UiT
THE ARCTIC UNIVERSITY OF NORWAY

TROLLing

The Tromsø Repository of Language and Linguistics

MONS

Scope and operation of an open repository for linguistic data

MØTE OM NORSK SPRÅK (MONS) 17, 22-24 NOVEMBER 2017, OS, BERGEN

Helene N. Andreassen, Philipp Conzett, Stein Høydalsvik, Laura Janda,
Leif Longva, Tore Nesset & Odu Obiajulu

ARE YOU ABOUT TO PUBLISH YOUR RESEARCH, AND WANT TO MAKE YOUR DATA AND CODE AVAILABLE, ALONGSIDE THE PAPER? THE TROMSØ REPOSITORY OF LANGUAGE AND LINGUISTICS (TROLLING) AIMS TO MEET THE LINGUISTIC COMMUNITY'S INCREASING WISH AND DEMAND FOR ARCHIVING, PUBLISHING, AND DISSEMINATION OF SCIENTIFIC RESEARCH DATA, AS WELL AS PROPER ATTRIBUTION TO THEM.

The raison d'être of TROLLing

Open science
Transparency, accessibility, and reusability of research, using web-based tools.

Sharing of research data is increasingly encouraged/required by many funders and publishers.

As open as possible, as closed as necessary
(The new Guidelines on Findable, Accessible, Interoperable and Reusable (FAIR) Data Management, European Commission 2016)

Where the data is not publicly available, (and there are often good reasons why it is not), authors should explain why it is not made available.
(Notes to Contributors, Language, LSA)

Tromsø in the forefront by joining forces

- Initiative taken by the internationally oriented UiT linguistics community.
- The University Library a natural partner, having a long tradition with open access archiving and publishing, and open access infrastructure.
- **TROLLing** developed in less than a year (launch June 2014), by a working group consisting of researchers in linguistics, and specialists within linguistics, open access and system development at the library.
- Development guided by scientific needs and international standards, with a three-member scientific advisory board contributing to overarching strategic discussions.
- Long-term preservation of data ensured by the Department of Information Technology.

Encouraging the researchers to publish their data: measures taken

- To incite a change, consciousness-raising and support are key elements in our contact with the linguistic research community.
- Outreach is carried out by faculty and other trained linguists at workshops, scientific conferences, and in contact with journal editors.
- Support services are provided by the University Library, including curation of datasets, individual training, and online teaching material.

Areas of use and reuse

- **TROLLing** data are used in a wide range of publications: Journal articles, edited book chapters, conference proceedings, and PhD and Master's theses.
- **TROLLing** data are already reused to generate new findings, e.g. in another person's doctoral work, where the original method has been extended and applied to new data.
- **TROLLing** data are further used in educational settings, to illustrate the language phenomenon under discussion.

The whos and whats of TROLLing

- Available to linguists world wide for upload (which requires registration) and download.
- Archive for **open** data, although some data files may be locked for a certain period.
- Archive for open **structural** data, annotated and organized to serve as empirical basis in linguistic research.

... corpus concordances from a diachronic study of German nominalization patterns.
hdl:10037.1/10285

... [video] recordings were produced for the Artifon project as part of visual illustrations for students learning Norwegian (phonetics).
hdl:10037.1/10056

... new experimental data on the acquisition of structures involving ditransitive verbs in two East Slavic languages: Russian and Ukrainian.
doi:10.5072/FK2/VA3BVU

Structure

- Built on Dataverse, an open source platform from Harvard University.
- User-friendly interface, with metadata templates based on **international standards** in compliance with DataCite.
- Linking to related publications, or raw or primary data stored elsewhere, possible through metadata registration.
- Data can be shared with colleagues and journal editors prior to publication via the feature Private URL.
- In-depth file descriptions required to ensure comprehensibility.
- Persistent file formats required to ensure accessibility.
- In 2018, TROLLing will apply for the CoreTrustSeal, which will warrant the archive's trustworthiness within the scientific community.

Visibility and retrieval

- Each dataset is automatically given a persistent identifier (DOI).
- **Cross-referencing:** The research paper contains a citation to the TROLLing post, and the research paper citation is part of the TROLLing metadata.
- **TROLLing** being part of a global open network, data are visible in DataCite, and will be included in other major search engines (Google Scholar, library indexes).
- **TROLLing** is registered in repository indexes, e.g. re3data.org, which improves the chances of potential users discovering it.

Attribution

- Sharing/reusing data according to **best practice** demands clear rules and guidelines, including an international license attached to the data.
- In the case of reuse, others must refer to the data in line with good academic practice.
- In **TROLLing**, a dataset citation string is automatically generated, based on the registered metadata.

