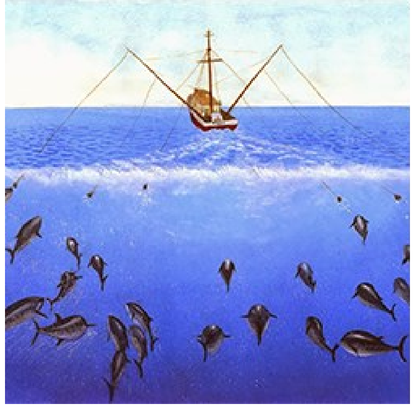
Adapting flexible metadata support in Dataverse to the needs of domain-specific repositories – the case of The Tromsø Repository of Language and Linguistics (TROLLing)

24 November 2021

Philipp Conzett
Helene N. Andreassen

University Library
UiT The Arctic University of Norway

ISKO UK
Knowledge Organization Research Observatory



TROLLing
The Tromsø Repository of Language and Linguistics

@TROLLingRepo
@PhilippConzett
@n_andreassen

CLARIN

SSHOC
social sciences & humanities open cloud

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

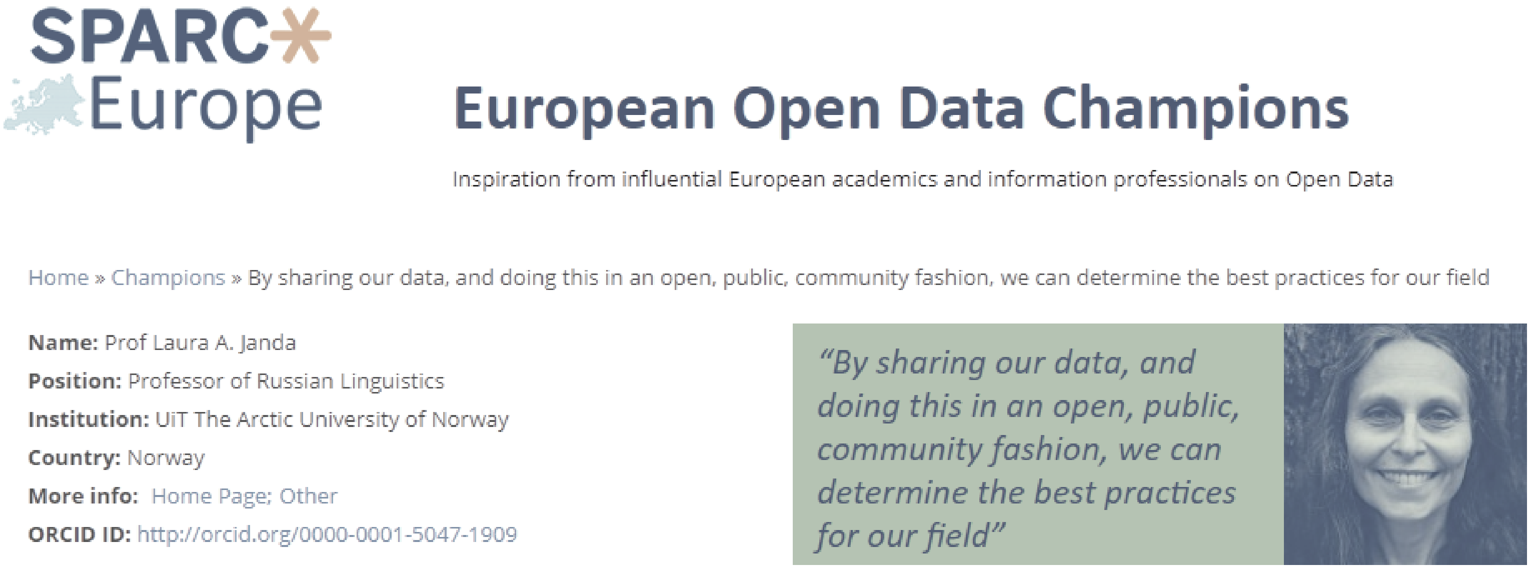
TROLLing: history

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.



SPARC Europe

European Open Data Champions

Inspiration from influential European academics and information professionals on Open Data

Home » 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

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

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
- …

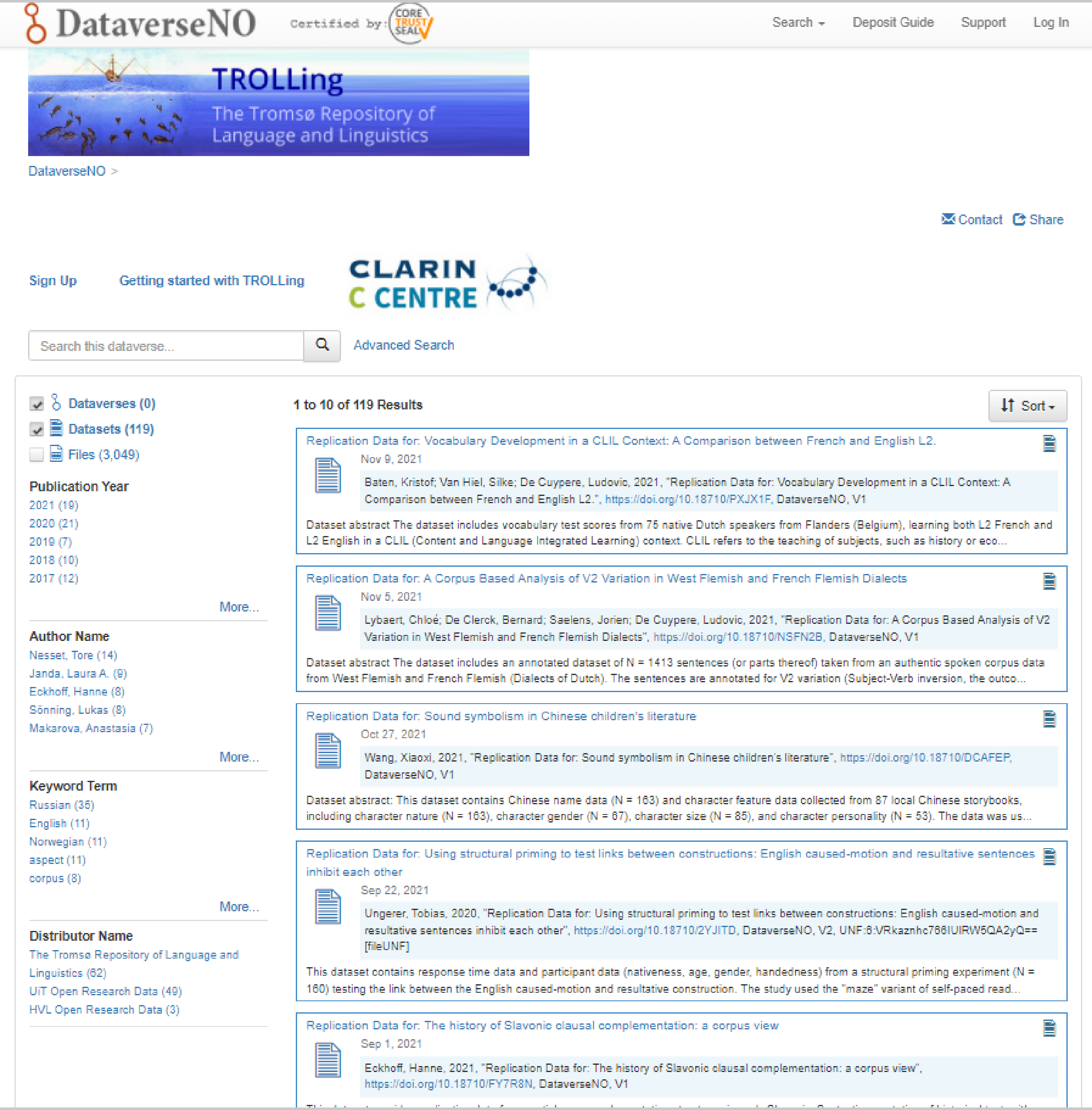
TROLLing: the infrastructure

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.



☑ Dataverses (0)
☑ Datasets (119)
□ Files (3,049)

Publication Year
2021 (19)
2020 (21)
2019 (7)
2018 (10)
2017 (12)
More...

Author Name
Nesset, Tore (14)
Janda, Laura A. (9)
Eckhoff, Hanne (8)
Sönning, Lukas (8)
Makarova, Anastasia (7)
More...

Keyword Term
Russian (35)
English (11)
Norwegian (11)
aspect (11)
corpus (8)
More...

