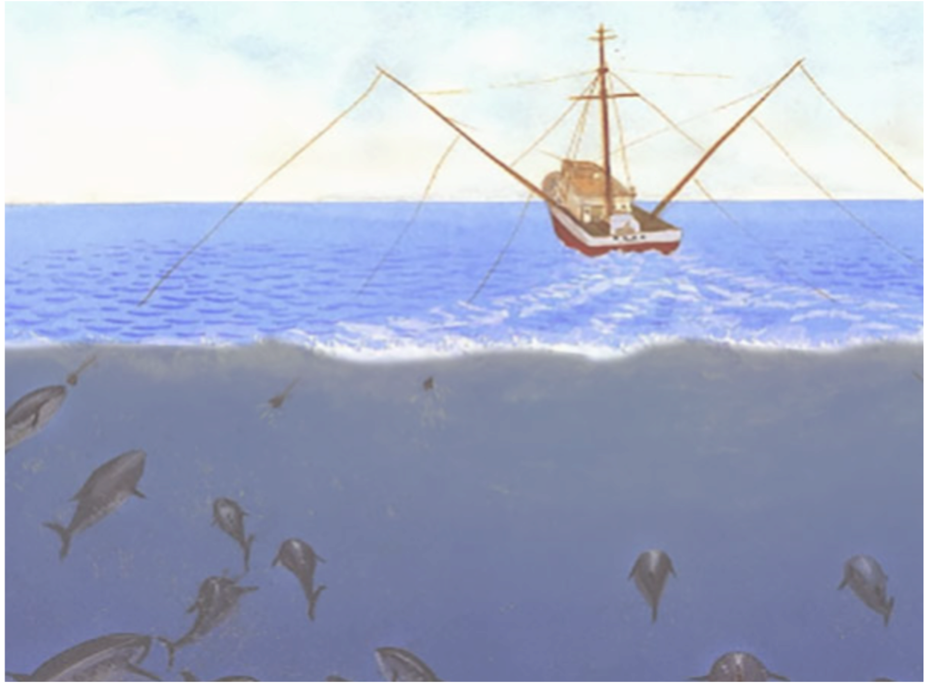
Developed and operated at UiT by the University Library and the IT Department



TROLLing: Infrastructure

Operated in line with the FAIR principles (Findable – Accessible – Interoperable – Reusable)

Since 2020, CoreTrustSeal certified as a sustainable and trusted research data repository



Archiving datasets in TROLLing

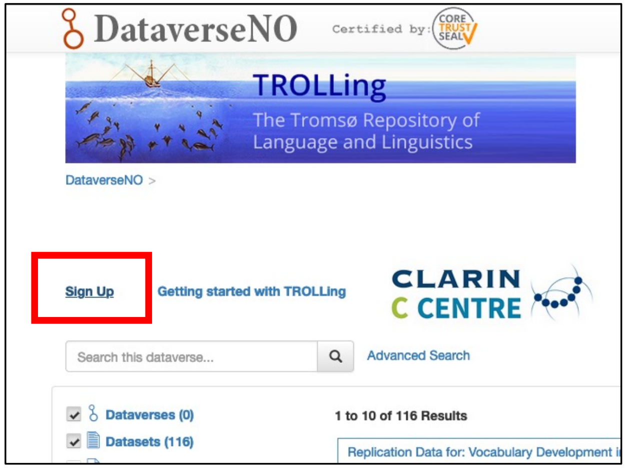
The work process

Preparation

Create TROLLing account
Read Deposit Guide
Write readme file
Check reuse restrictions
Format files

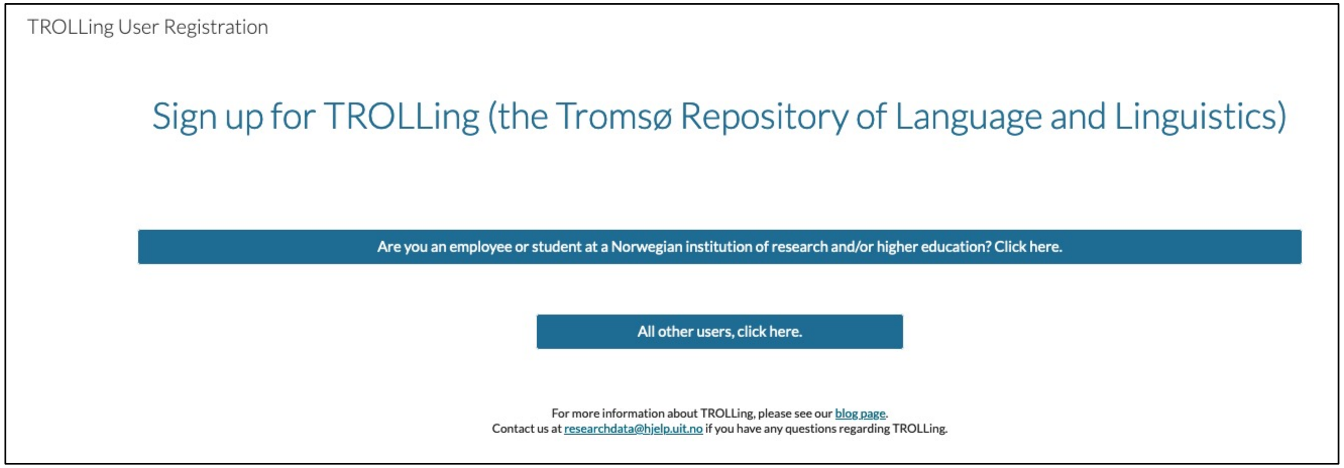
Creating a TROLLing account

All who wish to contribute with data need to be granted status as contributor.



1 to 10 of 116 Results

Replication Data for: Vocabulary Development i


TROLLing User Registration

Sign up for TROLLing (the Tromsø Repository of Language and Linguistics)

Are you an employee or student at a Norwegian institution of research and/or higher education? Click here.
All other users, click here.

