Core NRP Repository Systems

Illyria Brejchová, Eliška Blažková, Milan Janíček
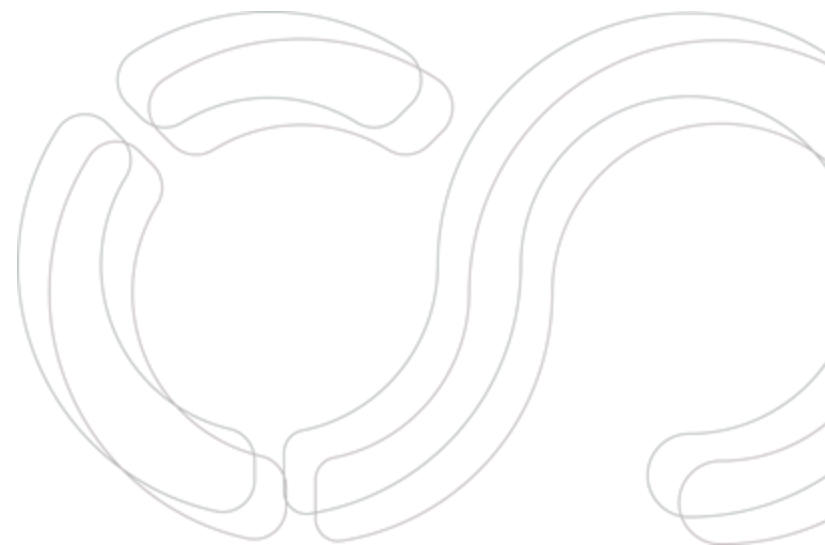
Illyria Brejchová

INVENIO
Powering Open Science

Framework
Open Source framework for
large-scale digital repositories

RDM
Turn-key Research Data Management repository

ILS
Integrated Library System

inveniosoftware.org

cesnet INVENIO
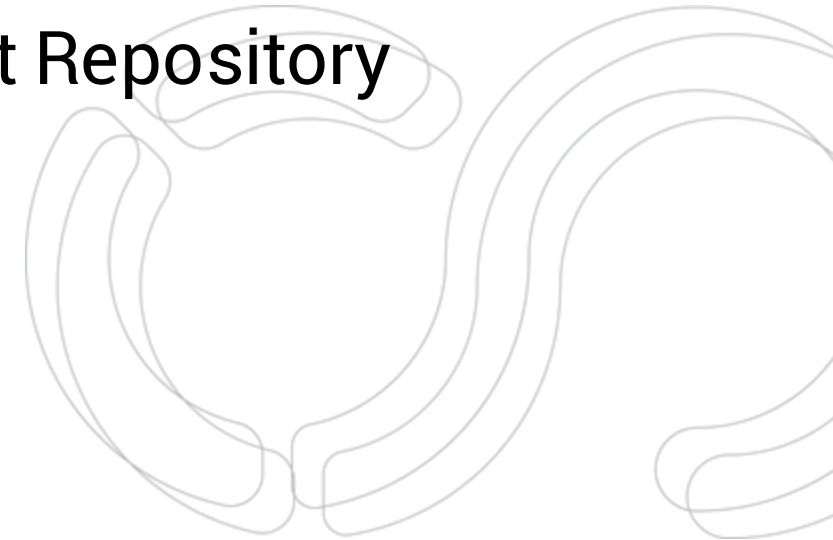
Example Invenio Instances

RDM:

- Zenodo
- Invenio RDM Sandbox

CESNET:

- Catch-all Data Repository
- MBDB
- National Document Repository (comming soon)

Differences Between InvenioRDM & CESNET Invenio

- E-INFRA **AAI** integration
- Integration with NRP **Kubernetes** and **S3** storage (ensuring DevOps and geographic replications)
- **Workflows** per community

Coming Soon:

- **CCMM**
- Some enhanced components to be contributed upstream to RDM (improved deposition form, dataset upload)

CESNET Invenio Levels of Service

A.

CCMM and basic UI adjustments

B.

Extended CCMM

C.