Distributor Name
The Tromsø Repository of Language and Linguistics (62)
UiT Open Research Data (49)
HVL Open Research Data (3)

1 to 10 of 119 Results

Replication Data for: Vocabulary Development in a CLIL Context: A Comparison between French and English L2.
Nov 9, 2021
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
Dataset abstract The dataset includes vocabulary test scores from 75 native Dutch speakers from Flanders (Belgium), learning both L2 French and L2 English in a CLIL (Content and Language Integrated Learning) context. CLIL refers to the teaching of subjects, such as history or eco...

Replication Data for: A Corpus Based Analysis of V2 Variation in West Flemish and French Flemish Dialects
Nov 5, 2021
Lybaert, Chloé; De Clerck, Bernard; Saelens, Jorien; De Cuypere, Ludovic, 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
Dataset abstract The dataset includes an annotated dataset of N = 1413 sentences (or parts thereof) taken from an authentic spoken corpus data from West Flemish and French Flemish (Dialects of Dutch). The sentences are annotated for V2 variation (Subject-Verb inversion, the outco...

Replication Data for: Sound symbolism in Chinese children's literature
Oct 27, 2021
Wang, Xiaoxi, 2021, "Replication Data for: Sound symbolism in Chinese children's literature", https://doi.org/10.18710/DCAFEP, DataverseNO, V1
Dataset abstract: This dataset contains Chinese name data (N = 163) and character feature data collected from 87 local Chinese storybooks, including character nature (N = 163), character gender (N = 67), character size (N = 85), and character personality (N = 53). The data was us...

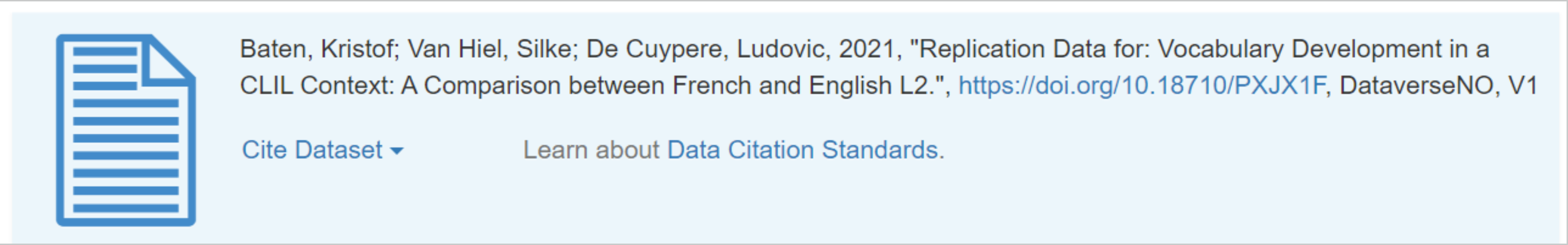
Replication Data for: Using structural priming to test links between constructions: English caused-motion and resultative sentences inhibit each other
Sep 22, 2021
Ungerer, Tobias, 2020, "Replication Data for: Using structural priming to test links between constructions: English caused-motion and resultative sentences inhibit each other", https://doi.org/10.18710/2YJITD, DataverseNO, V2, UNF:6:VRkaznhc786IUIRW5QA2yQ== [fileUNF]
This dataset contains response time data and participant data (nativeness, age, gender, handedness) from a structural priming experiment (N = 160) testing the link between the English caused-motion and resultative construction. The study used the "maze" variant of self-paced read...

Replication Data for: The history of Slavonic clausal complementation: a corpus view
Sep 1, 2021
Eckhoff, Hanne, 2021, "Replication Data for: The history of Slavonic clausal complementation: a corpus view", https://doi.org/10.18710/FY7R8N, DataverseNO, V1

TROLLing: the infrastructure

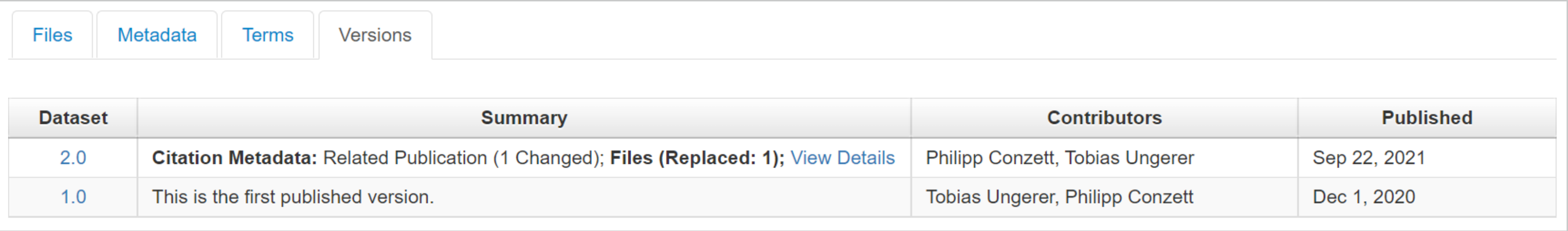
Being part of DataverseNO, TROLLing has since 2020 been CoreTrustSeal certified as a sustainable and trusted research data repository.

Some main technical features:



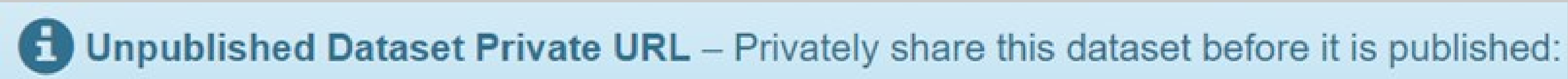
Baten, Kristof; Van Hiel, Silke; De Cuypere, Ludovic, 2021, "Replication Data for: Vocabulary Development in a CLIL Context: A Comparison between French and English L2.", https://doi.org/10.18710/PXJX1F, DataverseNO, V1

- ✓ automatically generated reference, including a
- ✓ Permanent identifier (DOI)



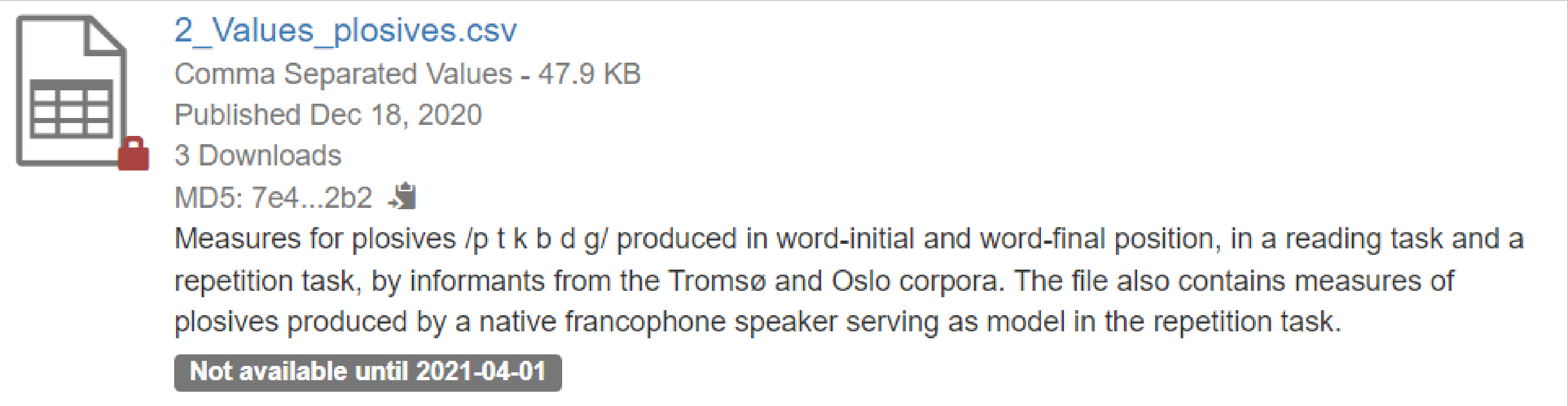
Dataset	Summary	Contributors	Published
2.0	**Citation Metadata:** Related Publication (1 Changed); **Files (Replaced: 1);** View Details	Philipp Conzett, Tobias Ungerer	Sep 22, 2021
1.0	This is the first published version.	Tobias Ungerer, Philipp Conzett	Dec 1, 2020

- ✓ Version control



Unpublished Dataset Private URL – Privately share this dataset before it is published:

- ✓ Private URL



2_Values_plosives.csv
Comma Separated Values - 47.9 KB
Published Dec 18, 2020
3 Downloads
MD5: 7e4...2b2