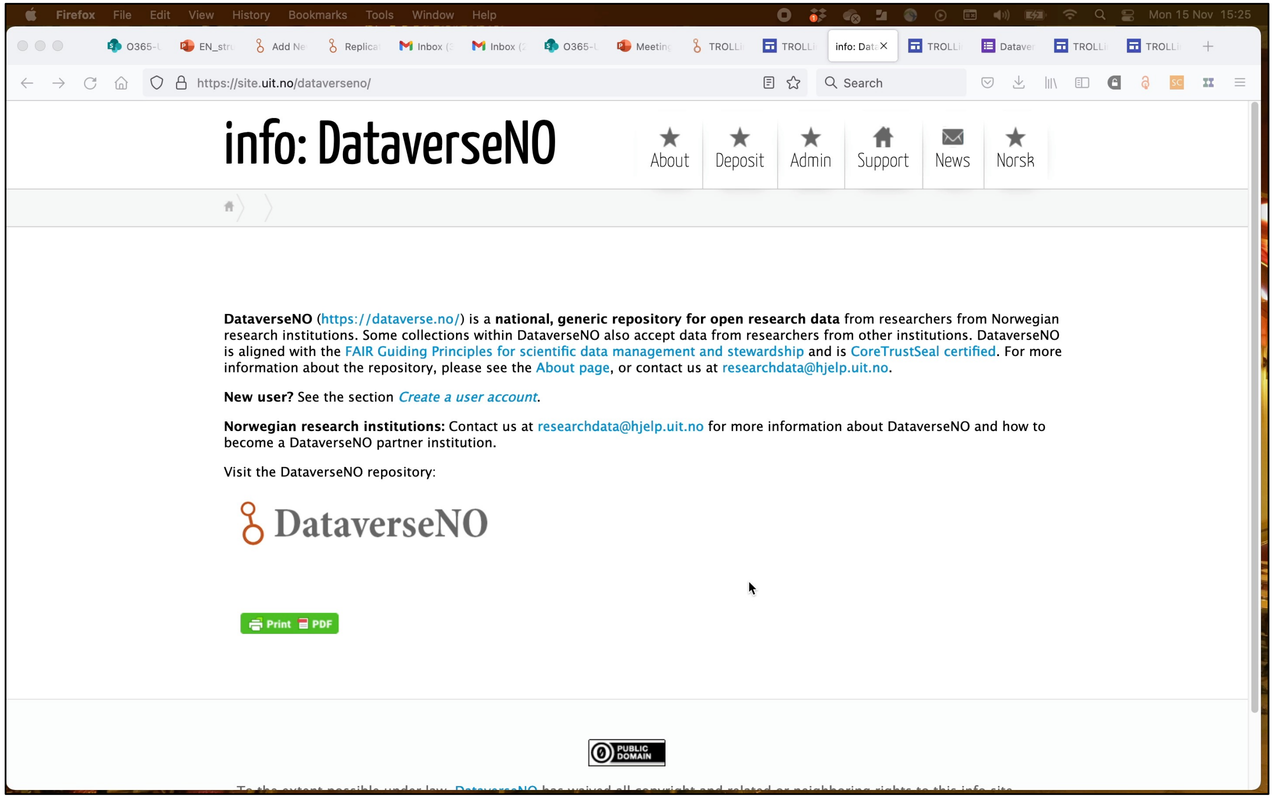
For more information about TROLLing, please see our blog page.
Contact us at researchdata@hjelp.uit.no if you have any questions regarding TROLLing.

TROLLing Deposit guide

Deposit guide common for the entire DataverseNO.

Useful to scroll through to see what needs to be prepared before your data can be deposited.

https://info.dataverse.no



DataverseNO (https://dataverse.no/) is a **national, generic repository for open research data** from researchers from Norwegian research institutions. Some collections within DataverseNO also accept data from researchers from other institutions. DataverseNO is aligned with the FAIR Guiding Principles for scientific data management and stewardship and is CoreTrustSeal certified. For more information about the repository, please see the About page, or contact us at researchdata@hjelp.uit.no.

New user? See the section *Create a user account*.

Norwegian research institutions: Contact us at researchdata@hjelp.uit.no for more information about DataverseNO and how to become a DataverseNO partner institution.

Visit the DataverseNO repository:

To the extent possible under law, DataverseNO has waived all copyright and related or neighboring rights to this info site

