Adapting flexible metadata support in Dataverse to the needs of domain-specific repositories – the case of The Tromsø Repository of Language and Linguistics (TROLLing)

24 November 2021

Philipp Conzett
Helene N. Andreassen
University Library
UiT The Arctic University of Norway

isko uk
Knowledge Organization Research Observatory

Parts of the work presented in this presentation have been funded by the EU Horizon 2020 Research and Innovation Program (2014-2020) under Grant Agreement No. 823782 (SSHOC) and the Research Council of Norway INFRASTRUKTUR Program under Grant Agreement No. 295700 (CLARINO+).
Outline of the presentation

1. What is TROLLing?
   - History, scope, infrastructure, support, numbers
2. Current metadata support in TROLLing
3. Future metadata support in TROLLing
Part 1: What is TROLLing?

The Tromsø Repository of Language and Linguistics
trolling.uit.no
TROLLing: history

Pre 2013: UiT University Library providing Open Access publication support.

Fall 2013: The UiT Library was contacted by Laura Janda and Tore Nesset, professors of Russian language at UiT asking for help to create a community-driven repository where linguists worldwide could archive and share their data and code to support the transparency and reproducibility of linguistic studies.

Establishment of working group and development of TROLLing; user guidelines, curation routines, outreach.

June 2014: TROLLing was launched, as (one of) the first open repository for linguistic research data.

European Open Data Champions

“By sharing our data, and doing this in an open, public, community fashion, we can determine the best practices for our field”

Name: Prof Laura A. Janda
Position: Professor of Russian Linguistics
Institution: UiT The Arctic University of Norway
Country: Norway
More info: Home Page; Other
ORCID ID: http://orcid.org/0000-0001-5047-1909
TROLLing: scope

All subdisciplines of linguistics

The international community

All types of data (but open)
  Raw data and processed data
  Text, image, audio, video, ...

All types of supplementary material
  Code/scripts
  Experimental protocol
  ...

5
TROLLing: the infrastructure

Based on the community-driven Dataverse software

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

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

For historical reasons still part of DataverseNO, an institution-based national generic repository for open research data. Will be moved to its own Dataverse installation in 2022.
TROLLing: the infrastructure

Being part of DataverseNO, TROLLing has since 2020 been CoreTrustSeal certified as a sustainable and trusted research data repository.
Some main technical features:

- automatically generated reference, including a Permanent identifier (DOI)
- Version control
- Private URL
- Embargo file access
TROLLing: the infrastructure

Since 2018, TROLLing has been a CLARIN C Centre, and basic citation metadata from TROLLing is harvested by the CLARIN Virtual Language Observatory (VLO).
Descriptive metadata harvested by more generic search engines such as Google Dataset Search and BASE Bielefeld.

Other search engines:
https://search.datacite.org/
http://b2find.eudat.eu/
TROLLing: publishing process

Preparation
- Deposit Guide
- Write readme file
- License control
- Formatting files

Creation
- Add metadata
- Submit data files
- If applicable, inform about desired anonymisation of the dataset

Submission
- Allow a few days for the data curation process

Curation
- Metadata quality control
- File format and readme check
- License control
- Address recommended changes
- Alternatively, explain why you don’t agree
- Re-submit

Revision
- 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
Deposit support

info.dataverse.no

Deposit guide:

---

**info: DataverseNO**

About | Deposit | Admin
--- | --- | ---
Prepare your data | Deposit your data | Refer to your data

**Prepare your data**

Before depositing your data in DataverseNO (including the different collections, e.g. UIT Open Research), make sure your dataset(s) comply with our guidelines below. DataverseNO accepts only research data in digital form.

Preparation guidelines for archiving finally include:

- Use consistent and comprehensible file names (see section 1 below).
- Save your data in a preferred file format(s) (see section 2 below).
- Describe your data in a flexible file (see section 3 below).

For more detailed guidelines, see below:

1. **File naming and organization**
2. **Preferred file formats**
3. How to describe your data
4. **File size**
5. **References**

For questions, comments or suggestions, see our support page.

---

**README file template:**

<Read text is included in angle brackets and should be deleted before saving.>

<DatawakeNO README File Template --- General --- Version: 2.1 (2021-01-07)>

This README file was generated on [YYYY-MM-DD] [YYYY-MM-DD] by [NAME].
Last updated: [YYYY-MM-DD].

