Dataverse

Dataverse as a CLARIN repository application

CLARIN Centre Meeting 2020

Philipp Conzett
UiT The Arctic University of Norway
and CLARINO

CLARIN ERIC

Outline

- ❏ **Background:** Why am I presenting this?
- ❏ **Dataverse:** Main features; How FAIR is Dataverse? Community; Development
- ❏ **Dataverse as a CLARIN repository application:** Summary
- ❏ **Questions and discussion**

Background

Why am I presenting this?

- Because my university (UiT The Arctic University of Norway) runs a repository for open data from linguistics, the Tromsø Repository of Language and Linguistics (TROLLing), which uses Dataverse as its repository application.

Repository: https://trolling.uit.no/
Info site: https://info.trolling.uit.no/

TROLLing in a nutshell - Brief history

- Initiated in 2013 by linguists at UiT
- Developed by the UiT Library together with linguists from UiT
- Launched in 2014
- CLARIN C Centre since 2018

CLARIN C CENTRE

- Special collection within DataverseNO since 2018

DataverseNO

- CoreTrustSeal certified since last week (24 March 2020)

CORE TRUST SEAL

TROLLing in a nutshell - Key features

- ❏ Open repository for linguistic data and (statistical) code
- ❏ Open and free for linguists worldwide
- ❏ All datasets are curated by subject librarians before publication
- ❏ Default license: CC0
- ❏ Based on Dataverse repository application; used virtually out-of-the-box; low technical development and maintenance costs for UiT

For more information, see Conzett (2019) and GÉANT and UNINETT (2019).

TROLLing in a nutshell - Key figures

- ❏ 160 registered users (= potential depositors)
- ❏ 84 published datasets
- ❏ 2 716 files
- ❏ Working on download statistics

(As of 31 March 2020)

Outline

- ❏ **Background:** Why am I presenting this?
- ❏ **Dataverse:** Main features; How FAIR is Dataverse? Community; Development
- ❏ **Dataverse as a CLARIN repository application:** Summary
- ❏ **Questions and discussion**

Dataverse: Main features for data management

- ❏ **Persistent Identifiers:** Support for DOI and Handle at dataset and file level
- ❏ **Citation:** Generated references at dataset and file level
- ❏ **Versioning:** Versioning of datasets
- ❏ **Metadata:** Schemas for general metadata (e.g. Dublin Core) and domain-specific metadata (e.g. DDI), customizable metadata schemas
- ❏ **Embargo:** Restrict file access for a period of time
- ❏ **File storage:** Different systems: Local, Swift (OpenStack), S3 (Amazon)
- ❏ Coming: Support for sensitive data (DataTags)

Adapted from Durand (2020)

Dataverse

Dataverse: Main features for user management

- ❏ **Sign-in:** Multiple sign-in options: Native, Shibboleth, OAuth (ORCID, Github, Google, Microsoft), Open ID Connect
- ❏ **Collections:** Possible to create collections and sub-collections
- ❏ Branding and widgets

Adapted from Durand (2020)

Dataverse

Dataverse: Main features for workflows

- ❏ Different curation and publishing workflows may be configured
- ❏ Private URLs for access to unplublished datasets, e.g. for peer review
- ❏ Data upload / download options:
  - ❏ Browser / FileUploader
  - ❏ Dropbox
  - ❏ Rsync (for big data “packages”)
  - ❏ Remote Storage (TRSAs)

Adapted from Durand (2020)

Dataverse

Dataverse: Main features for interoperability

- ❏ APIs
  - ❏ SWORD
  - ❏ Native
  - ❏ Metrics
- ❏ Harvesting (OAI-PMH)
  - ❏ Server
  - ❏ Client
- ❏ Modular external tools
  - ❏ Explore and configure
  - ❏ Scope: Dataset / datafile

From Durand (2020)

Dataverse

How FAIR is Dataverse?