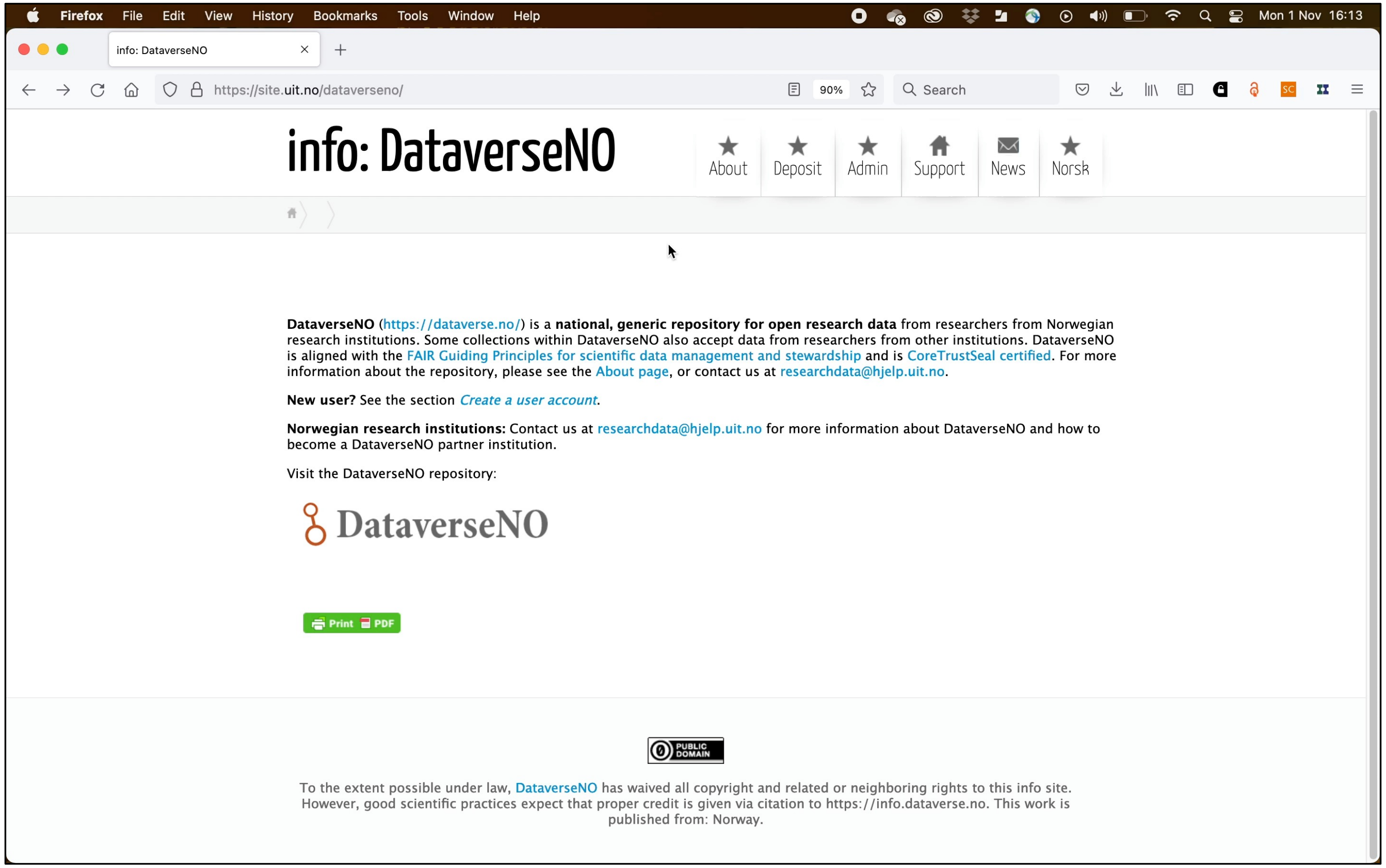
Readme file

- Enough information to understand correctly the data set
- A human readable introduction and description

- General information about the project
  - Title, project period, description, funding sources, project participants, contact information.
- About the dataset
  - License for reuse, related datasets, data sources.

- Methodology
  - Data collection, data processing, data quality etc.
- Content overview
  - File overview, relation between files, version of the dataset.
- Information about the data
  - Column headings, abbreviations, units of measure, contextual information for interpretation of the data set.

Readme file template



Firefox File Edit View History Bookmarks Tools Window Help
Mon 1 Nov 16:13
info: DataverseNO
https://site.uit.no/dataverseno/
90%
Search
About
Deposit
Admin
Support
News
Norsk

info: DataverseNO

DataverseNO (https://dataverse.no/) is a **national, generic repository for open research data** from researchers from Norwegian research institutions. Some collections within DataverseNO also accept data from researchers from other institutions. DataverseNO is aligned with the FAIR Guiding Principles for scientific data management and stewardship and is CoreTrustSeal certified. For more information about the repository, please see the About page, or contact us at researchdata@hjelp.uit.no.

New user? See the section *Create a user account*.

Norwegian research institutions: Contact us at researchdata@hjelp.uit.no for more information about DataverseNO and how to become a DataverseNO partner institution.

Visit the DataverseNO repository:

DataverseNO

https://info.dataverse.no

Readme file examples

Alaee, M., Rasekh-Mahand, M., & Tehrani-Doost, M. (2021). Replication Data for: Eye Behavior during Syntactic Movement Evidence for Processing Approach to Persian Syntax. https://doi.org/10.18710/TZBAOR, *DataverseNO*, V1.

Lybaert, C., De Clerck, B., Saelens, J., & De Cuypere, L. (2021). Replication Data for: A Corpus Based Analysis of V2 Variation in West Flemish and French Flemish Dialects. https://doi.org/10.18710/NSFN2B, *DataverseNO*, V1.

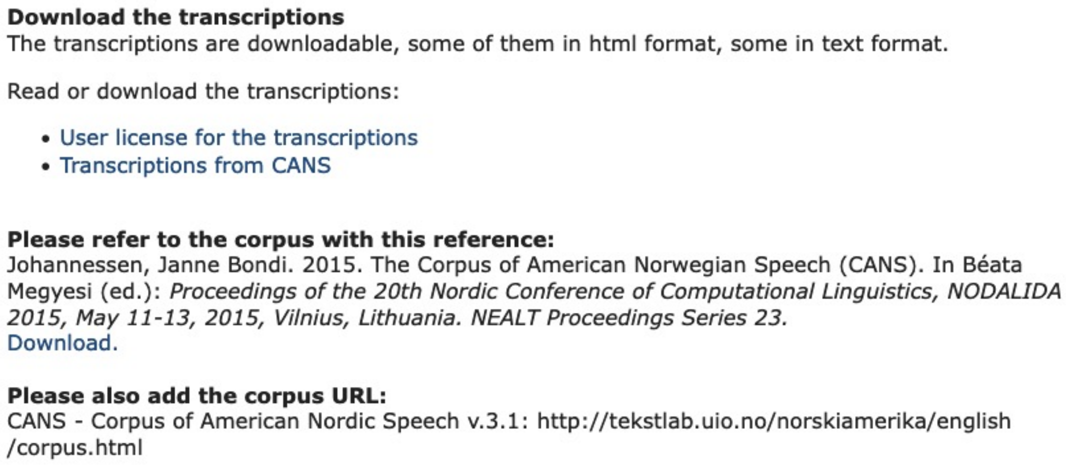
Wang, X. (2021). Replication Data for: Sound symbolism in Chinese children's literature. https://doi.org/10.18710/DCAFEP, *DataverseNO*, V1.

Reuse restrictions

You are responsible for ensuring that the data uploaded to TROLLing can be shared

GDPR

Third-party data



Download the transcriptions
The transcriptions are downloadable, some of them in html format, some in text format.

Read or download the transcriptions:

- User license for the transcriptions
- Transcriptions from CANS

Please refer to the corpus with this reference:
Johannessen, Janne Bondi. 2015. The Corpus of American Norwegian Speech (CANS). In Béata Megyesi (ed.): *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. NEALT Proceedings Series 23.*
Download.