-----------------------------
GENERAL INFORMATION
-----------------------------
// Title of Dataset:
// DOI:
// Contact Information
// The person to be contacted for questions about the dataset:
// Name:
// Institution:
// Email:
// ORCID:

<Whenever applicable, the following information should be registered in the metadata schema of DataverseNO. In the text below, remove fields/lines that are not applicable, and leave the rest unchanged.>

// Contributors: See metadata field Contributors.
// Kind of Data: See metadata field Kind of Data.
// Date of data collection/production: See metadata field Date of Collection.
// Geographic Location: See metadata section Geographic Coverage.
// Funding Sources: See metadata section Grant Information.
// Description of dataset:

<Short description of what the dataset is about, including reference to related project(s) and publication(s), if applicable. Should correspond to the information entered in the metadata fields Description and Related Publication.>
TROLLing: repository managers and curators

Helene N. Andreassen
PhD in French Phonology
Responsible for the UiT training program in research data management
Co-chair of the Linguistics Data Interest Group (Research Data Alliance)

Philipp Conzett
MA in Nordic Linguistics
Part of the repository management of DataverseNO
Member of the Steering Committee of the Global Dataverse Community Consortium
TROLLing collaboration

**CLARIN** – Common Language Resources and Technology Infrastructure, a European Research Infrastructure Consortium (ERIC)

**COST** – European Cooperation in Science and Technology: European network for Web-centred linguistic data science

**SSHOC** – Social Sciences and Humanities Open Cloud – a Horizon 2020 project

**RDA** – Research Data Alliance Linguistics Data Interest Group
TROLLing: numbers

Contributors
(as of 30 January 2021, when TROLLing reached 100 published datasets)

82 contributing authors
Representing a total of 42 research organizations
From 17 countries in 4 continents
TROLLing: numbers
(as of 24 November 2021)

Data

116 datasets containing 3,026 files
39 languages represented
Mostly supporting / replication data (articles and books)
Data from PhD and MA dissertations

Several datasets anonymised and shared with editors/peer reviewers together with a submitted journal or book manuscript
TROLLing: numbers
(as of 17 November 2021)

Usage

In total, 2302 dataset downloads
At average 4.25 downloads per dataset
Part 2: Current metadata support in TROLLing
Metadata registration in Dataverse

Metadata are registered in two rounds:
Round 1: all **mandatory (M)** and a few **recommended (R)** fields
Round 2: other recommended fields and optional fields (e.g. Social Science and Humanities Metadata)

Deposit Guidelines contain more information about the mandatory and recommended fields.

<table>
<thead>
<tr>
<th>Round 1: Citation Metadata:</th>
<th>Round 2: Citation Metadata:</th>
<th>Geospatial Metadata:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Title (M)</td>
<td>- Language (R)</td>
<td>- Geographic Coverage (R)</td>
</tr>
<tr>
<td>- Author (M), ORCID (R)</td>
<td>- Contributor (R)</td>
<td>- Geographic Bounding Box (R)</td>
</tr>
<tr>
<td>- Contact (M)</td>
<td>- Grant Information (R)</td>
<td></td>
</tr>
<tr>
<td>- Description (M)</td>
<td>- Time Period Covered (R)</td>
<td></td>
</tr>
<tr>
<td>- Keyword (M)</td>
<td>- Date of Collection (R)</td>
<td></td>
</tr>
<tr>
<td>- Related Publication (R)</td>
<td>- Kind of Data (R)</td>
<td></td>
</tr>
</tbody>
</table>

| | | |
| | | |
Need for more domain-specific metadata support

Example 1:

**Language**

*Currently:* only language of description

*Need:* also language that is investigated (currently added as keyword)
Need for more domain-specific metadata support

Example 2:

**Contributor**

**Currently:** only general/academic contributor roles

**Need:** also language research-specific roles, e.g., the OLAC Role Vocabulary, as recommended, e.g., in Tromsø Recommendations for Citation of Research Data in Linguistics ([https://doi.org/10.15497/rda00040](https://doi.org/10.15497/rda00040))