Own metadata model

Metadata in Invenio

- Defined in YAML -> generated python code, can include mixins
- Huge impact on the entire system configuration
- Domain specific metadata – best to define new record model linked to CCMM
- CESNET Invenio documentation
  - https://nrp-cz.github.io/docs/customize/model_backend/model
  - https://github.com/NRP-CZ/ccmm-invenio

3 December 2025

```yaml
CCMMProductionDataset:
  type: object
  properties:
    # iri:
    # IRI is not used inside schema as it is

    # we use RDM publication date as it is, publication_year is derived from it
    publication_date:
      type: RDMPublicationDate
      label:
        en: Publication date
        cs: Datum publikace

    # we use RDM version
    version:
      type: RDMRecordVersion
      label:
        en: Record Version
        cs: Verze záznamu

    # taken from RDM
    title:
      type: RDMTitle
      label:
        en: Title
        cs: Název

    publisher:
      type: RDMPublisher
      label:
        en: Publisher
        cs: Vydavatel
      hint:
        en: The entity responsible for making the resource available.
        cs: Entita odpovědná za zpřístupnění zdroje.
      help:
        en: Provide the name of the organization or individual that published the dat
        cs: Uveďte název organizace nebo jednotlivce, který publikoval datovou sadu.
```
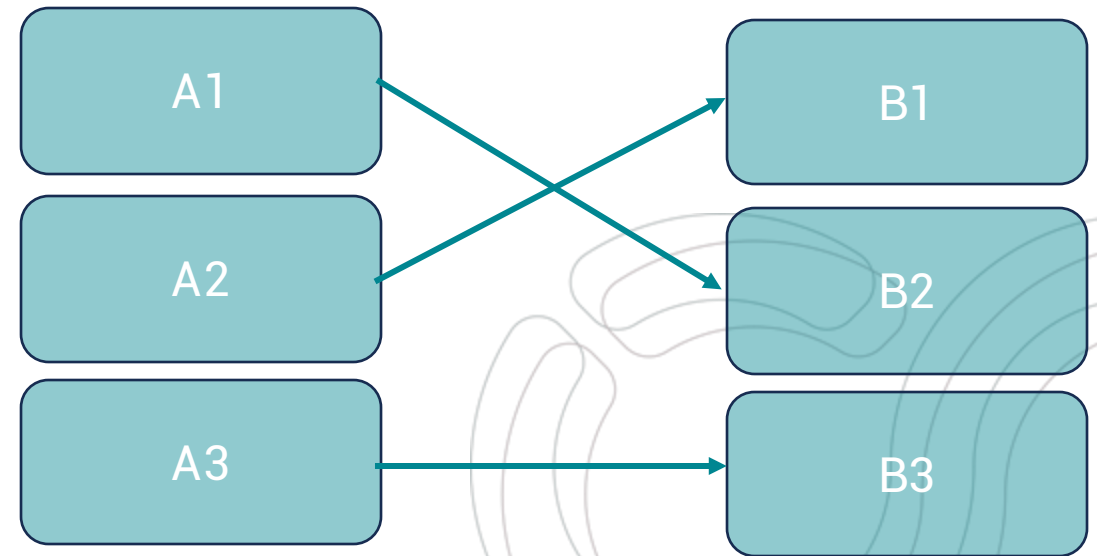
Metadata Mappings

- Parsers defined internally (python Marshmallow)
  - DataCite
  - CCMM for NMA
- Or defined externally
  - e.g. XSLT transformation

Coming soon:
- OpenAIRE Graph

Persistent Identifiers

- DOI by default
- Concept/Parent DOI
  - New DOI for each version (when files change)
- Possible to have a different prefix per community

https://inveniordm.docs.cern.ch/operate/customize/dois/#parent-or-concept-dois

Versions

Version v1   Jun 15, 2024
10.81088/a07dc-8bp18

Cite all versions? You can cite all versions by using the DOI 10.81088/t5n1r-6ex44. This DOI represents all versions, and will always resolve to the latest one. Read more.

Communities and User Roles

- **Reader**: A reader is a member of the community and can view restricted records owned by the community.

- **Curator**: A curator can in addition to a reader also edit/accept/decline records in a community.

- **Manager**: A manager can in addition to curators also manage members of a community.