Please also add the corpus URL:
CANS - Corpus of American Nordic Speech v.3.1: http://tekstlab.uio.no/norskiamerika/english/corpus.html

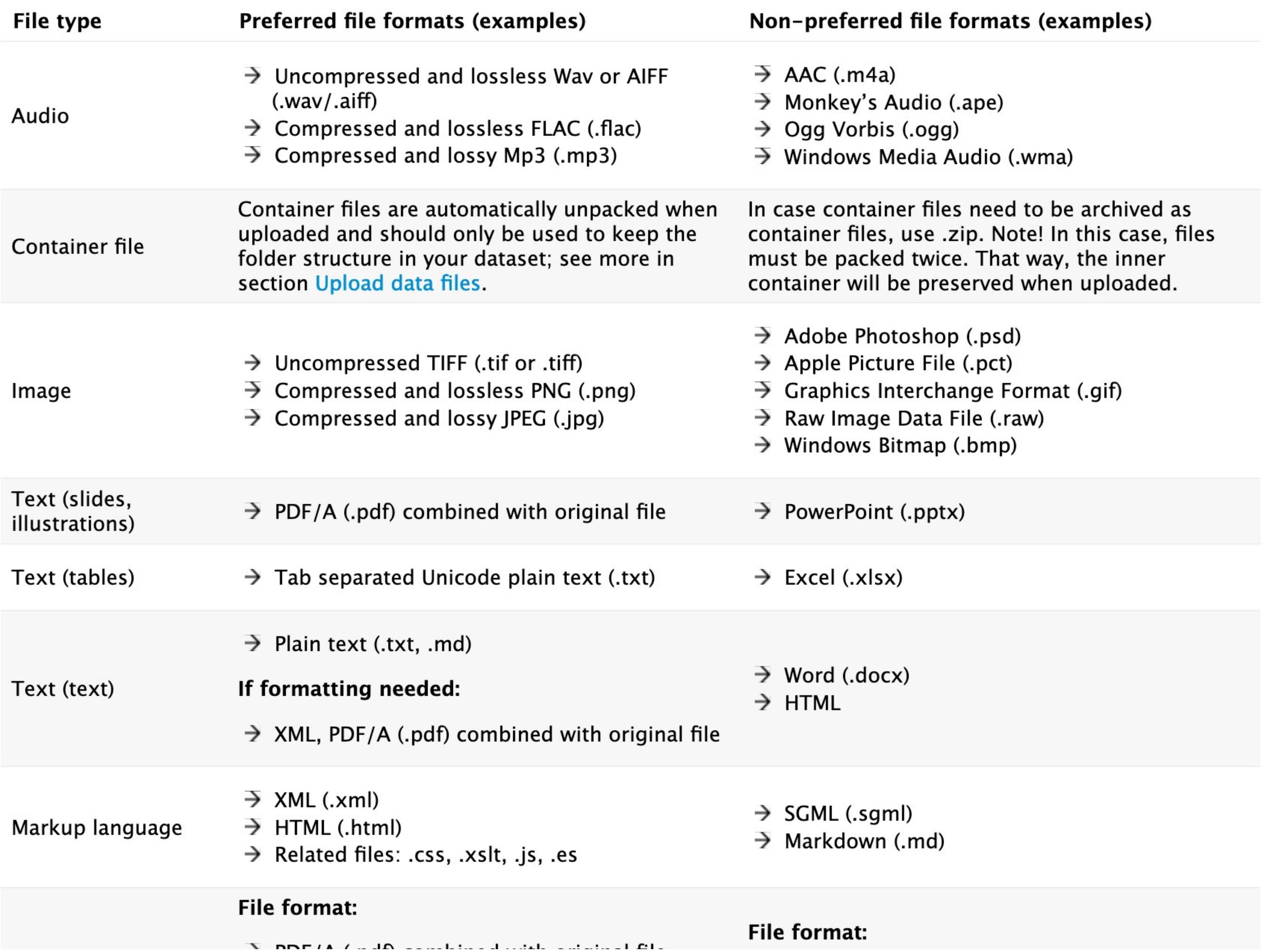
File formatting

Characteristics of preferred file formats

- non-proprietary
- open, with documented international standards
- in common usage by the research community
- using standard character encodings (e.g. ASCII, UTF-8)
- uncompressed (space permitting)