<table>
<thead>
<tr>
<th>OLAC Role Vocabulary:</th>
</tr>
</thead>
<tbody>
<tr>
<td>annotator</td>
</tr>
<tr>
<td>author</td>
</tr>
<tr>
<td>compiler</td>
</tr>
<tr>
<td>consultant</td>
</tr>
<tr>
<td>data_inputter</td>
</tr>
<tr>
<td>depositor</td>
</tr>
<tr>
<td>developer</td>
</tr>
<tr>
<td>editor</td>
</tr>
<tr>
<td>illustrator</td>
</tr>
<tr>
<td>interpreter</td>
</tr>
<tr>
<td>interviewer</td>
</tr>
<tr>
<td>participant</td>
</tr>
<tr>
<td>performer</td>
</tr>
<tr>
<td>photographer</td>
</tr>
<tr>
<td>recorder</td>
</tr>
<tr>
<td>researcher</td>
</tr>
<tr>
<td>research_participant</td>
</tr>
<tr>
<td>responder</td>
</tr>
<tr>
<td>signer</td>
</tr>
<tr>
<td>singer</td>
</tr>
<tr>
<td>speaker</td>
</tr>
<tr>
<td>sponsor</td>
</tr>
<tr>
<td>transcriber</td>
</tr>
<tr>
<td>translator</td>
</tr>
</tbody>
</table>

(Source: [http://www.language-archives.org/REC/role.html](http://www.language-archives.org/REC/role.html))

<table>
<thead>
<tr>
<th>Dataverse Contributor Roles:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collector</td>
</tr>
<tr>
<td>Data Curator</td>
</tr>
<tr>
<td>Data Manager</td>
</tr>
<tr>
<td>Editor</td>
</tr>
<tr>
<td>Funder</td>
</tr>
<tr>
<td>Hosting Institution</td>
</tr>
<tr>
<td>Project Leader</td>
</tr>
<tr>
<td>Project Manager</td>
</tr>
<tr>
<td>Project Member</td>
</tr>
<tr>
<td>Related Person</td>
</tr>
<tr>
<td>Researcher</td>
</tr>
<tr>
<td>Research Group</td>
</tr>
<tr>
<td>Rights Holder</td>
</tr>
<tr>
<td>Sponsor</td>
</tr>
<tr>
<td>Supervisor</td>
</tr>
<tr>
<td>Work Package Leader</td>
</tr>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>
Need for more domain-specific metadata support

Example 3:

**CMDI compatibility**

**Currently:** only some basic citation metadata is harvested by CLARIN Virtual Language Observatory (VLO)

**Need:** full CMDI compatibility
Part 3: Future metadata support in TROLLing
Domain-specific metadata schema(s)

Metadata Fields

- Citation Metadata (Required)
- Geospatial Metadata
- Social Science and Humanities Metadata
- Astronomy and Astrophysics Metadata
- Life Sciences Metadata
- Journal Metadata

- Language and Linguistic Metadata
Language and Linguistic metadata

Examples:

**CLARIN Core Metadata**

CMDI compatible
Recommended by CLARIN metadata WG (work in progress)

**European Language Grid (ELG) Metadata Schema**

ELG = “primary platform for Language Technology in Europe”

---

**CLARIN CMDI**

Source: [https://www.clarin.eu/content/component-metadata](https://www.clarin.eu/content/component-metadata)

---

**ELG Metadata Schema**

Language and Linguistic metadata

Examples:

External Controlled Vocabularies
- OLAC Role Vocabulary
- META-SHARE Ontology, e.g., modalityType
- ...

META-SHARE Ontology:
modalityType

Modality Type

IRI: http://w3id.org/meta-share/meta-share/ModalityType
A classification of modalities represented in the resource or processed by a tool/service

is in range of

modality_type

has members

body_gesture, combination_of_modalities, facial_expression, other, sign_language, spoken_language, unspecified, voice, written_language

Source: http://w3id.org/meta-share/meta-share
Challenges

1. How to implement complex metadata schemas (e.g. ELG)?
2. How to ensure maintenance of (complex) metadata schemas?
3. How to ensure sustainability of external controlled vocabulary services?
4. How to support interoperability on file-level?

1036 metadata fields.
Only one out of 10 (?) ELG sub-schemas!
Possible approaches

1. Use CLARIN Core Metadata for (small) supporting/replication datasets; use ELG Metadata for larger resources such as corpora.

2. Formalize and strengthen the role of the Global Dataverse Community Consortium (GDCC) to maintain Dataverse-related resources.

3. Use recognized vocabulary services, or if not available, have them run them by CLARIN, GDCC or another suitable organization.

4. For tabular data, consider adopting the Cross-Linguistic Data Formats initiative (CLDF).
Thank you for your attention!

Philipp Conzett
Helene N. Andreassen

University Library
UiT The Arctic University of Norway

TROLLing
The Tromsø Repository of Language and Linguistics

Parts of the work presented in this presentation have been funded by the EU Horizon 2020 Research and Innovation Program (2014-2020) under Grant Agreement No. 823782 (SSHOC) and the Research Council of Norway INFRASTRUKTUR Program under Grant Agreement No. 295700 (CLARINO+).