- **Owner**: An owner has full administrative access to a community, and can change all settings as well as delete the community.

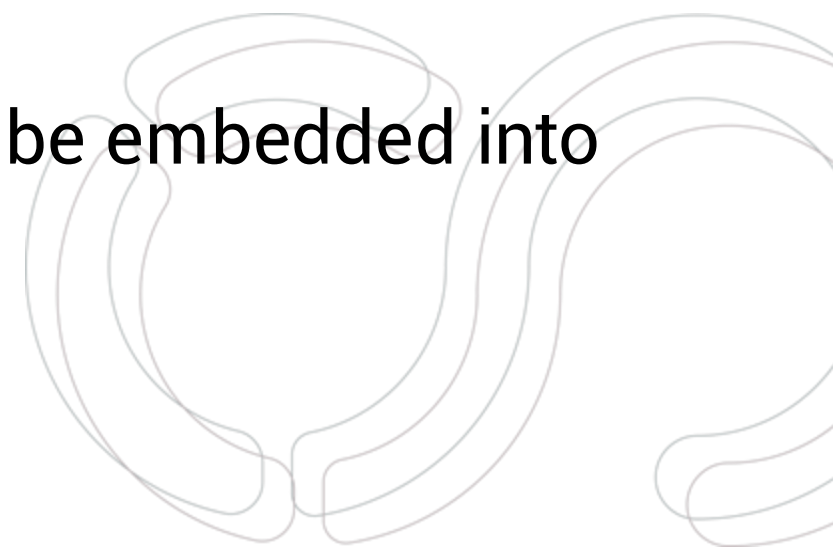- **It is possible to add your own**

Workflows

- Per community
- Permissions and requests
- Python
- Documentation and examples:
- https://nrp-cz.github.io/docs/customize/workflows

```python
can_read_draft = _can_read_anytime + [
    IfInState(
        ["draft", "submitted"],
        then_=[
            PrimaryCommunityMembers(),
            AccessGrant("preview"),
            AccessGrant("edit"),
            SecretLinks("preview"),
            SecretLinks("edit"),
        ],
    ),
]
```

Invenio User Interface

- Hybrid frontend architecture
  - Client side interaction with React (Semantic UI)
  - Server side with Flask and JinjaX

- Reuse as much as possible from Invenio RDM and CESNET Invenio
  - Library of Figma components

- Custom components (e.g. data previewers) can be embedded into the system

- Documentation: https://nrp-cz.github.io/docs/customize/repository_ui

Select the community where you want to submit your record.   **Select a community**

Files

☐ Metadata-only record ?

Storage available   0 out of 10 files   0 bytes out of 100.00 MB

Drag and drop files   - or -   ⬆ **Upload files**

⚠ File addition, removal or modification are not allowed after you have published your upload.

Basic information

▊ **Digital Object Identifier** *

Do you already have a DOI for this upload?   ◉ Yes, I already have one   ○ No, I need one

Copy/paste your existing DOI here...

A DOI allows your upload to be easily and unambiguously cited. Example: 10.1234/foo.bar

🏷 **Resource type** *

▼

🏷 **Title** *

Draft ⓘ

💾 Save draft   👁 Preview

⬆ Publish

↗ Share

🛡 **Visibility** *

Full record

Public   Restricted

Files only

Public   Restricted

🔓 **Public**
The record and files are publicly accessible.