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

File formatting



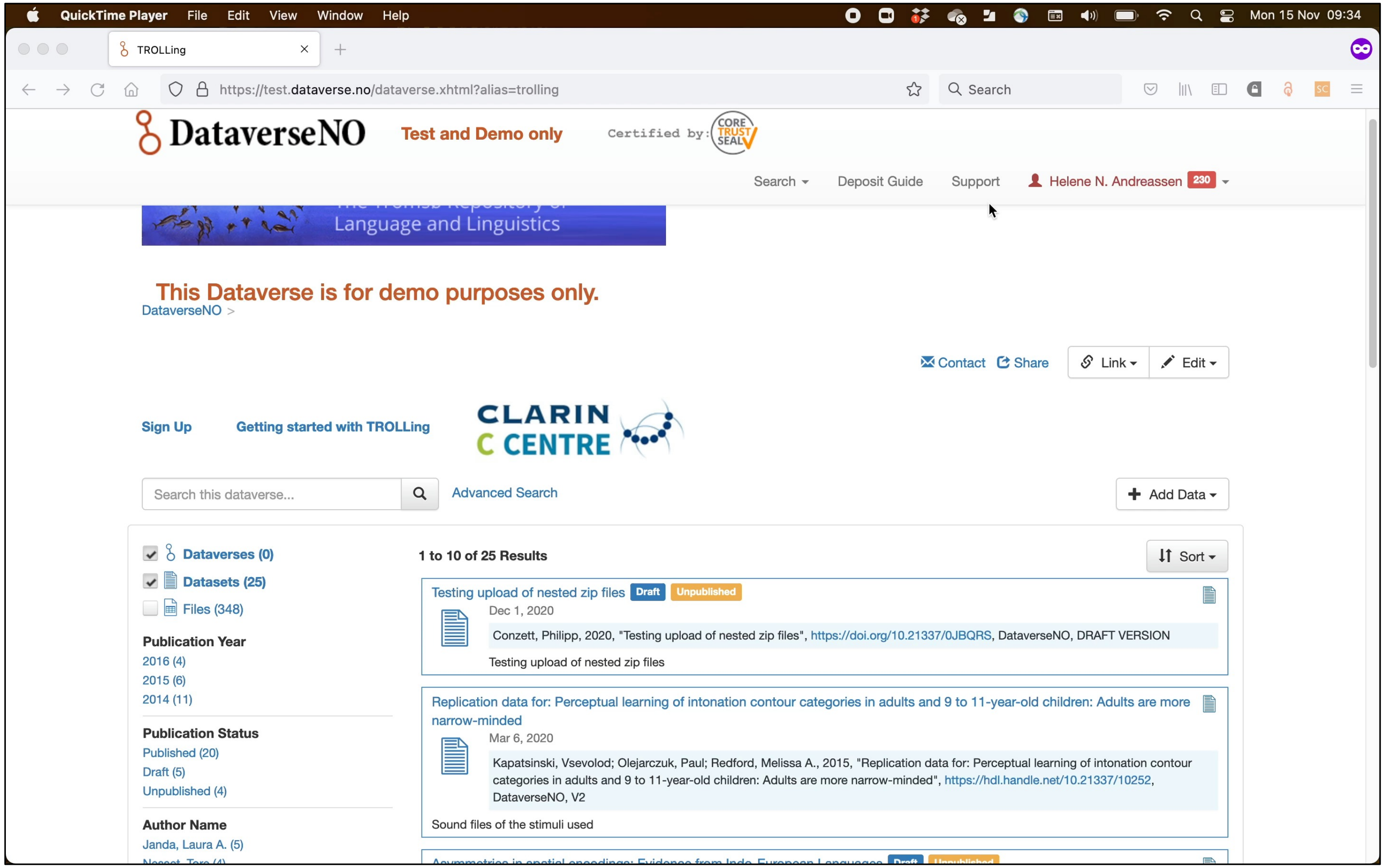
File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Audio	→ Uncompressed and lossless Wav or AIFF (.wav/.aiff) → Compressed and lossless FLAC (.flac) → Compressed and lossy Mp3 (.mp3)	→ AAC (.m4a) → Monkey's Audio (.ape) → Ogg Vorbis (.ogg) → Windows Media Audio (.wma)
Container file	Container files are automatically unpacked when uploaded and should only be used to keep the folder structure in your dataset; see more in section Upload data files.	In case container files need to be archived as container files, use .zip. Note! In this case, files must be packed twice. That way, the inner container will be preserved when uploaded.
Image	→ Uncompressed TIFF (.tif or .tiff) → Compressed and lossless PNG (.png) → Compressed and lossy JPEG (.jpg)	→ Adobe Photoshop (.psd) → Apple Picture File (.pct) → Graphics Interchange Format (.gif) → Raw Image Data File (.raw) → Windows Bitmap (.bmp)
Text (slides, illustrations)	→ PDF/A (.pdf) combined with original file	→ PowerPoint (.pptx)
Text (tables)	→ Tab separated Unicode plain text (.txt)	→ Excel (.xlsx)
Text (text)	→ Plain text (.txt, .md) **If formatting needed:** → XML, PDF/A (.pdf) combined with original file	→ Word (.docx) → HTML
Markup language	→ XML (.xml) → HTML (.html) → Related files: .css, .xslt, .js, .es	→ SGML (.sgml) → Markdown (.md)
	File format: → PDF/A (.pdf) combined with original file	**File format:**

https://site.uit.no/dataverseno/deposit/prepare/#what-are-preferred-file-formats

Creation

Add metadata
Upload data files
Inform about anonymisation (if applicable)

Metadata: Data about data



This Dataverse is for demo purposes only.

DataverseNO >

☑ Dataverses (0)
☑ Datasets (25)
□ Files (348)

Publication Year
2016 (4)
2015 (6)
2014 (11)

Publication Status
Published (20)
Draft (5)
Unpublished (4)

Author Name
Janda, Laura A. (5)
Nesset, Tore (4)

1 to 10 of 25 Results

Testing upload of nested zip files Draft Unpublished
Dec 1, 2020
Conzett, Philipp, 2020, "Testing upload of nested zip files", https://doi.org/10.21337/0JBQRS, DataverseNO, DRAFT VERSION
Testing upload of nested zip files

Replication data for: Perceptual learning of intonation contour categories in adults and 9 to 11-year-old children: Adults are more narrow-minded
Mar 6, 2020
Kapatsinski, Vsevolod; Olejarczuk, Paul; Redford, Melissa A., 2015, "Replication data for: Perceptual learning of intonation contour categories in adults and 9 to 11-year-old children: Adults are more narrow-minded", https://hdl.handle.net/10.21337/10252, DataverseNO, V2
Sound files of the stimuli used

Asymmetries in spatial encodings: Evidence from Indo-European Languages Draft Unpublished

[We go live]

Add more metadata

Upload files

File naming

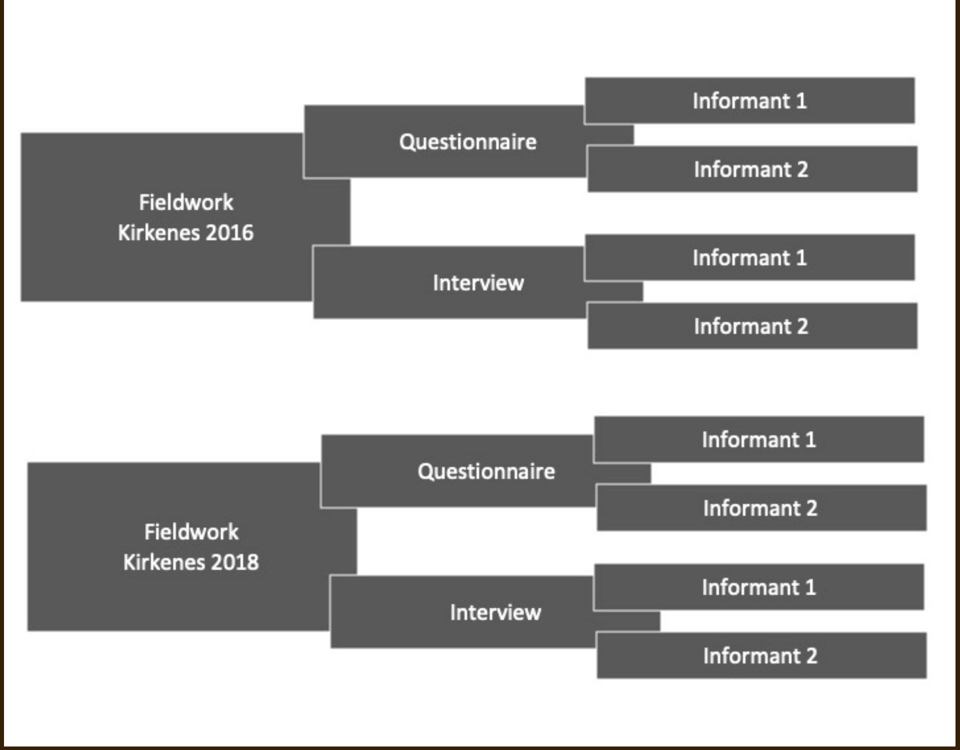
DATE
2016_interview_Inf1
2016_questionnaire_Inf1
2018_interview_Inf1
2018_questionnaire_Inf1

