EUDAT B2FIND
A Cross-Discipline Metadata Service and Discovery Portal

Heinrich Widmann, DKRZ
DI4R 2016,
Krakow, 28 September 2016

www.eudat.eu

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECTe-Infrastructures. Contract No. 654065
Outline

• EUDAT and the B2 Service Suite
• Guidelines and Concepts
• B2FIND – EUDAT’s Discovery Service
  • MD Ingestion and the B2FIND Schema
  • Disciplines, Communities and the MD catalogue
  • Data Access Identifiers
  • Discovery Portal
• Outlook and Summary
EUDAT and the B2 Service Suite
EUDAT

• The project European Data Infrastructure (EUDAT)
  • funded by the EU Horizon2020 program
  • started in 2011, now in 2\textsuperscript{nd} phase 'EUDAT2020', will end 2018
  • \( \geq 2018 \) : agreement of cooperation

• Motivation : Manage the rising tide of research data
  ➢ Improve Interoperability in a wide cross-disciplinary scope

• Objective : Build up a Collaborate Data Infrastructure,
  ➢ based on common data services
  ➢ driven by requirements of the research communities
B2 Service Suite

→ http://www.eudat.eu/services
Guidelines and Concepts
• **The FAIR principles**

  ➢ **B2FIND approach**

  • **Findability**: “Ease with which information can be found”
    ➢ Powerful and easy-to-use search features and functionalities
  • **Accessibility**: “Ability to access data stored within repositories”
    ➢ Unique and persistent identification and resolvability of data objects
  • **Interoperability**: “Ability of multiple systems with different structures to exchange data with minimal loss of content” (NISO)
    ➢ Comprehensive cross-disciplinary MD catalogue based on common standards and by minimising loss of information
  • **Reuseability**: “Ability to re-use data created by others”
    ➢ Cross-discipline approach and catalogue covering multiple sources
Levels of Interoperability

Heterogeneity

Research Communities (Data Provider)

Data Repositories (e.g. B2SHARE/B2SAVE or Aggregator as DataCite)

Service Provider (e.g. EUDAT-B2FIND)

Homogeneity

Schema A

Collect and extract MD

Schema B

B2FIND harvest and mapping

Information Loss

Info Loss

MD generation

MD generation

MD generation

B2FIND

DataCite
B2FIND
MD Ingestion and
Common Schema
B2FIND Ingestion Workflow

- **Data provider (Community)**
  - Harvest specification:
    - OAI-URL
    - OAI subsets
    - MD formats

- **MD Generation and Specification**
  - Harves specification:
    - OAI-URL
    - OAI subsets
    - MD formats

- **MD Harvesting**
  - Mapping specification:
    - XPATH rules
    - Community specific MD schemas and ...
  - For joining B2FIND only a few preconditons has to be fulfilled
    - Harvesting endpoint
    - Spec. of MD format
  - Guarantee data synchronisation by frequent and incremental data harvesting

- **Mapping and Validation**

- **Uploading and Indexer**

- **Search and Data Access**

- **User (Scientist or Researcher)**
  - For joining B2FIND only a few preconditions has to be fulfilled
    - Harvesting endpoint
    - Spec. of MD format
  - Guarantee data synchronisation by frequent and incremental data harvesting
<table>
<thead>
<tr>
<th>Metadata Type</th>
<th>B2FIND Field name</th>
<th>Allowed values</th>
<th>Semantic definition</th>
<th>Level of Obligation</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>General information</td>
<td>Title</td>
<td>Free text (unicode)</td>
<td>A name or title a resource is known</td>
<td>Mandatory</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Free text</td>
<td>Additional info</td>
<td>Recommended</td>
<td>0-1</td>
</tr>
<tr>
<td>Data Access</td>
<td>Source</td>
<td>Valid URL or URN</td>
<td>Unique link to data resource</td>
<td>Mandatory (1)</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td>PID</td>
<td>Persistent Identifier</td>
<td>+ persistent and resolvable</td>
<td></td>
<td>1-3</td>
</tr>
<tr>
<td></td>
<td>DOI</td>
<td>Digital Object Identifier</td>
<td>+ citable</td>
<td></td>
<td>0-1</td>
</tr>
<tr>
<td>Provenance data</td>
<td>Creator</td>
<td>‘;’-sep. list of names</td>
<td>Main researchers involved in data prod.</td>
<td>Recommended</td>
<td>0-n</td>
</tr>
<tr>
<td></td>
<td>Discipline</td>
<td>List of values from CV (Controlled Vocab)</td>
<td>Field of research</td>
<td>Recommended</td>
<td>0-n</td>
</tr>
<tr>
<td>Publication Year</td>
<td>YYYY</td>
<td></td>
<td>The year data are published</td>
<td>Recommended</td>
<td>1</td>
</tr>
<tr>
<td>Formal data</td>
<td>Temporal Coverage</td>
<td>Interval of 2 DTimes [ Begin, End ]</td>
<td>The temporal limits of a date-time</td>
<td>Optional</td>
<td>1-n</td>
</tr>
<tr>
<td></td>
<td>Spatial Coverage</td>
<td>Spatial box or point [[minlat,minlon...]]</td>
<td>The spatial limits of a place</td>
<td>Optional</td>
<td>1-n</td>
</tr>
</tbody>
</table>
B2FIND
Disciplines, Communities and
MD Catalogue
The Facet 'Discipline'
Controlled Vocabulary