Options

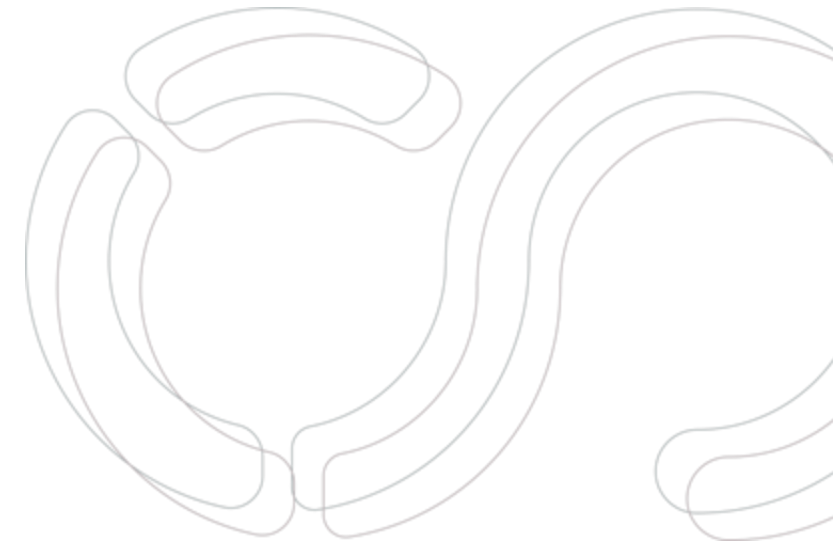☐ Apply an embargo ⏱
Record or files protection must be **restricted** to apply an embargo.

DevOps & Developing Custom Components

- Testing, Staging and Production Environment
- "On click" test deployment in progress
- Error logging (GlitchTip)
- GitHub
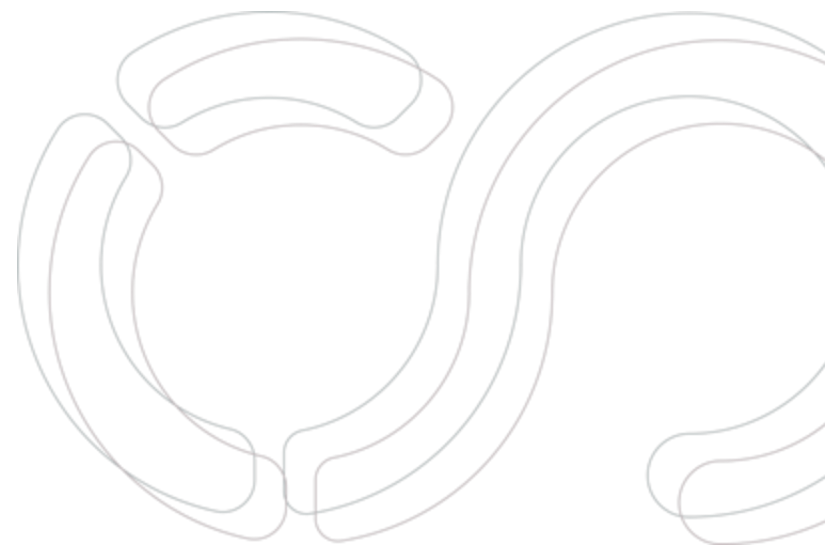 - Repository instance code
  - Contributed code

Milan Janíček

What is DSpace?

Key Features of DSpace

- Free open source software
- Completely customizable to fit user needs
- Manage and preserve all format of digital content (PDF, Word, JPEG, MPEG, TIFF files)
- Apache SOLR based search for metadata and full text contents

- UTF-8 Support
- Granular group based access control
- Optimized for Google Scholar indexing
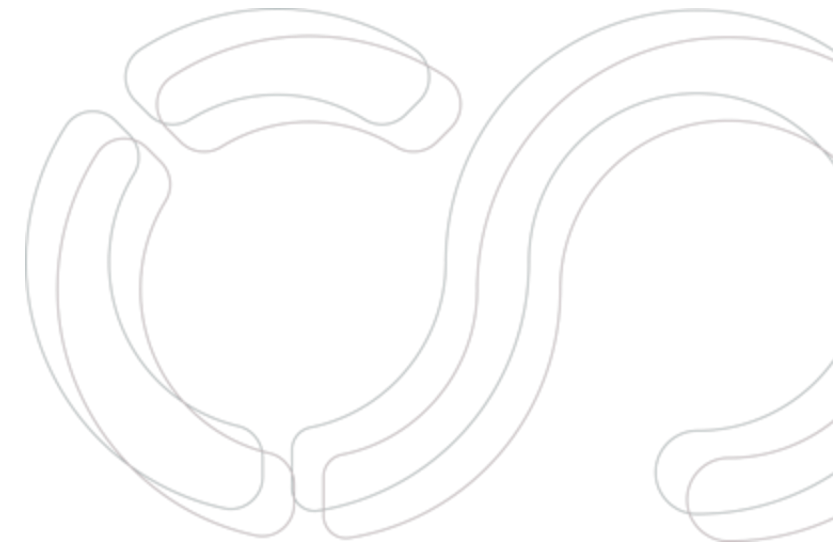- Interface available in 22 languages

https://dspace.org/features/

What is DSpace?