TYPE
Interview_Inf1_2016
Interview_Inf2_2016
Questionnaire_Inf1_2016
Questionnaire_Inf2_2016

INFORMANT (OR TOPIC)
Inf1_interview_2016
Inf1_questionnaire_2016
Inf2_interview_2016
Inf2_questionnaire_2016

FORCED ORDER
01_Overview_fieldwork_2016
02_Inf1_questionnaire_2016
02_Inf1_interview_2016
03_Inf1_questionnaire_2018

Folder structuring

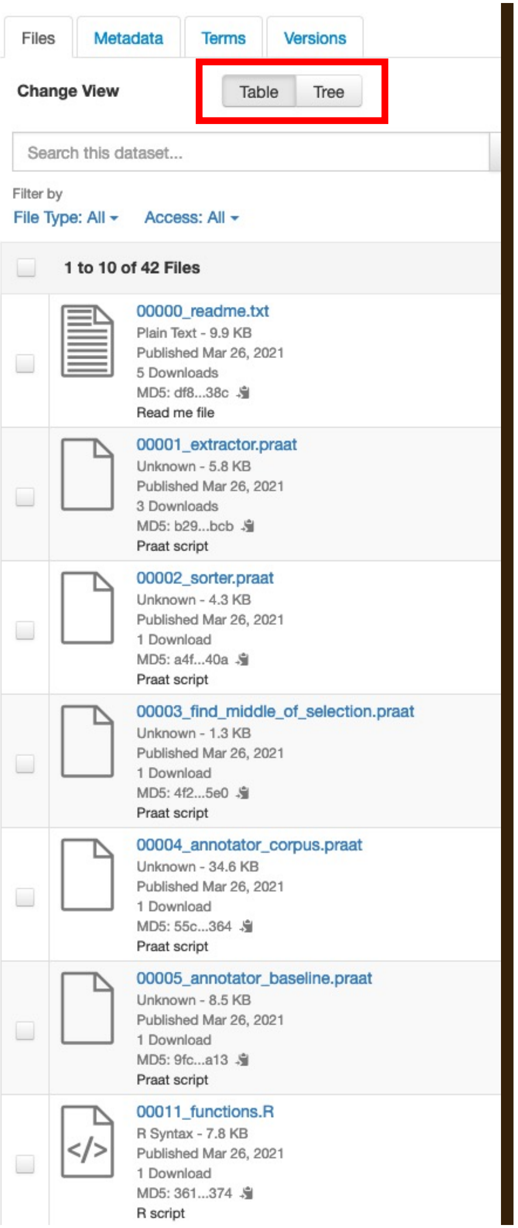


Creating a folder structure in TROLLing:
https://site.uit.no/dataverseno/deposit/deposit-your-data/#upload-data-files

Folder structuring

Example:
Strand, Bror-Magnus S., 2021, "Replication data for: Playing with fire compounds", https://doi.org/10.18710/09GQFO, DataverseNO, V1, UNF:6:myhi9R8BuhLnHuc+kHflSw== [fileUNF]

Creating a folder structure in TROLLing:
https://site.uit.no/dataverseno/deposit/deposit-your-data/#upload-data-files



1 to 10 of 42 Files

□ 00000_readme.txt
Plain Text - 9.9 KB
Published Mar 26, 2021
5 Downloads
MD5: df8...38c
Read me file

□ 00001_extractor.praat
Unknown - 5.8 KB
Published Mar 26, 2021
3 Downloads
MD5: b29...bcb
Praat script

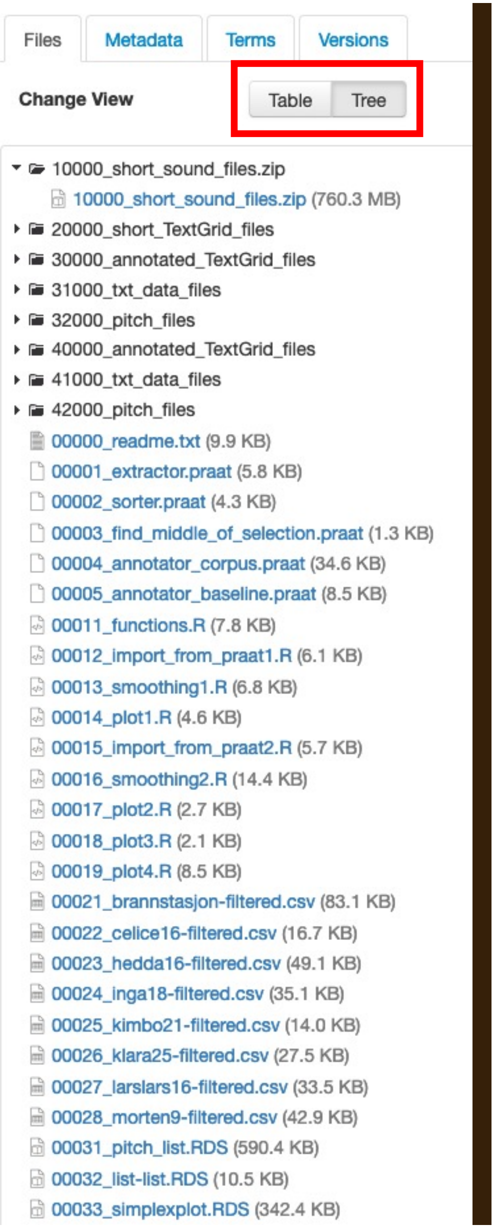
□ 00002_sorter.praat
Unknown - 4.3 KB
Published Mar 26, 2021
1 Download
MD5: a4f...40a
Praat script