“Fields of Knowledge”
/

Humanities
  - Arts
  - History
  - Linguistics

Social sciences
  - Archaeology

Natural sciences
  - Biology
  - Physics
  - Earth Sciences
  - Elementary Particle Physics

Professionals
  - Engineering
  - Material science
  - Crystallography

taken from “List of Academic disciplines”
Coverage of Disciplines in B2FIND

- Humanities: 51%
- Social Sciences: 10%
- Natural Sciences: 8%
- Not Stated: 29%
- Professions: 2%
B2FIND MD Catalogue

Ingestion status

- 17 communities
- > 450,000 MD records
B2FIND
Data Access
## Data Access Identifiers

### Resolvability and 'Levels of aggregation'

The image shows a diagram illustrating the relationship between data collection, resolution, and access through various identifiers. The diagram includes a table comparing different types of identifiers:

<table>
<thead>
<tr>
<th>Type</th>
<th>Unique</th>
<th>Persistent</th>
<th>Resolvable</th>
<th>Citable</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PID</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>URL (Source)</td>
<td>✓</td>
<td>?</td>
<td>?</td>
<td>✗</td>
</tr>
</tbody>
</table>

The table indicates that DOI is the most robust identifier, offering both unique and persistent characteristics, and enabling resolvability and citability. PID also has unique and persistent properties, but it lacks citability. URL (Source) includes only unique properties, and neither resolvability nor citability is indicated.

The diagram also highlights the transition from B2FIND metadata to resource through XML, with stricter policies governing the use of these identifiers.
Coverage of Data Access Identifiers

- DOI + PID: 450 (1%)
- DOI: 123370 (28%)
- PID: 121959 (27%)
- URL: 197931 (45%)
B2FIND
Discovery Portal
B2FIND provides ‘faceted’ search for:
- Free text
- Geo spatial
- Temporal coverage
- Publication year
- Textual facets as:
  - Tags
  - Creator
  - Discipline etc.

Dataset view provides display of metadata:
- Spatial extent
- Table of field-value pairs
- Links to data resources
Outlook and Summary
Outlook

• Handle scalability and granularity issues
  ➢ ‘Levels of aggregation’
• Metrics for Key Indicators and Metadata Quality
  ➢ Establish content-related quality assurance
• Add further search and distribution channels, e.g.
  ➢ Use linked data : Potential for semantic enrichment
  ➢ ‘Annotation’ functionality : Users link datasets to external reference materials (vocabularies, ontologies, etc.)
  ➢ Query-based Taxonomies : Enabling hierarchical search, e.g. in trees of ‘Disciplines’
Summary

- **EUDAT-B2FIND**
  - established an operative service based on agreed standards and guidelines as the FAIR principles,
  - provides a discovery portal with powerful search functionalities and
  - is based on a unique catalogue of research data, combining many heterogeneous and cross-discipline sources
- Improved interoperability is achieved by homogenisation to a common metadata schema
- Further efforts are made to address the demands of the communities and data projects, to adapt the system for future challenges
Thank you for your attention!

Links:
- info: http://eudat.eu/b2find
- portal: http://b2find.eudat.eu

Contact
- www.eudat.eu/support-request
- widmann@dkrz.de