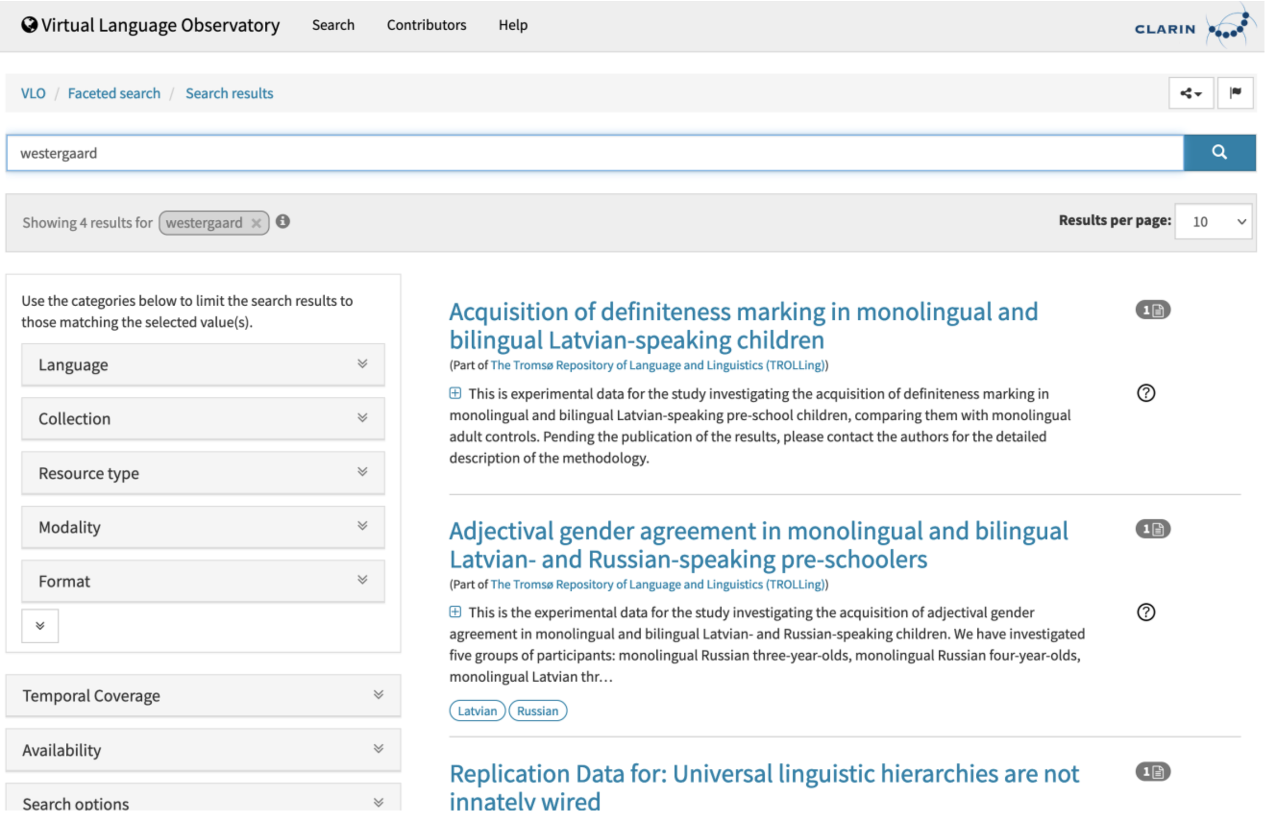
Measures for plosives /p t k b d g/ produced in word-initial and word-final position, in a reading task and a repetition task, by informants from the Tromsø and Oslo corpora. The file also contains measures of plosives produced by a native francophone speaker serving as model in the repetition task.

Not available until 2021-04-01

- ✓ 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).


Virtual Language Observatory
Search
Contributors
Help
CLARIN
VLO / Faceted search / Search results

westergaard

Showing 4 results for westergaard

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Acquisition of definiteness marking in monolingual and bilingual Latvian-speaking children

(Part of The Tromsø Repository of Language and Linguistics (TROLLing))

⊞ This is experimental data for the study investigating the acquisition of definiteness marking in monolingual and bilingual Latvian-speaking pre-school children, comparing them with monolingual adult controls. Pending the publication of the results, please contact the authors for the detailed description of the methodology.

Adjectival gender agreement in monolingual and bilingual Latvian- and Russian-speaking pre-schoolers

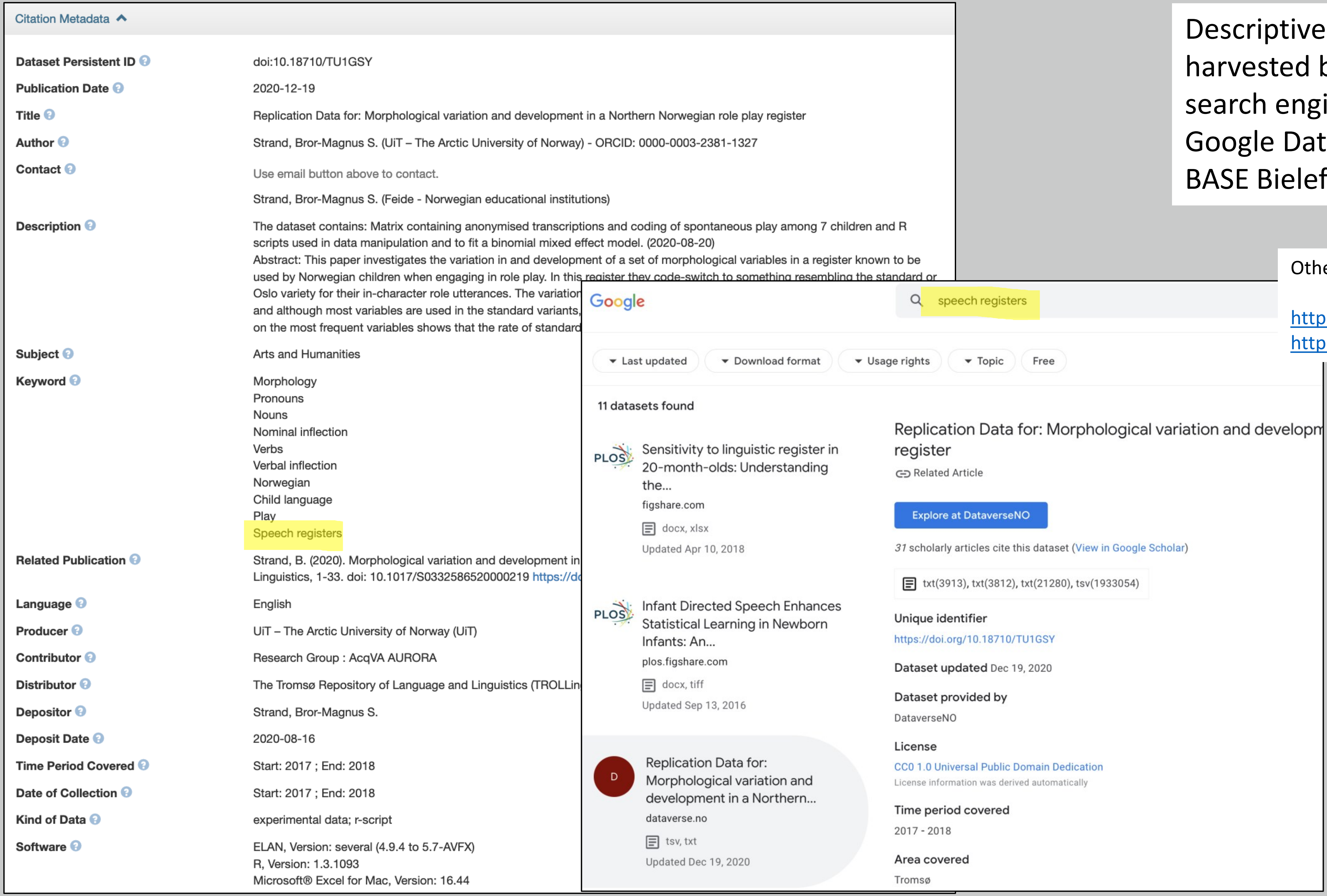
(Part of The Tromsø Repository of Language and Linguistics (TROLLing))

⊞ This is the experimental data for the study investigating the acquisition of adjectival gender agreement in monolingual and bilingual Latvian- and Russian-speaking children. We have investigated five groups of participants: monolingual Russian three-year-olds, monolingual Russian four-year-olds, monolingual Latvian thr…

Latvian Russian

Replication Data for: Universal linguistic hierarchies are not innately wired

Descriptive metadata harvested by more generic search engines such as Google Dataset Search and BASE Bielefeld.