Adapted from Crosas (2020):

- ❑ Currently strong support for **F**indable, **A**ccessible, and **R**eusable principles
- ❑ Currently weak support for **I**nteroperable principles
- ❑ Continuously improving its FAIR alignment and thereby contributing to **increased FAIRness** of the data published in Dataverse

From Crosas (2020):

FM [AID*]	Question	Dataverse Q'aire	Dataverse Optimized
Identifier type	1	DOI	DOI
F1A	2		
F1B	Not tested in Q'aire		
F2A	4A		
F2A	4B		
F3	5B		
F4	6A		
F4	6B		
A1.1	7A		
A1.2	8A		
A1.2	8B	N/A	N/A
A2	9		
I1	10		
I2	11		
I3	12		
R1.1	13		
R1.2	14A		

The Dataverse Community: Where?

- 55 installations around the world (as of 31 March 2020):

Source: https://dataverse.org/

The Dataverse Community

Who is contributing?

- ❏ Developers, researchers, librarians, data scientists (several hundreds in total)

How are they contributing?

Adapted from Durand (2020)

- ❏ Code (100+ contributors)
- ❏ UI/UX testing & interviews
- ❏ Almost daily discussing issues in the Dataverse Google Group
- ❏ Participating in Dataverse Community Calls every second week
- ❏ Dataverse Community Meeting once a year at Harvard
- ❏ Workshops & trainings, e.g. European Dataverse Workshop 2020 at UiT
- ❏ **G**lobal **D**ataverse **C**ommunity **C**onsortium to coordinate community efforts

Continuous development of Dataverse

- Dataverse Roadmap: https://www.iq.harvard.edu/roadmap-dataverse-project
- SSHOC task 5.2: Hosting and sharing data repositories (Wittenberg and Tykhonov, 2020)
  - Goal: Building mature research data repository infrastructure for the European Open Science Cloud
  - Based on Dataverse
  - Based on requirements from involved communities
  - Including support for CMDI metadata and controlled vocabularies for linguistic data
- Similar efforts in other projects and networks: CLARIAH+, CLARINO+, COST Action: European network for Web-centred linguistic data science, ...

Outline

- ❏ **Background:** Why am I presenting this?
- ❏ **Dataverse:** Main features; How FAIR is Dataverse? Community; Development
- ❏ **Dataverse as a CLARIN repository application:** Summary
- ❏ **Questions and discussion**

Dataverse as a CLARIN repository application

What can Dataverse offer to CLARIN?

- ❏ Functional repository application for research data
- ❏ Increasing interoperability support for linguistic data
- ❏ Strong support from an international developer and user community

Thank you!

Outline

- ❑ **Background:** Why am I presenting this?
- ❑ **Dataverse:** Main features; How FAIR is Dataverse? Community; Development
- ❑ **Dataverse as a CLARIN repository application:** Summary
- ❑ **Questions and discussion**

References

Crosas, Mercè. 2020. “Fair Principles and Beyond: Implementation in Dataverse”. *Septentrio Conference Series*, no. 2 (March). https://doi.org/10.7557/5.5334.

Conzett, Philipp. 2019. “Disciplinary Case Study: The Tromsø Repository of Language and Linguistics (TROLLing)”. https://doi.org/10.5281/zenodo.2668775.

Durand, Gustavo. 2020. “Dataverse’s Approach to Technical Community Engagement”. *Septentrio Conference Series*, no. 2. https://doi.org/10.7557/5.5424.

GÉANT, and UNINETT. 2019. ‘Why TROLLing Is the Thing to Do for Linguists’. In The Field. May 2019. https://www.inthefieldstories.net/why-trolling-is-the-thing-to-do-for-linguists/.

Wittenberg, Marion, and Vyacheslav Tykhonov. 2020. “Dataverse in the European Open Science Cloud”. *Septentrio Conference Series*, no. 2 (March). https://doi.org/10.7557/5.5421.