□ 00003_find_middle_of_selection.praat
Unknown - 1.3 KB
Published Mar 26, 2021
1 Download
MD5: 4f2...5e0
Praat script

□ 00004_annotator_corpus.praat
Unknown - 34.6 KB
Published Mar 26, 2021
1 Download
MD5: 55c...364
Praat script

□ 00005_annotator_baseline.praat
Unknown - 8.5 KB
Published Mar 26, 2021
1 Download
MD5: 9fc...a13
Praat script

□ 00011_functions.R
R Syntax - 7.8 KB
Published Mar 26, 2021
1 Download
MD5: 361...374
R script

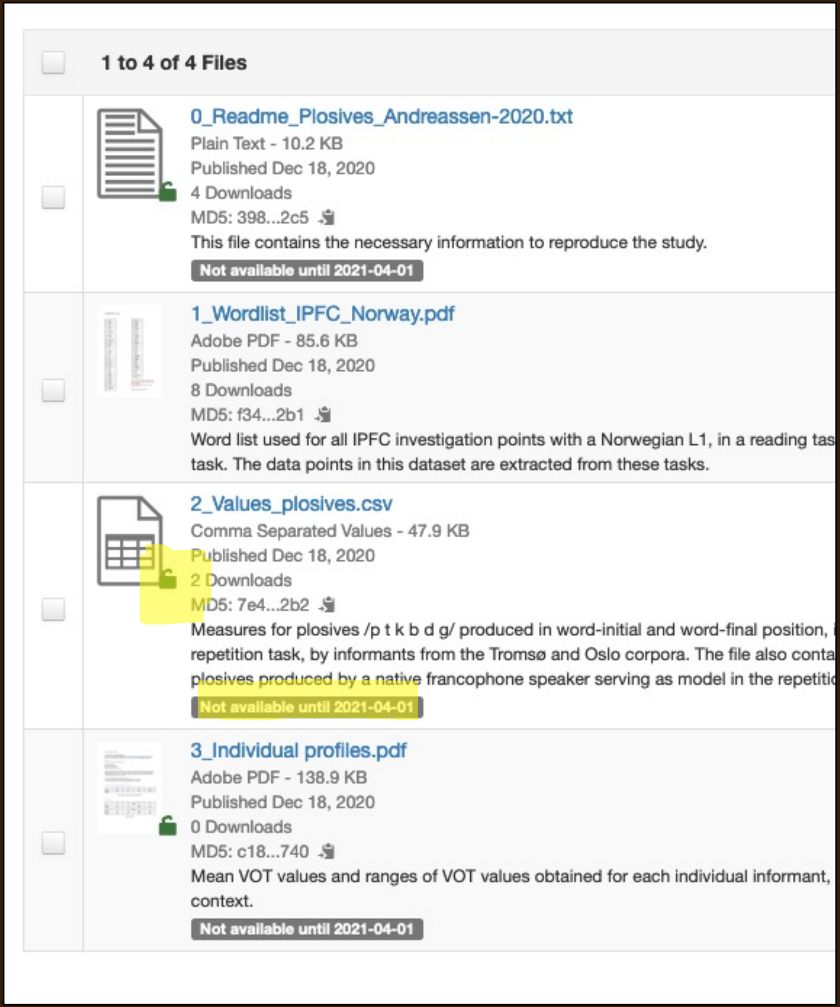


- 10000_short_sound_files.zip
  - 10000_short_sound_files.zip (760.3 MB)
- 20000_short_TextGrid_files
- 30000_annotated_TextGrid_files
- 31000_txt_data_files
- 32000_pitch_files
- 40000_annotated_TextGrid_files
- 41000_txt_data_files
- 42000_pitch_files
- 00000_readme.txt (9.9 KB)
- 00001_extractor.praat (5.8 KB)
- 00002_sorter.praat (4.3 KB)
- 00003_find_middle_of_selection.praat (1.3 KB)
- 00004_annotator_corpus.praat (34.6 KB)
- 00005_annotator_baseline.praat (8.5 KB)
- 00011_functions.R (7.8 KB)
- 00012_import_from_praat1.R (6.1 KB)
- 00013_smoothing1.R (6.8 KB)
- 00014_plot1.R (4.6 KB)
- 00015_import_from_praat2.R (5.7 KB)
- 00016_smoothing2.R (14.4 KB)
- 00017_plot2.R (2.7 KB)
- 00018_plot3.R (2.1 KB)
- 00019_plot4.R (8.5 KB)
- 00021_brannstasjon-filtered.csv (83.1 KB)
- 00022_celice16-filtered.csv (16.7 KB)
- 00023_hedda16-filtered.csv (49.1 KB)
- 00024_inga18-filtered.csv (35.1 KB)
- 00025_kimbo21-filtered.csv (14.0 KB)
- 00026_klara25-filtered.csv (27.5 KB)
- 00027_larslars16-filtered.csv (33.5 KB)
- 00028_morten9-filtered.csv (42.9 KB)
- 00031_pitch_list.RDS (590.4 KB)
- 00032_list-list.RDS (10.5 KB)
- 00033_simplexplot.RDS (342.4 KB)

Embargo?

Possibility to restrict access to data files for a certain period of time.

This is something you can administer when you upload data files.