Citation Metadata

Dataset Persistent ID	doi:10.18710/TU1GSY
Publication Date	2020-12-19
Title	Replication Data for: Morphological variation and development in a Northern Norwegian role play register
Author	Strand, Bror-Magnus S. (UiT – The Arctic University of Norway) - ORCID: 0000-0003-2381-1327
Contact	Use email button above to contact. Strand, Bror-Magnus S. (Feide - Norwegian educational institutions)
Description	The dataset contains: Matrix containing anonymised transcriptions and coding of spontaneous play among 7 children and R scripts used in data manipulation and to fit a binomial mixed effect model. (2020-08-20) Abstract: This paper investigates the variation in and development of a set of morphological variables in a register known to be used by Norwegian children when engaging in role play. In this register they code-switch to something resembling the standard or Oslo variety for their in-character role utterances. The variation and although most variables are used in the standard variants, on the most frequent variables shows that the rate of standard
Subject	Arts and Humanities
Keyword	Morphology Pronouns Nouns Nominal inflection Verbs Verbal inflection Norwegian Child language Play Speech registers
Related Publication	Strand, B. (2020). Morphological variation and development in Linguistics, 1-33. doi: 10.1017/S0332586520000219 https://do
Language	English
Producer	UiT – The Arctic University of Norway (UiT)
Contributor	Research Group : AcqVA AURORA
Distributor	The Tromsø Repository of Language and Linguistics (TROLLin
Depositor	Strand, Bror-Magnus S.
Deposit Date	2020-08-16
Time Period Covered	Start: 2017 ; End: 2018
Date of Collection	Start: 2017 ; End: 2018
Kind of Data	experimental data; r-script
Software	ELAN, Version: several (4.9.4 to 5.7-AVFX) R, Version: 1.3.1093 Microsoft® Excel for Mac, Version: 16.44

11 datasets found

Sensitivity to linguistic register in 20-month-olds: Understanding the...
figshare.com
docx, xlsx
Updated Apr 10, 2018

Infant Directed Speech Enhances Statistical Learning in Newborn Infants: An...
plos.figshare.com
docx, tiff
Updated Sep 13, 2016

Replication Data for: Morphological variation and development in a Northern...
dataverse.no
tsv, txt
Updated Dec 19, 2020

Replication Data for: Morphological variation and developm register

Related Article

Explore at DataverseNO

31 scholarly articles cite this dataset (View in Google Scholar)

txt(3913), txt(3812), txt(21280), tsv(1933054)

Unique identifier
https://doi.org/10.18710/TU1GSY

Dataset updated Dec 19, 2020

Dataset provided by
DataverseNO

License
CC0 1.0 Universal Public Domain Dedication
License information was derived automatically

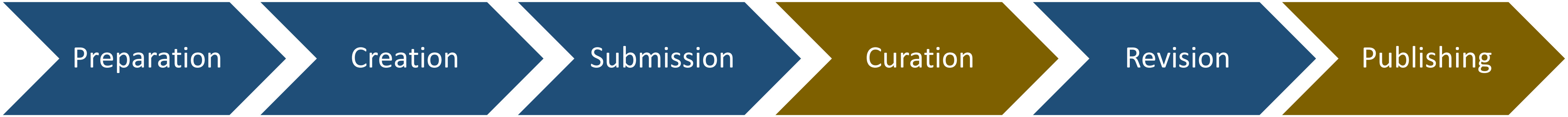
Time period covered
2017 - 2018

Area covered
Tromsø

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 anononymisation of the dataset

Submission

- Allow a few days for the data curation process

Curation

- Metadata quality control
- File format and readme check
- License control

Revision

- Address recommended changes
- Alternatively, explain why you don’t agree
- Re-submit

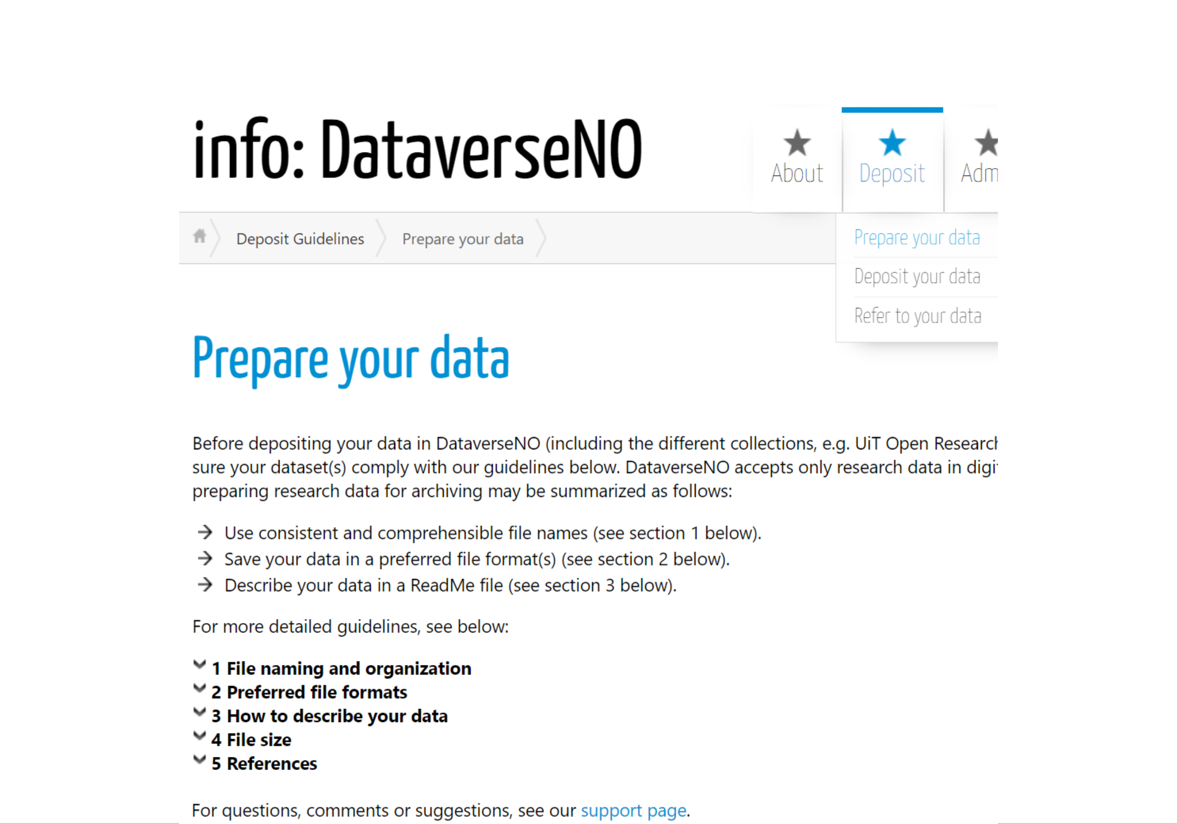
Publishing

- 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.

Deposit support

info.dataverse.no

Deposit guide:



info: DataverseNO

Prepare your data

Before depositing your data in DataverseNO (including the different collections, e.g. UiT Open Researcl sure your dataset(s) comply with our guidelines below. DataverseNO accepts only research data in digi preparing research data for archiving may be summarized as follows:

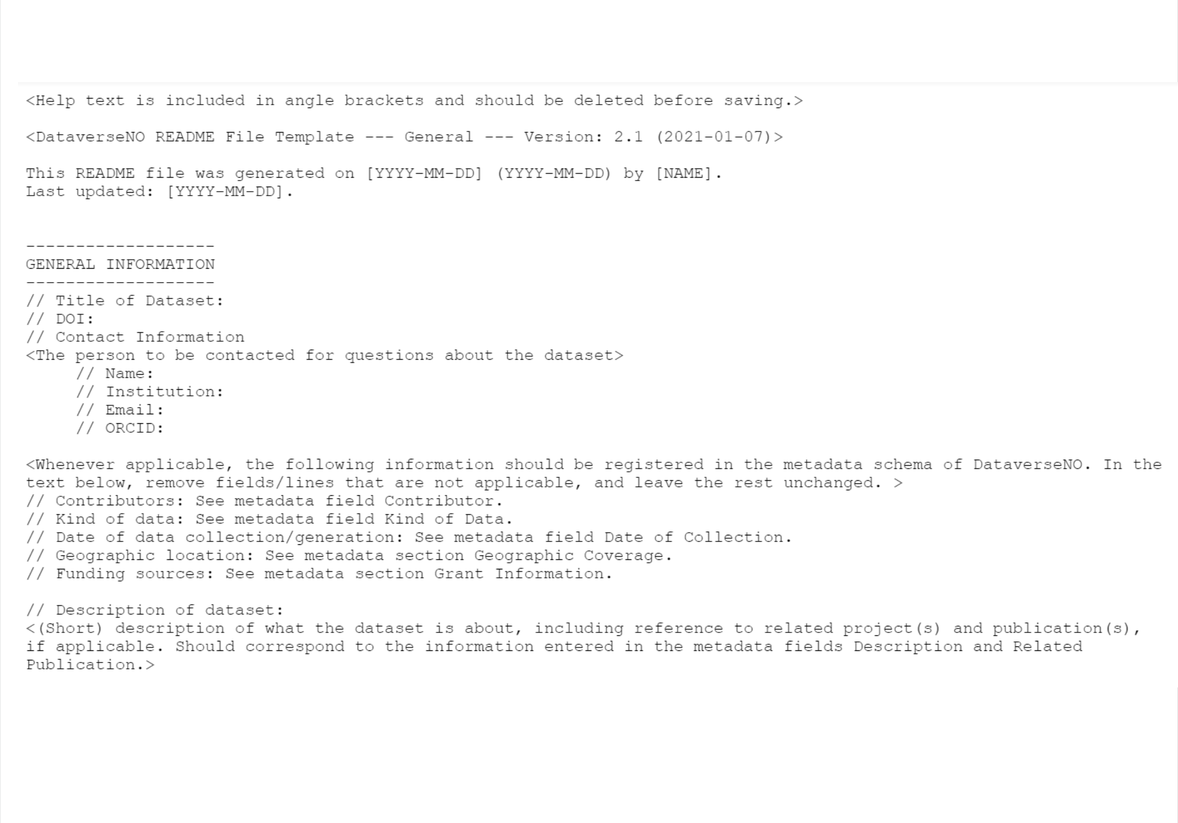
- → Use consistent and comprehensible file names (see section 1 below).
- → Save your data in a preferred file format(s) (see section 2 below).
- → Describe your data in a ReadMe file (see section 3 below).

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:



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

<DataverseNO 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 Contributor.
// Kind of data: See metadata field Kind of Data.
// Date of data collection/generation: 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)