Holliday, Jeffrey J.; Turnbull, Rory; Eychenne, Julien, 2016, "K-SPAN (Korean Surface Phones and Neighborhoods)", doi:10.18710/TWM79F, Neighborhoods)", doi:10.18710/TWM79F, DataverseNO, V2, UNF:6:NWbRmiBvO5wWcDN2QHCQJw==

Version control

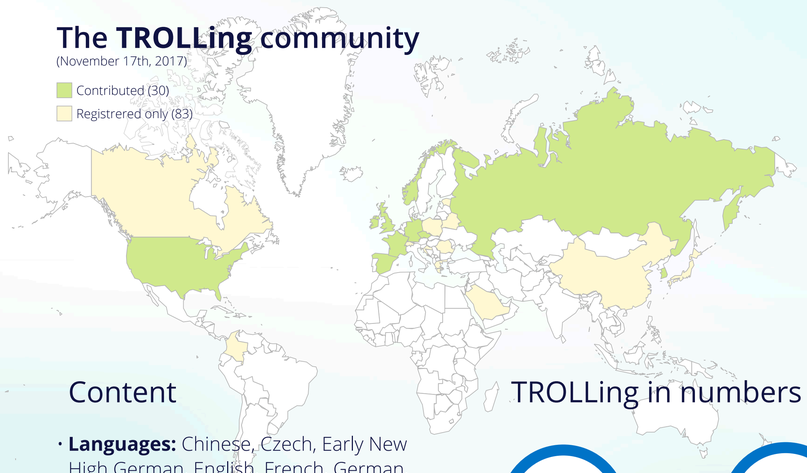
- Datasets can be updated at any time by the contributors themselves.
- When a new version is published, the dataset citation string is automatically updated to include the new version number.
- Old versions remain accessible.
- All updates are documented to facilitate between-version comparisons.

2.1	**Citation Metadata:** Keyword (8 Changed); Contact (1 Changed); **Additional Citation Metadata:** (4 Changed); View Details
2.0	**Files (Added: 1; Removed: 1);** View Details
1.0	This is the first published version.

hdl:10037.1/10294

The **TROLLing** community

(November 17th, 2017)



Content

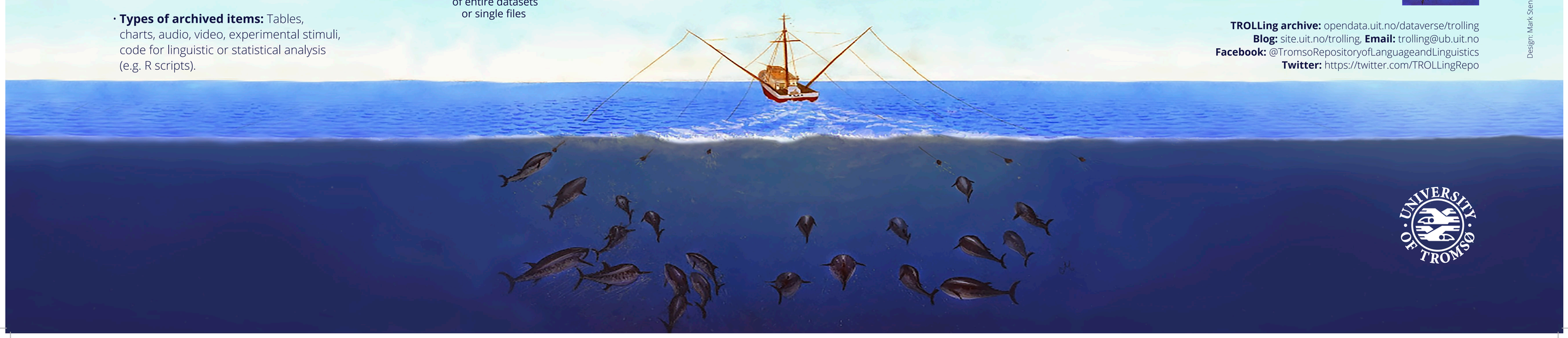
- **Languages:** Chinese, Czech, Early New High German, English, French, German, Icelandic, Korean, Latin, Latvian, Middle Russian, Nizaa, North Saami, Norwegian, Olbanian, Old Church Slavonic, Old East Slavic, Romanian, Russian, Spanish, Ukrainian.
- **Fields of study:** Phonetics, phonology, morphology, syntax, semantics, lexicon (synchronic, diachronic, acquisition, language technology, sociolinguistics).
- **Types of archived items:** Tables, charts, audio, video, experimental stimuli, code for linguistic or statistical analysis (e.g. R scripts).

TROLLing in numbers

60 Published datasets

560 Files

3503 Downloads of entire datasets or single files



TROLLing
The Tromsø Repository of Language and Linguistics

TROLLing archive: opendata.uit.no/dataverse/trolling
Blog: site.uit.no/trolling, **Email:** trolling@ub.uit.no
Facebook: @TromsoRepositoryofLanguageandLinguistics
Twitter: https://twitter.com/TROLLingRepo

UNIVERSITY OF TROMSØ