☐ 1 to 4 of 4 Files

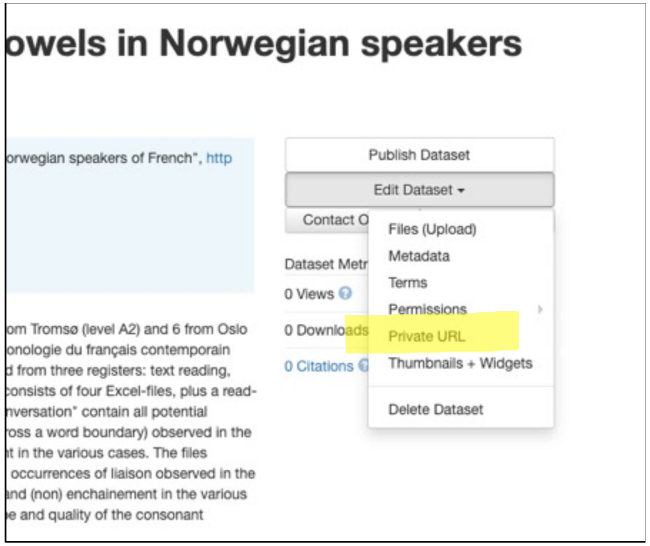
☐ 0_Readme_Plosives_Andreassen-2020.txt
Plain Text - 10.2 KB
Published Dec 18, 2020
4 Downloads
MD5: 398...2c5
This file contains the necessary information to reproduce the study.
Not available until 2021-04-01

☐ 1_Wordlist_IPFC_Norway.pdf
Adobe PDF - 85.6 KB
Published Dec 18, 2020
8 Downloads
MD5: f34...2b1
Word list used for all IPFC investigation points with a Norwegian L1, in a reading tas
task. The data points in this dataset are extracted from these tasks.

☐ 2_Values_plosives.csv
Comma Separated Values - 47.9 KB
Published Dec 18, 2020
2 Downloads
MD5: 7e4...2b2
Measures for plosives /p t k b d g/ produced in word-initial and word-final position, i
repetition task, by informants from the Tromsø and Oslo corpora. The file also conta
plosives produced by a native francophone speaker serving as model in the repetitio
Not available until 2021-04-01

☐ 3_Individual profiles.pdf
Adobe PDF - 138.9 KB
Published Dec 18, 2020
0 Downloads
MD5: c18...740
Mean VOT values and ranges of VOT values obtained for each individual informant,
context.
Not available until 2021-04-01

Anonymisation



owels in Norwegian speakers

orwegian speakers of French", http

om Tromsø (level A2) and 6 from Oslo
onologie du français contemporain
d from three registers: text reading,
consists of four Excel-files, plus a read-
nversation" contain all potential
oss a word boundary) observed in the
t in the various cases. The files
occurrences of liaison observed in the
nd (non) enchainement in the various
e and quality of the consonant

- Submission of manuscript for double-blind peer review?

- In the *Related Publication* field:
    - Inform that you wish an anonymised version of your dataset.

- Once the dataset is curated:
    - We create an **anonymised version** with a **private URL** that you may send to the editors alongside your manuscript.
    - NB! Your dataset is NOT published yet, and may still be modified (or deleted).

Submission

When your dataset is ready, you click *Submit for review*

Allow a few days for the data curation process.

Curation

TROLLing curators: Helene N. Andreassen & Philipp Conzett

(Part of a larger group in the library working on research data management.)

Checking of

Metadata
Readme file
File format
License control

Creation of curator report with suggested modifications

Revision

Make recommended changes
Re-submit dataset

Rarely more than 3 rounds before all details are in place

Publishing

Only curators can publish datasets.

All modifications after initial publication need to go through/be approved by us.

- this includes removal of embargo on files.

For anonymised datasets, inform us when the (non-anonymised) dataset can be published.

Publishing

To improve transparency, you should

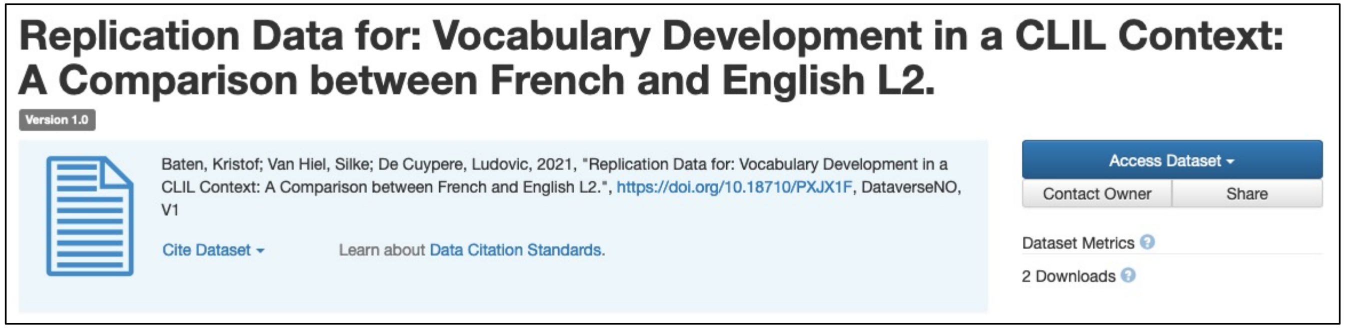
- link to your published dataset via a data availability statement

- cite it in your own paper the way you cite other sources

Citation of datasets

Citing the dataset in the paper allows the reader to easily access it.

Each TROLLing dataset comes with a fully fledged bibliographic reference.



Replication Data for: Vocabulary Development in a CLIL Context: A Comparison between French and English L2.

Baten, Kristof; Van Hiel, Silke; De Cuypere, Ludovic, 2021, "Replication Data for: Vocabulary Development in a CLIL Context: A Comparison between French and English L2.", https://doi.org/10.18710/PXJX1F, DataverseNO, V1

Want to learn more about data citation in linguistics? Check out the *Tromsø Recommendations for citation of research data in linguistics* (https://doi.org/10.15497/rda00040)

Plans

- AcqVA Aurora data as a sub-collection within TROLLing

- Own page with TROLLing info on the AcqVA Aurora webpages
  - Links
  - Templates
  - Tutorials

- Move TROLLing to its own installation (outside of DataverseNO)
- Market it to scientific publishers as a trustworthy, discipline-specific repository

Repository

trolling.uit.no

User guide

info.dataverse.no

Info

info.trolling.uit.no

Support

researchdata@hjelp.uit.no

You may also contact Helene or Philipp directly.

TROLLing

The Tromsø Repository of Language and Linguistics

Archiving research data: Whys and hows

Meeting with AcqVA Aurora, 15 November 2021

Helene N. Andreassen & Leif Longva, UiT University Library