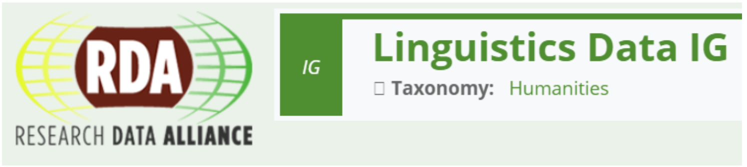
Philipp Conzett

MA in Nordic Linguistics

Part of the repository management of DataverseNO

Member of the Steering Committee of the Global Dataverse Community Consortium

Photo: private



Linguistics Data IG

IG

□ Taxonomy: Humanities

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

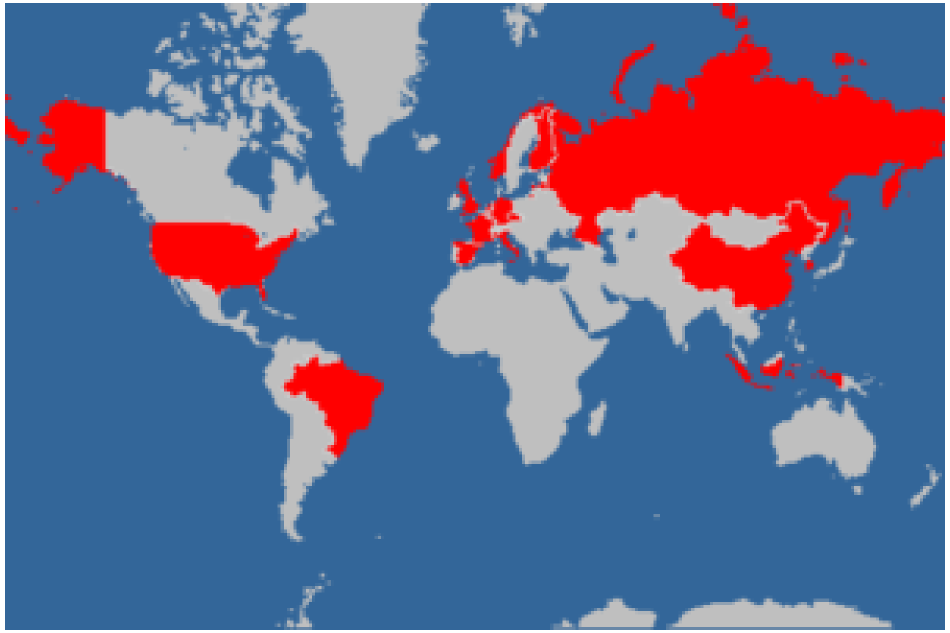
Contributors

(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

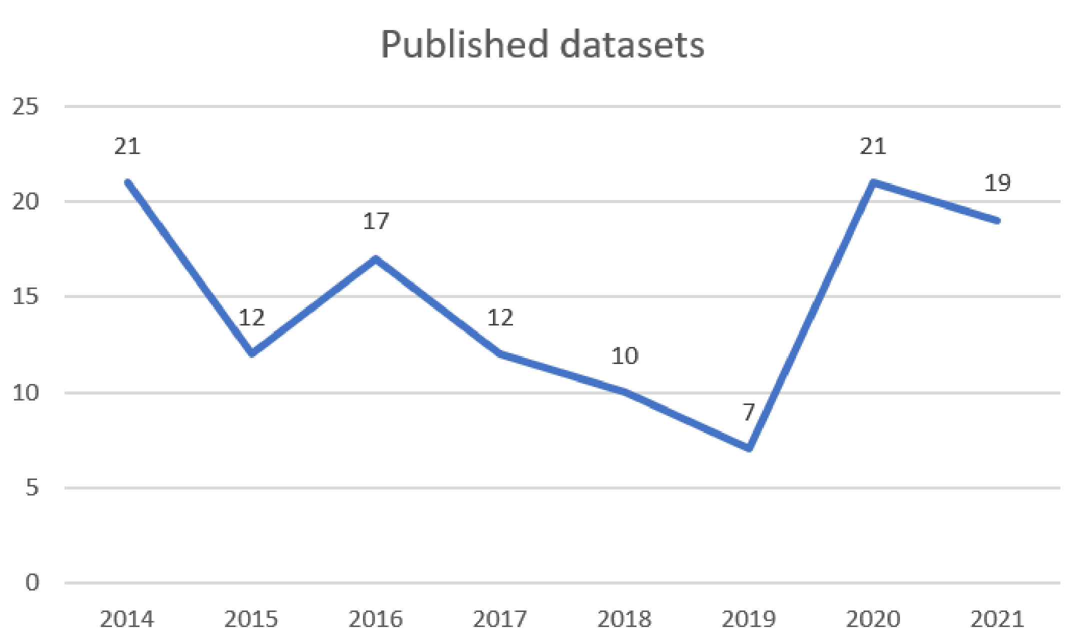
116 datasets containing 3 026 files

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

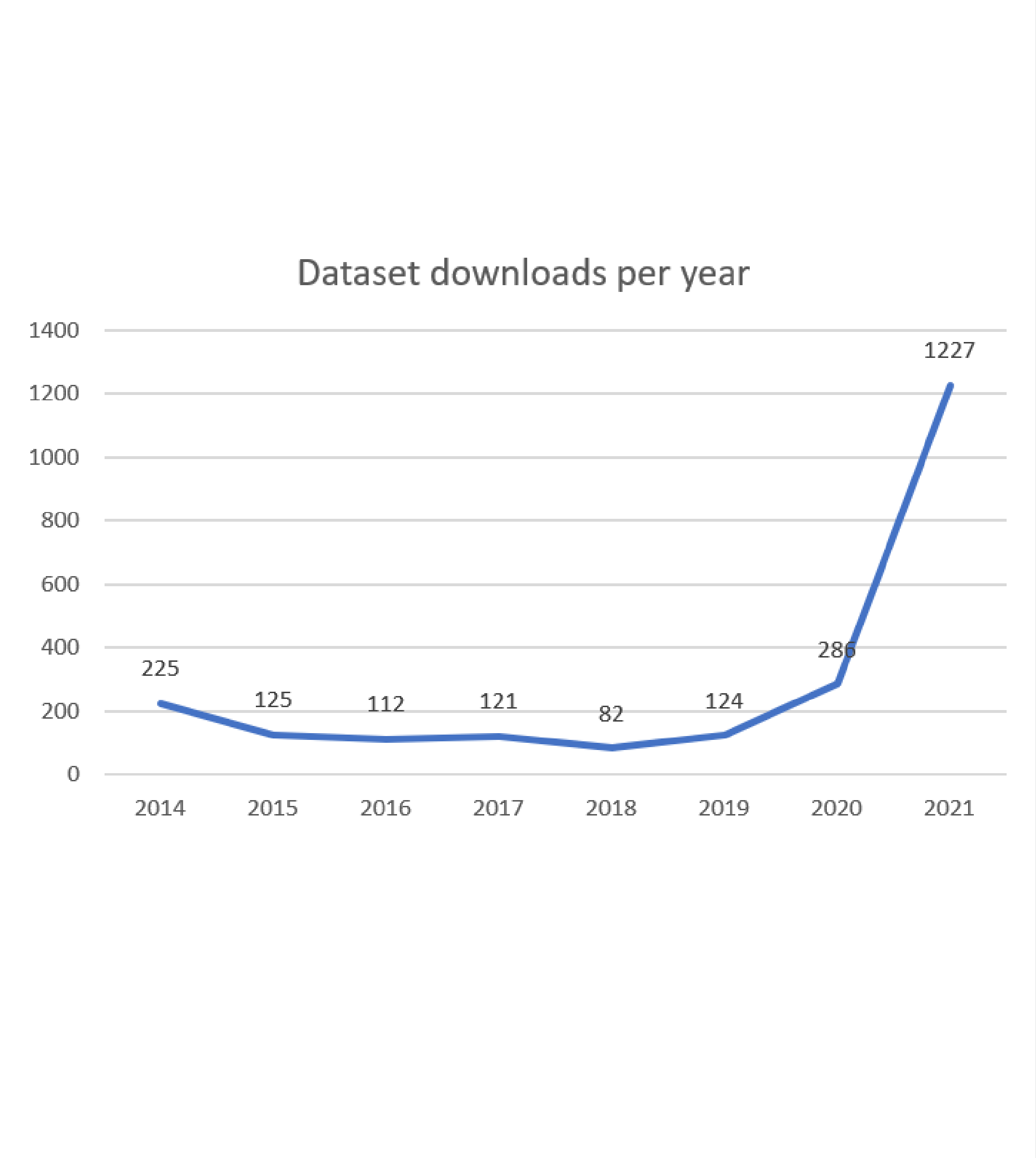


TROLLing: numbers
(as of 17 November 2021)

Usage

In total, 2302 dataset downloads

At average 4.25 downloads per dataset

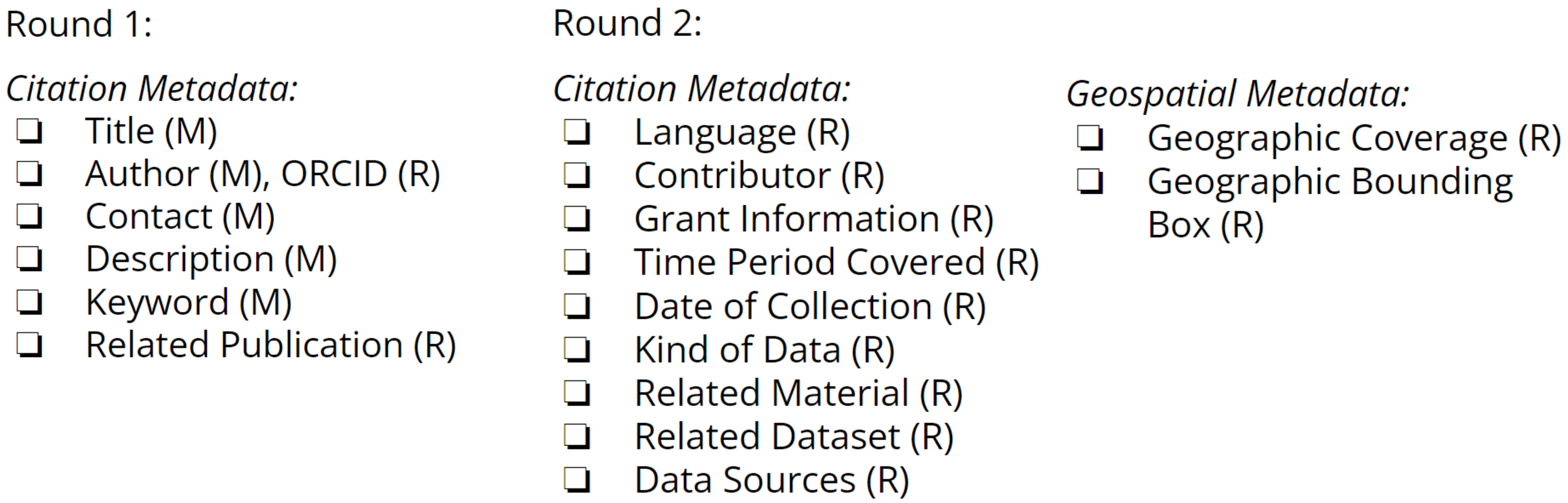

Dataset downloads per year

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.



Round 1:

Citation Metadata:

- ❑ Title (M)
- ❑ Author (M), ORCID (R)
- ❑ Contact (M)
- ❑ Description (M)
- ❑ Keyword (M)
- ❑ Related Publication (R)

Round 2:

Citation Metadata:

- ❑ Language (R)
- ❑ Contributor (R)
- ❑ Grant Information (R)
- ❑ Time Period Covered (R)
- ❑ Date of Collection (R)
- ❑ Kind of Data (R)
- ❑ Related Material (R)
- ❑ Related Dataset (R)
- ❑ Data Sources (R)

Geospatial Metadata:

- ❑ Geographic Coverage (R)
- ❑ Geographic Bounding Box (R)

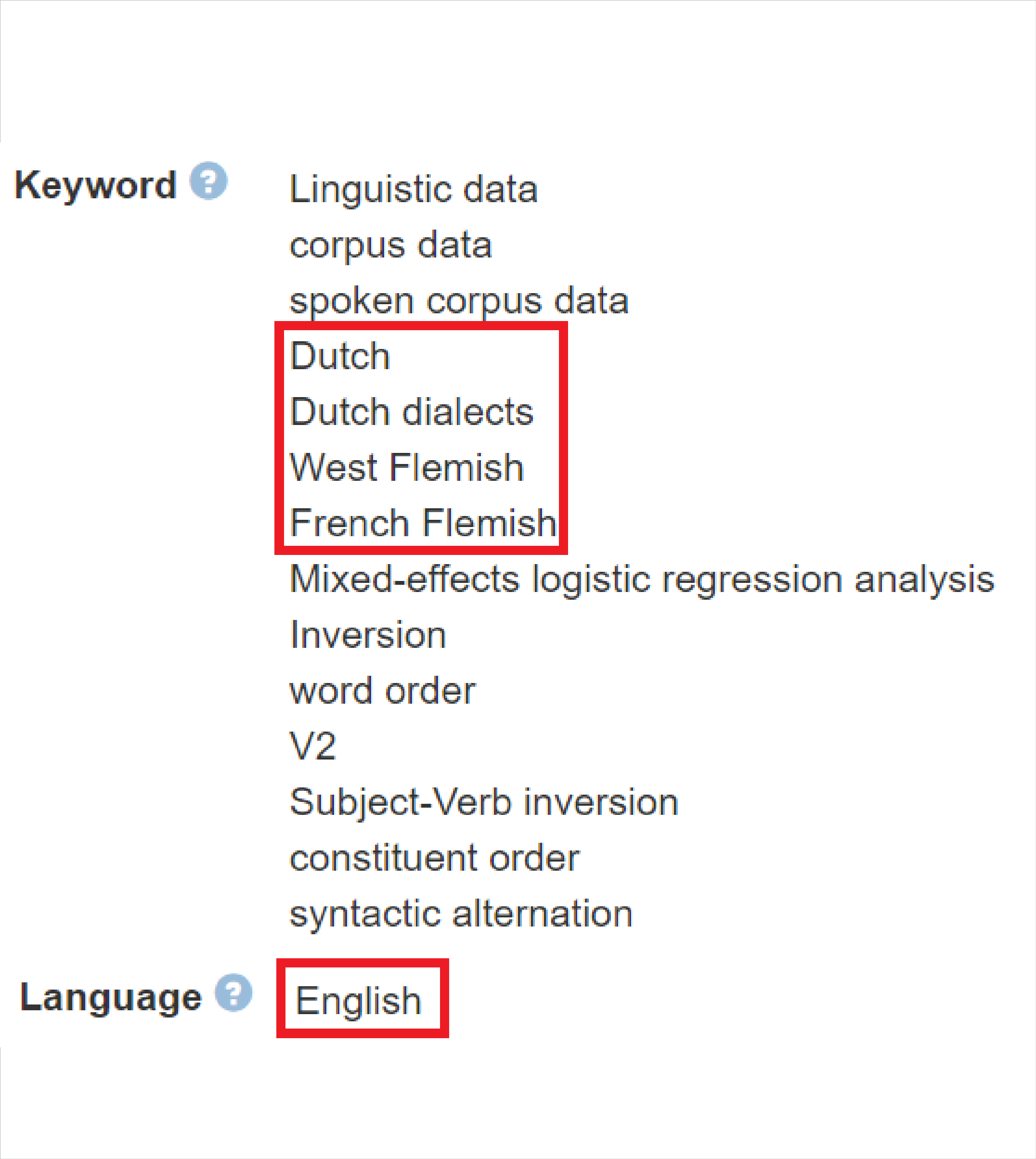
Need for more domain-specific metadata support

