EUDAT-B2FIND
A FAIR and Interdisciplinary Discovery Portal for Research Data

Heinrich Widmann, DKRZ
Claudia Martens, DKRZ
Open Science Days,
Berlin, 17 October 2017
Outline

• What is EUDAT-B2FIND?
• Which guidelines are followed?
• How we tackle the issue ‘diversity vs. interoperability’?
• How is the service implemented?
• Which data and domains does B2FIND cover?
• How are the metadata collected?
• How can you search the Discovery Portal?
What is EUDAT?

- The project EUDAT (see http://eudat.eu) established a Common Data Infrastructure providing several generic services for interdisciplinary data management, driven by the communities.
What is EUDAT-B2FIND?

• B2FIND is the metadata service of EUDAT consisting of
  • An interdisciplinary metadata catalogue that spans a large number heterogeneous datasets
    • harvested from various research communities and stored through the EUDAT service B2SHARE
    • covering a wide range of highly diverse disciplines
      ➢ mapped to a common and unified schema
  • An open search portal allowing researchers
    • to find collections of scientific resources in a wide spread and cross-domain search space
    • to access those resources through the given references in the metadata
Which guidelines are followed?
The FAIR principles → B2FIND approach

- **Findability** → *Discovery Portal* with powerful search features based on rich metadata catalogue

- **Accessibility** → *Persistent Identifiers* for unique resolvability of data objects

- **Interoperability** → *Interdisciplinary Catalogue* based on *Common Standards*, *Vocabs* and *Schema*

- **Reuseability** → *Licenses*, *Provenance* and *Domain-relevant Information* is provided
How can data interoperability be improved?

Levels of Interoperability

- **Heterogeneity**
  - Research Communities (Data Provider)
  - Cross-disciplinary Data Repositories (e.g., B2SHARE or DataCite)
- **Homogeneity**
  - Service Provider (e.g., EUDAT-B2FIND)

**Mandatory**

**Optional / Recommended**

**Discipline specific**

- Schema A
  - Collect and extract MD
- Schema B
  - Harvest MD
- Schema C
  - Extented templates for Domain specific fields

- B2FIND harvest and mapping
How is the service implemented?

- uses standard APIs (as OAI-PMH) to harvest from various data providers
- performs a adaptable mapping of the domain specific metadata onto the unified B2FIND schema
- is based on CKAN (see ckan.org), an open source repository and portal software
  - CKAN comes with a SOLR indexer, a Postgres DB and a power and RESTful API
  - Extensible by open available and own developed CKAN extensions
<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><strong>Mandatory</strong></td>
<td>1</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td>Free text</td>
<td>Additional info</td>
<td><strong>Recommended</strong></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></td>
<td>0-1-1-3</td>
</tr>
<tr>
<td></td>
<td>PID</td>
<td>Persistent Identifier</td>
<td>+ persistent and resolvable</td>
<td><strong>Mandatory (1)</strong></td>
<td></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><strong>Recommended</strong></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><strong>Recommended</strong></td>
<td>0-n</td>
</tr>
<tr>
<td></td>
<td>Publication Year</td>
<td>YYYY</td>
<td>The year data are published</td>
<td><strong>Recommended</strong></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><strong>Optional</strong></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><strong>Optional</strong></td>
<td>1-n</td>
</tr>
</tbody>
</table>
Which data and domains does B2FIND cover?

Ingestion Status of the MD Catalogue

- 19 communities
- > 600000 MD records
How can you search the Discovery Portal?

B2FIND Discovery Portal
Faceted Search and Data Access

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
How are the metadata collected?

B2FIND Ingestion Workflow

→ https://github.com/EUDAT-Training/B2FIND-Training

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

**MD Provider**

**Data provider (Community)**

**MD Generation and Specification**

**MD 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
Summary

EUDAT-B2FIND
• established an operative service based on agreed standards and guidelines as the FAIR principles
• provides a discovery portal with powerful faceted search functionalities
• is based on a joint catalogue of research data, combining many heterogeneous and cross-discipline sources
• achieves improved interoperability by homogenisation to a common metadata schema
• follows a low barrier approach for data providers to expose metadata to B2FIND
Thank you for your attention!

Links:
- Info and Docs: [http://eudat.eu/b2find](http://eudat.eu/b2find)
- Guidelines for data providers: [http://b2find.eudat.eu/guidelines](http://b2find.eudat.eu/guidelines)
- B2FIND Portal: [http://b2find.eudat.eu](http://b2find.eudat.eu)
- Support: [www.eudat.eu/support-request](http://www.eudat.eu/support-request)
- Contact: widmann@dkrz.de, martens@dkrz.de