- core parts of the system
  - backend – API (Java), db (PostgreSQL), full-text search engine (SOLR), frontend (AngularJS)

- repository components
  - metadata (DublinCore, extensible fields)
  - authority control (incl. controlled vocabularies) - https://wiki.lyrasis.org/display/DSDOC7x/Authority+Control+of+Metadata+Values
  - PIDs (handle, DOI)
  - communities, collections & access control, worklows
  - APIs
  - data interchange protocols (OAI-PMH)

DSpace Versions

- [version<7]
  - old, different user interface

- [version>=7] - 7.6.x, 8.x, 9.x,
  - "modernized" - heavily refactored - version
  - separated backend (API) & frontend (JavaScript)
  - entities – extended data model
    - Person, Project, Publication, Journal

- you can find repositories built on different versions of DSpace
  - 1.x - https://dml.cz/ (Czech Digital Mathematics Library (DML-CZ))
  - 6.x - https://publications.cuni.cz/

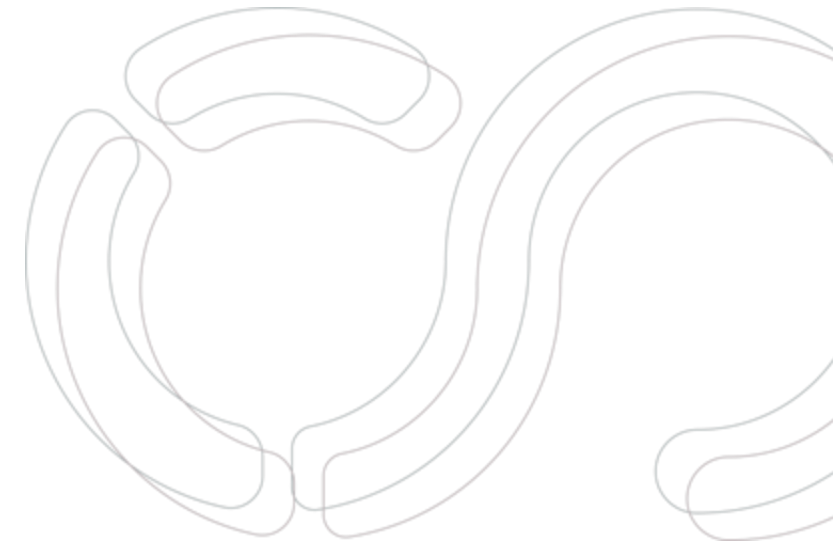 - 7.x - https://repozitar.techlib.cz/
  - 9.x - https://dspace.vut.cz/

DSpace-CRIS

- CRIS – current research information system (Czech equivalents – OBD, ASEP..)

- modification of DSpace, by 4science

- added CRIS functionality (including entities)
  - example -Technischen Universität Hamburg
  - further examples - listed here