Example 1:

Language

<u>Currently:</u> only language of description

<u>Need:</u> also language that is investigated (currently added as keyword)



Keyword

Linguistic data
corpus data
spoken corpus data
Dutch
Dutch dialects
West Flemish
French Flemish
Mixed-effects logistic regression analysis
Inversion
word order
V2
Subject-Verb inversion
constituent order
syntactic alternation

Language English

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)

Dataverse Contributor Roles:

- Data Collector
- Data Curator
- Data Manager
- Editor
- Funder
- Hosting Institution
- Project Leader
- Project Manager
- Project Member
- Related Person
- Researcher
- Research Group
- Rights Holder
- Sponsor
- Supervisor
- Work Package Leader
- Other

OLAC Role Vocabulary:

- annotator
- author
- compiler
- consultant
- data_inputter
- depositor
- developer
- editor
- illustrator
- interpreter
- interviewer
- participant
- performer
- photographer
- recorder
- researcher
- research_participant
- responder
- signer
- singer
- speaker
- sponsor
- transcriber
- translator

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

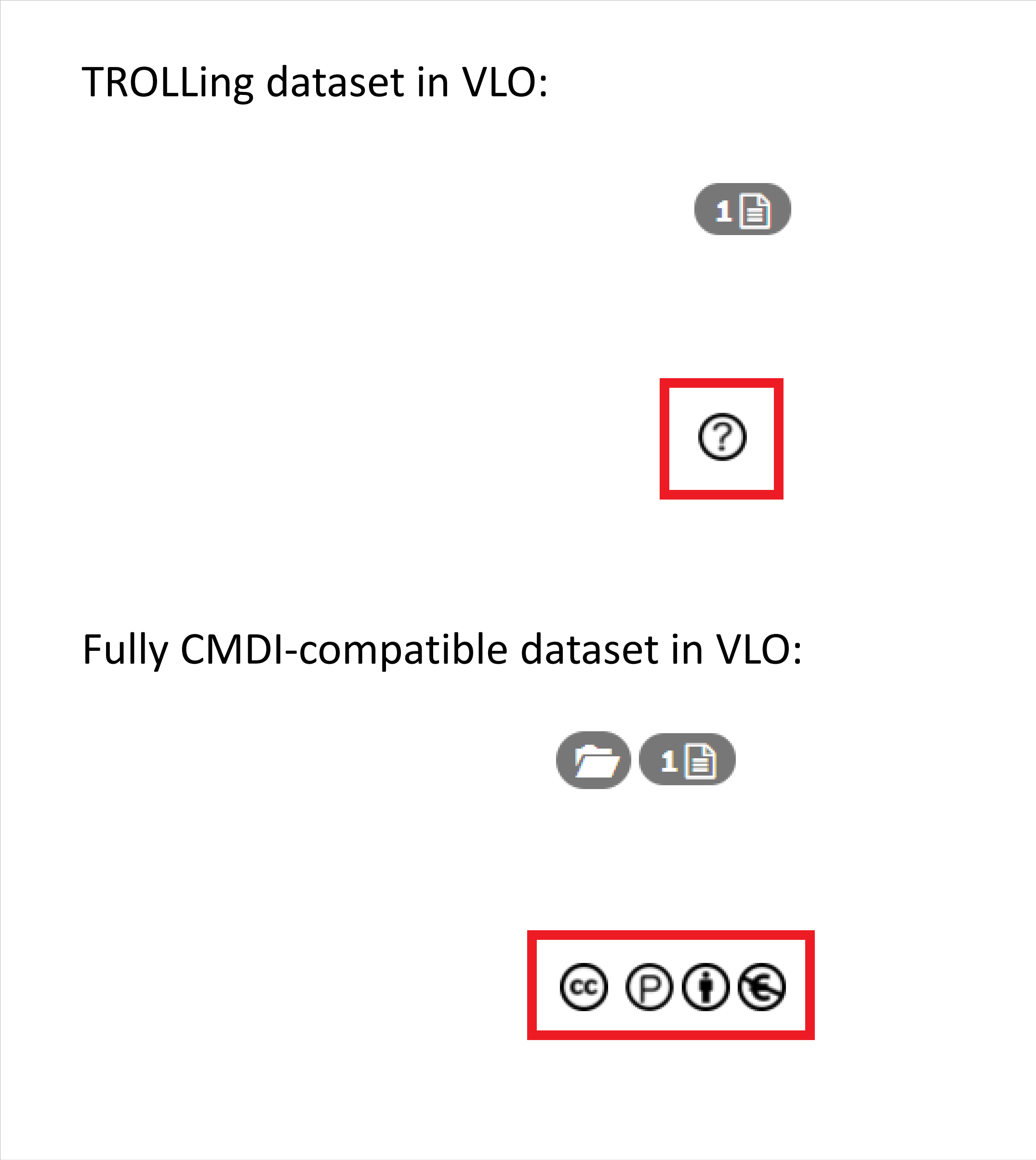
Need for more domain-specific metadata support

Example 3:

CMDI compatibility

<u>Currently:</u> only some basic citation metadata is harvested by CLARIN Virtual Language Observatory (VLO)

<u>Need:</u> full CMDI compatibility

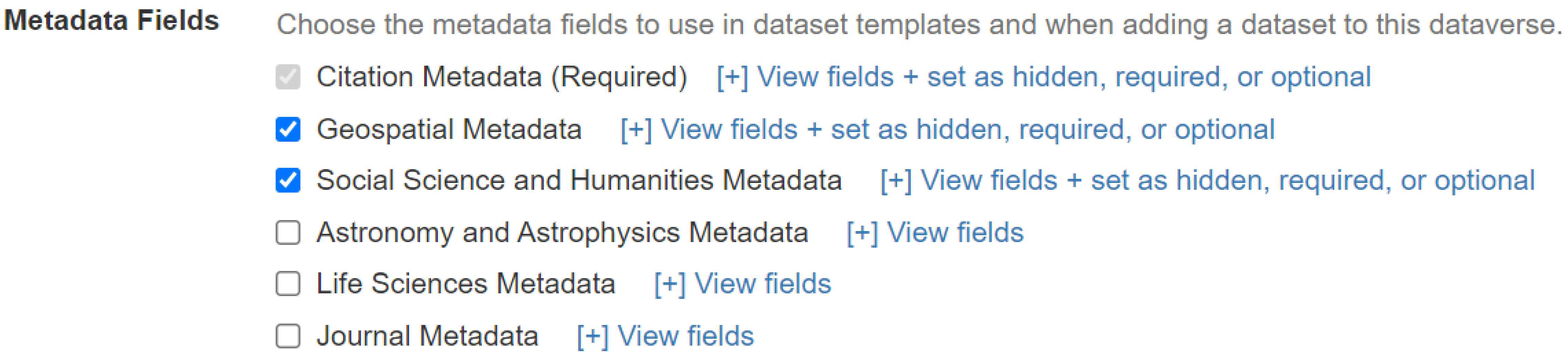


TROLLing dataset in VLO:

Fully CMDI-compatible dataset in VLO:

Part 3: Future metadata support in TROLLing

Domain-specific metadata schema(s)



Metadata Fields Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- ☑ Citation Metadata (Required) [+] View fields + set as hidden, required, or optional
- ☑ Geospatial Metadata [+] View fields + set as hidden, required, or optional
- ☑ Social Science and Humanities Metadata [+] View fields + set as hidden, required, or optional
- □ Astronomy and Astrophysics Metadata [+] View fields
- □ Life Sciences Metadata [+] View fields
- □ Journal Metadata [+] View fields

☑ Language and Linguistic Metadata

Language and Linguistic metadata

Examples:

