TROLLing defining, building, and operating an open archive for linguistic data

Helene N. Andreassen, PhD
UiT The Arctic University of Norway

2nd Workshop on Standards for Data Citation and Attribution in Linguistics

University of Texas at Austin, April 8-10, 2016
Outline

A variety of competencies and skills

Idea
Ambition

Cooperation

Operational service
UiT and open access

UiT will be recognized by a culture for active dissemination through open channels for publishing, as well as through exhibitions, journals and the media.

UiT strategic plan 2020
Digital – above all!

Main ambition Explore and develop the digital possibilities, and use these to strengthen our services to employees and students

Strategy Take a central position in the work with archiving and dissemination of research data, locally and nationally

University Library strategic plan 2020

Photo: Rune Ytreberg
How TROLLing came to be

1. Inquiry in 2013 from the UiT linguistics community turns the library’s ambition to work on open research data into action

2. Working group put together, consisting of researchers and subject specialists in linguistics, OA specialists and system developers

3. Establishment of a three-member scientific advisory board

4. Development guided by scientific needs and international solutions

5. Launch in June 2014

Photos: flickr.com/photos/kimgskytte
TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

• International service, open to researchers across the world for upload and download

• Maintained and curated by the University Library
  – Relevance of uploaded data
  – Quality and comprehensiveness of metadata
  – Description and format of uploaded data

opendata.uit.no
The platform

Guiding principle: Be future-oriented, and think bigger!

TROLLing built on the Dataverse platform (https://dataverse.harvard.edu/)
- Allows adding of other types of datasets using the same tools and templates
- Facilitates harvesting of data by international services
- Complies with DataCite (https://www.datacite.org/)
Adaption of metadata template
How to optimize retrieval of data

- Topic specification
  - Field
  - Time-depth
  - Topic
  - Free-text keywords

- Description
  - Abstract
  - File content
Setting of requirements
How to optimize reuse of data

• Description
  – In template
  – In read-me file
  – In data file (column headings, file name, etc.)

• Persistent file format
  – Non-proprietary
  – Open
  – Standard character encoding (UTF-8)
Update to come

Dataverse Version 4

- Recently released
- Trolling in the process of being migrated

- Important improvements
  - Richer and more flexible metadata template
  - Tagging on file level, improving the search function
  - Improved metrics: views, downloads, citations, shares
**Citing the data**

**Built-up of dataset citation**

- Persistent identifier
  - Doi shortly available

- Data description
  - “Replication data” or other

- Version indicator
  - Previous versions accessible

**Requirements on reuse of data**

- Standard license selected: CC0
  - Meet the potential problem of attribution stacking

- Citation in line with good academic practice
  - Use the reference as provided
  - (Add subset info if appropriate)

---

If you use these data, please add the following citation to your scholarly references. Why cite?


Citation Format

Results found in this publication can be replicated using these data.

Original Publication

TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

• International service, open to researchers across the world for upload and download

• Maintained and curated by the University Library at UiT

• Assisting and educating the users
  – User guides
  – Instruction videos

  ➢ Blog interface for communication
  ➢ Cooperation with faculty

The TROLLing Blog

Getting started

User guides:
- TROLLing essentials (pdf)
- TROLLing data format essentials (pdf)

Instruction videos:
- How to create an account (YouTube)
- How to create a study (YouTube)
- How to upload data (YouTube)
- How to search for data (YouTube)

site.uit.no/trolling
TROLLing
The Tromsø Repository of Language and Linguistics

Archive for open linguistic data and statistical code

• International service, open to researchers across the world for upload and download

• Maintained and curated by the University Library at UiT

• Development of user guides and promotional material in cooperation with faculty

• Marketing in every channel possible
  – Promotion material
  – YouTube

➤ Cooperation with faculty, graphic designers and video producers

https://www.youtube.com/watch?v=uEf0c0NT9_A
Outreach
Conferences and meetings: presentations and workshops

Laura Janda
• Palatalisation Workshop, CASTL/UiT, 2014.

Helene N. Andreassen
• Journées FLOraL (Français Langue ORAle et Linguistique), Paris, 2014.

Philipp Conzett & Leif Longva
• emtacl15 - emerging technologies in academic libraries, Trondheim, 2015.

Philipp Conzett & Obiajulu Odu
Outreach
Approaching the publishers

- Encouraging from above
  - Journal editorial boards
    - put TROLLing into guidelines
  - Cristin (Norwegian National Research Information System)
    - create a category “data collection” or “dataset”/make it count

- Encouraging from below
  - Networks (TROLLing team and UiT linguists)
  - UiT based journals
  - OJS-Dataverse plugin (TBT)
  - Individual projects
Outreach
Visibility in social media

Facebook
– where “everything now happens”

• Updates
  – New uploads
  – Presentations/workshops
  – Technical information

• Collaborative management
  – TROLLing curators
  – Faculty research assistant
User activity in TROLLing (per April 7, 2016)

Numbers
• 40 studies
• 1394 downloads
• 105 registered users
  – 19 countries
  – Europe, Asia, North- and South-America

Contributors
• 24 unique contributors
  – 5 countries
  – Europe, North-America

Associated publications
• OA journals
• Paid journals
• No publication
• PhD thesis
• Master thesis
## User activity in TROLLing (per April 7, 2016)

<table>
<thead>
<tr>
<th>Content</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Subfields&lt;br&gt;  - Semantics, syntax, morphology, phonology, phonetics&lt;br&gt;  - Synchronic, diachronic, first and second language acquisition</td>
<td>• Types of data&lt;br&gt;  - Tables, charts&lt;br&gt;  - Audio, video&lt;br&gt;  - Scripts, experimental method</td>
</tr>
<tr>
<td>• Languages&lt;br&gt;  - Czech, Old Church Slavonic, Russian, Ukrainian&lt;br&gt;  - French, Romanian, Spanish&lt;br&gt;  - German, Norwegian&lt;br&gt;  - Saami</td>
<td></td>
</tr>
</tbody>
</table>
Fully operational service and why curation is still necessary

TROLLing identity, a clear definition

It is a place for open, structured datasets belonging to the science of langue

- Yes
  - Structured, well described, openly accessible datasets

- No
  - Metadata only
  - Primary data
  - Sensitive data
  - Bibliographies, dictionaries, national anthems
    - (To be continued)

BUT

- Researchers have little time
- Researchers are not used to think about data management

VIA CURATION

- Assistance and training
- Consistent optimization

Member
TROLLing Scientific Advisory Board

«Guidelines so easy that my grandmother would not have any problems uploading data.»
To learn more about TROLLing

- Visit the archive at opendata.uit.no
- Visit the blog at site.uit.no/trolling
- Contact us at trolling@ub.uit.no
- **New idea** Join us in a TROLLing webinar, where we can have a look at the archive together, live, all while being located in different parts of the world
Thank you for your attention*

Helene
helene.n.andreassen@uit.no

TROLLing
trolling@ub.uit.no

*Thanks to Philipp Conzett, Stein Høydalsvik, Laura Janda, and Leif Longva (UiT) for useful information and constructive comments