- discussions about future DSpace & DSpace-CRIS merge are underway
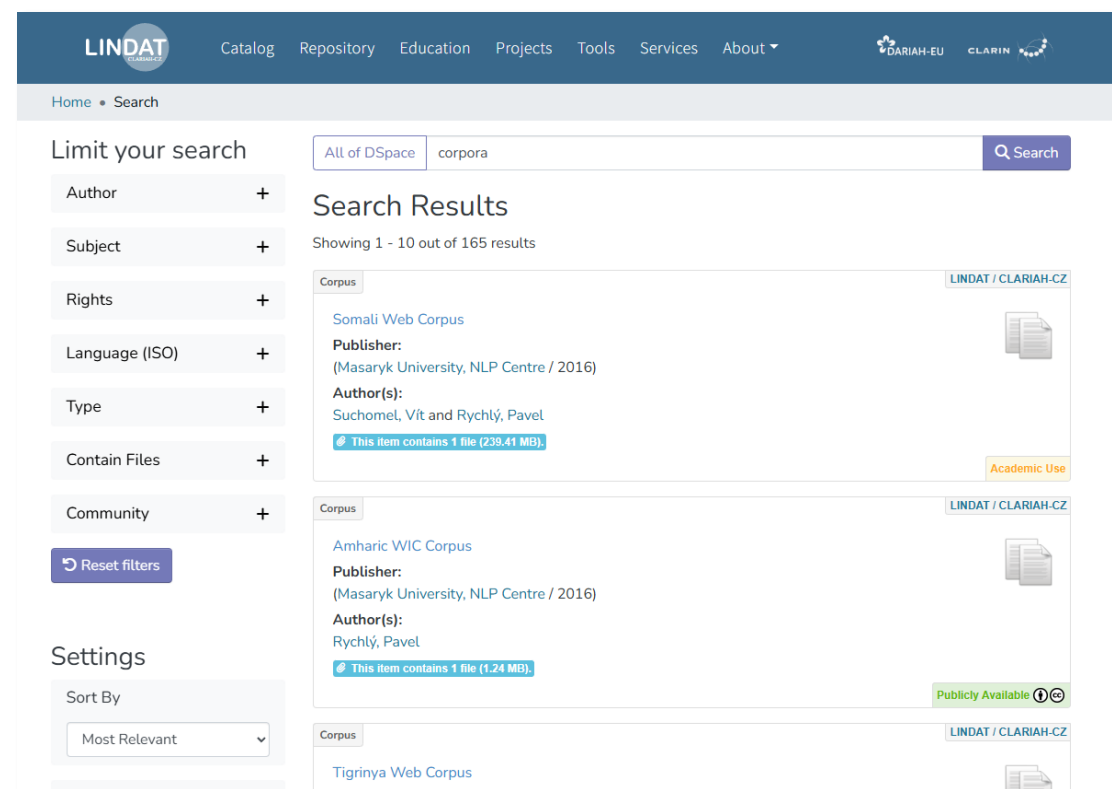
DSpace Ecosystem

- project lead
  - LYRASIS (USA, non-profit)

- service providers
  - Atmire (Belgium)
  - 4science (Italy)
  - The Library Code (Germany)
  - PCG Academia (Poland)
  - dataquest (Slovakia)
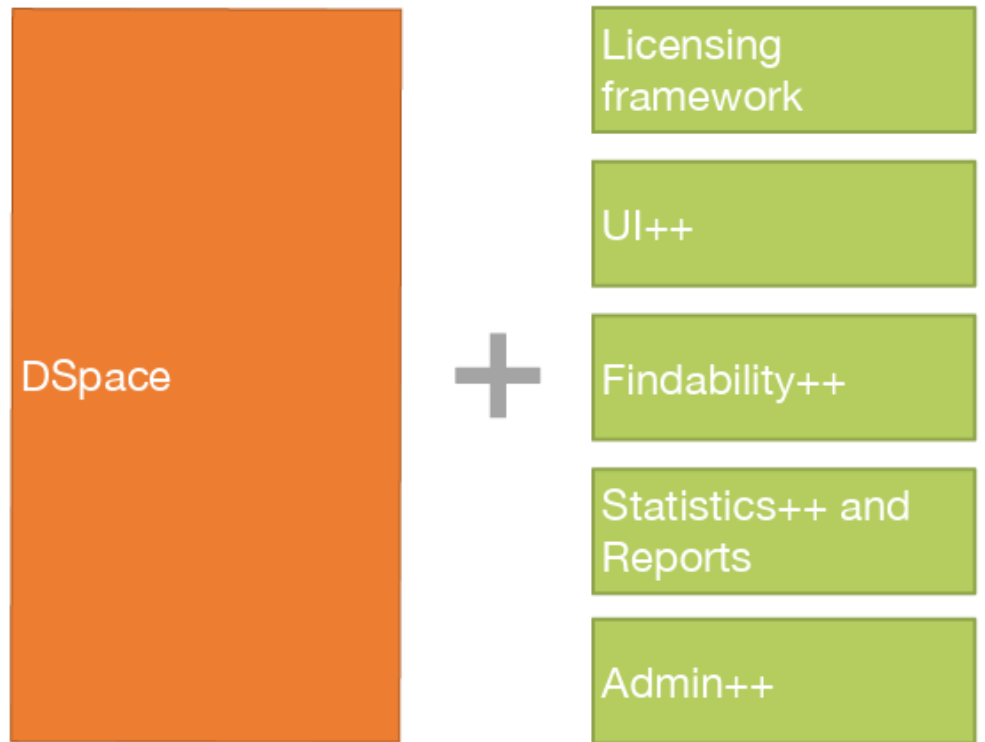 - …