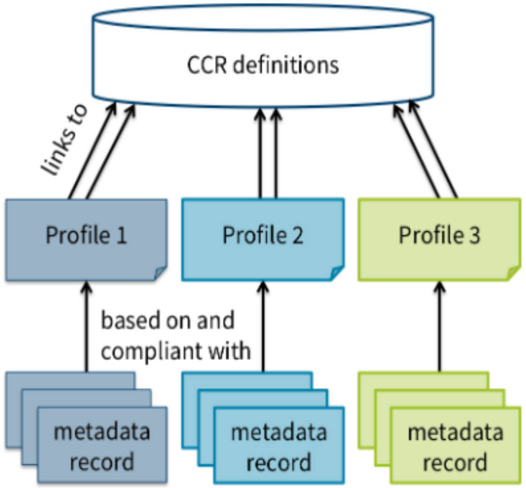
CLARIN Core Metadata

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

European Language Grid (ELG) Metadata Schema

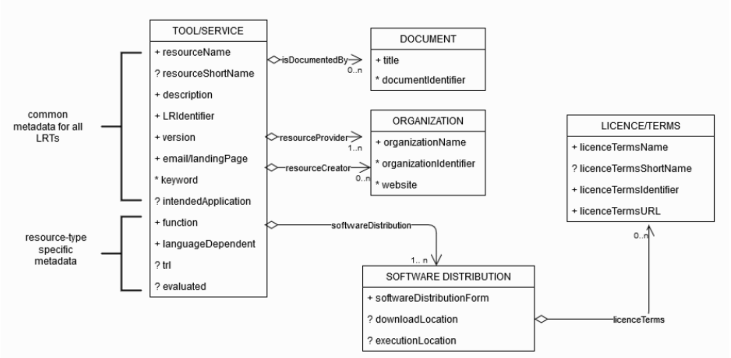
ELG = “primary platform for Language Technology in Europe”

CLARIN CMDI



Source: https://www.clarin.eu/content/component-metadata

ELG Metadata Schema



Source: https://european-language-grid.readthedocs.io/en/release1.1.1/all/A1_Metadata/Metadata.html

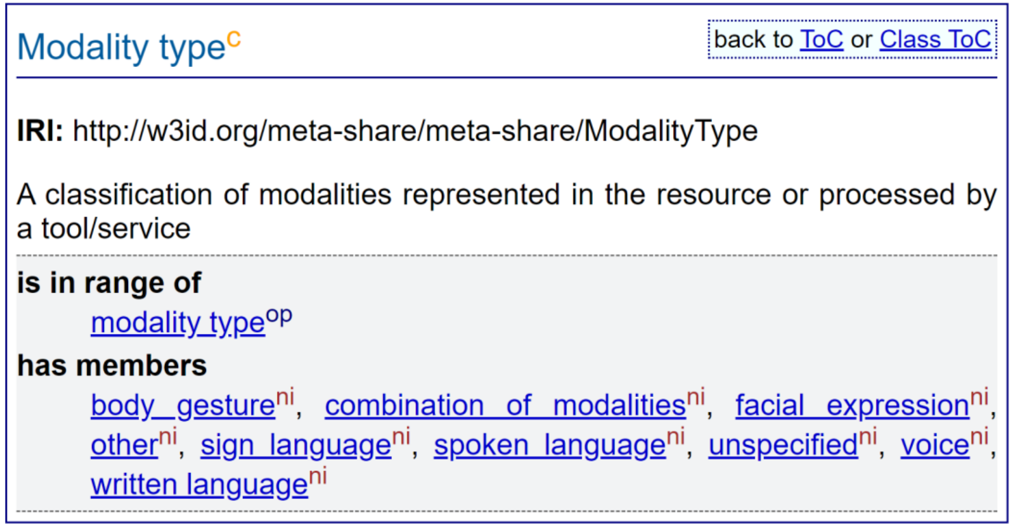
Language and Linguistic metadata

Examples:

External Controlled Vocabularies

- OLAC Role Vocabulary
- META-SHARE Ontology, e.g., modalityType
- …

META-SHARE Ontology: modalityType



Modality type^c

back to ToC or Class ToC

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^op

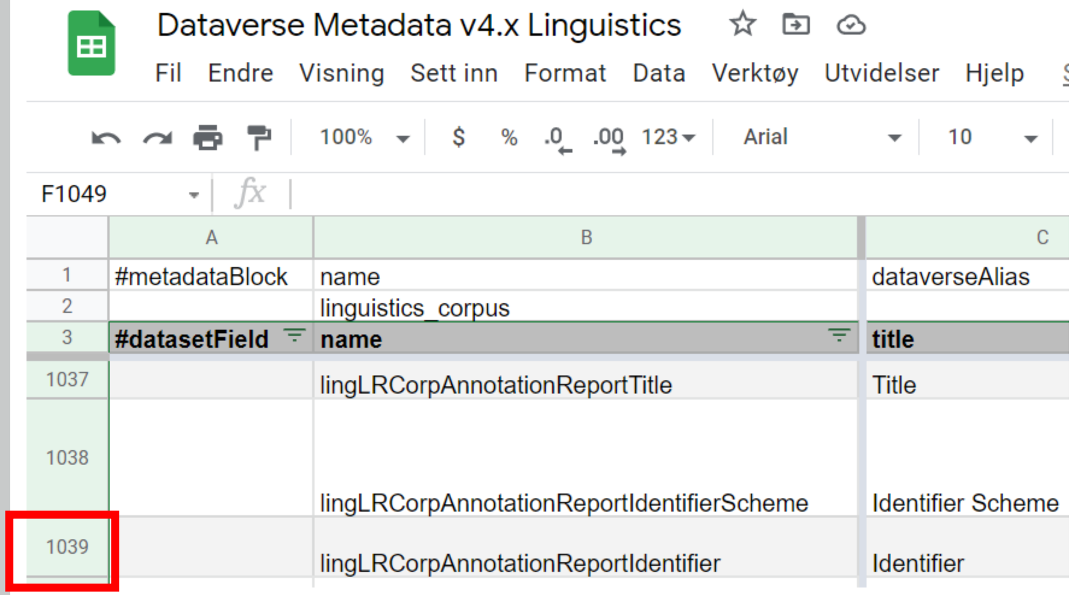
has members

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

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?


Dataverse Metadata v4.x Linguistics
Fil Endre Visning Sett inn Format Data Verktøy Utvidelser Hjelp
100% $ % .0 .00 123 Arial 10
F1049
fx

	A	B	C
1	#metadataBlock	name	dataverseAlias
2		linguistics_corpus	
3	#datasetField	name	title
1037		lingLRCorpAnnotationReportTitle	Title
1038		lingLRCorpAnnotationReportIdentifierScheme	Identifier Scheme
1039		lingLRCorpAnnotationReportIdentifier	Identifier

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

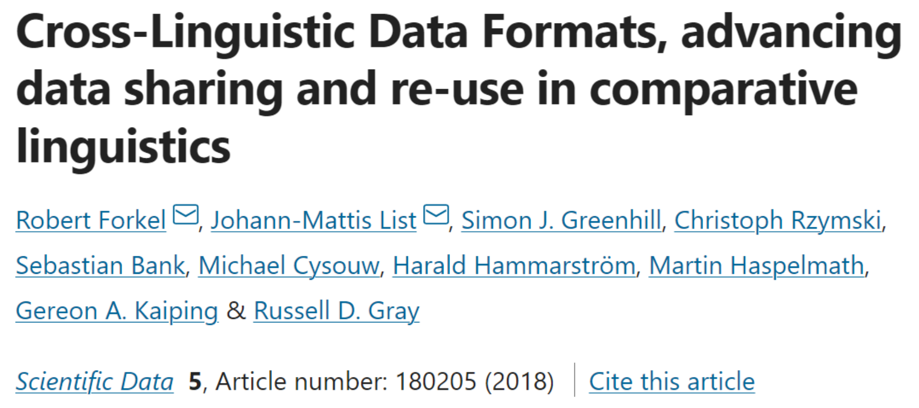
Possible approaches

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).

CLARIN

EUROPEAN LANGUAGE GRID

GDCC The Global Dataverse Community Consortium
Supporting Dataverse repositories Around the World



Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray

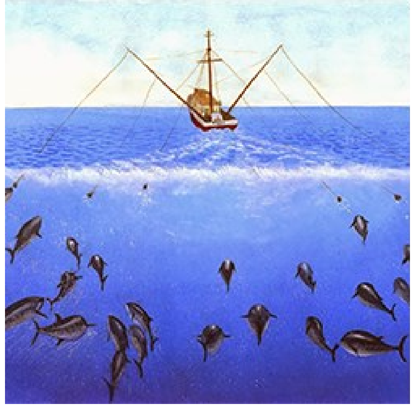
Scientific Data **5**, Article number: 180205 (2018) | Cite this article

Thank you for your attention!

Philipp Conzett
Helene N. Andreassen

University Library
UiT The Arctic University of Norway

THE ARCTIC UNIVERSITY OF NORWAY • UiT •



TROLLing

The Tromsø Repository of Language and Linguistics

CLARIN

SSHOC
social sciences & humanities open cloud

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+).