CLARIN-DSpace

- LINDAT/CLARIAH-CZ Repository
  - linguistics data -> digital humanities
  - https://lindat.mff.cuni.cz/repository/home

- modified version (fork) of DSpace
  - https://github.com/ufal/clarin-dspace
  - currently based on 7.6.x DSpace

CLARIN-DSpace

DSpace

+

Licensing framework

UI++

Findability++

Statistics++ and Reports

Admin++

schema by Pavel Straňák

- examples of added functionality
  - licensing framework
    - research data in linguistics need special licenses
  - findability
    - data interchange (research data search services)
  - user interface
    - specific features for research data

CLARIN-DSpace

- there is no CLARIN-DSpace based pilot repository in NRP
  - … but there are other instances in production
  - ÚFAL has active experience with operating research data repository for many years

- DSpace is being developed by an active international community
  - new functionalities will be added also via DSpace/DSpace-CRIS development
  - NRP specific functionalities are mostly added on top of the original code
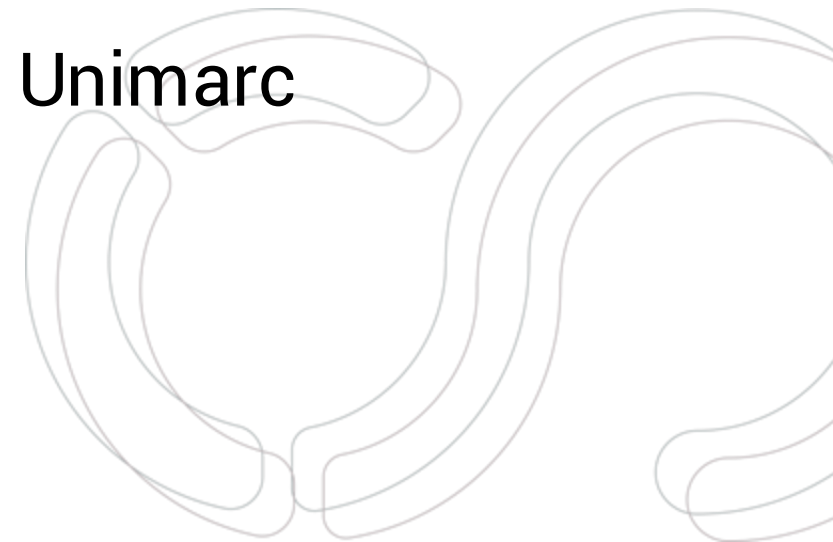
Eliška Blažková

What is ASEP Data repository platform?

- One of the three core repository platforms

- Developed by Library of the Czech Academy of Sciences

- ASEP data repository - part of a broader ecosystem:
  - Data repository
  - Bibliographic database
  - CRIS System (automated reporting to RIV)
  - Analytics module

- Data repository originally started out as an in-house module for Academy of Sciences in 2018

- Partially Open source (open front-end, partially open back-end)

Features

- Persistent identifiers: Handle and DOI
- Dataset versioning
- Ability to link dataset to it's research paper
- Metadata natively stored in Datacite compatible Unimarc
- Metadata exports to JSON, XML and CSV
- Rich metadata with Schema.org
- Supports Z39.50 and OAI-PMH protocols

New features, coming soon™

- Ability to archive DMPs, with PID (Handle), linking them with their source data
- Ability to do that directly from a DSW/FW instance
- New and improved workflow for publishing scientific data
- New controlled vocabularies - ROR, ORCID, SPDX licenses