



B2FIND: EUDAT Metadata Service

Daan Broeder, et al.
EUDAT Metadata Task Force





EUDAT Joint Metadata Domain of Research Data

- Deliver a service for searching and browsing metadata across communities
 - Appropriate terminology for users of all disciplines when specifying queries – *possibly adaptive?*
 - Access to the data when allowed – *single auth/autz?*
 - Useful visualization of results – *community provided?*
 - Commenting facility to exchange experiences
- Use existing technologies: OAI-PMH, SOLR/Lucene, etc.
- Expected challenges
 - Suitable catalog and indexing system for >> 1M records
 - Semantic interoperability problems
 - Granularity issues



Overall plan

- Import metadata from other EUDAT services: B2SHARE, B2SAFE
- Look for stable metadata providers from communities
 - EUDAT core communities: ENES, CLARIN, EPOS
 - other interested communities: GBIF, CESSDA, BBMI,...
 - *other projects aggregating metadata: DataOne, DataCite, Europeana*
 - community input:
 - What are useful dimensions for searching & browsing?
 - What are useful metadata collections?
- Also outreach to emerging communities
 - Help setup a metadata infrastructure, harvest their metadata ...

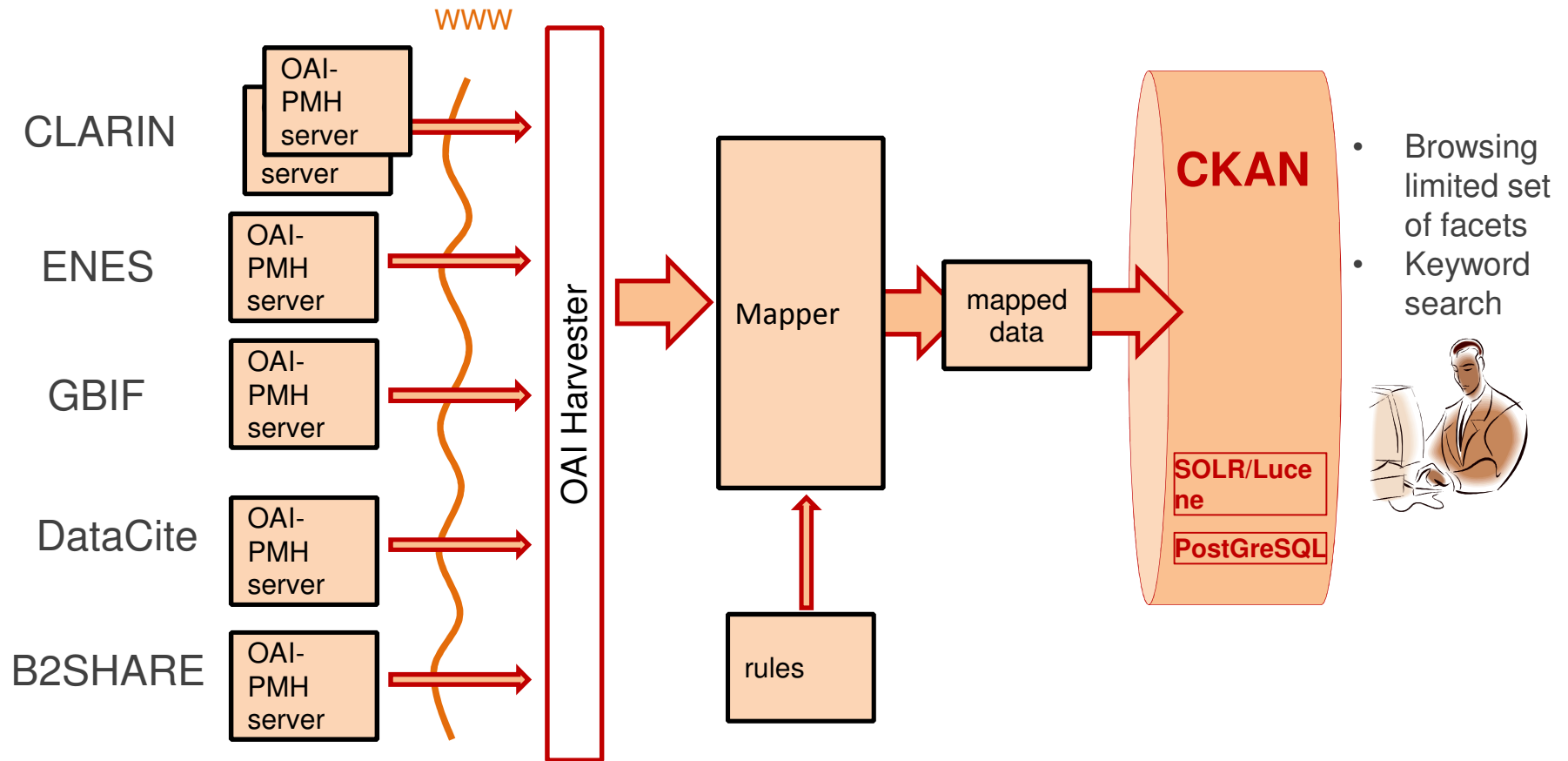


EUDAT Metadata Catalog version II

- Using CKAN as catalog software
 - Open Knowledge Foundation software
 - Choice made after some appraisals: large community, available documentation, proven track record
 - All should be modular & pluggable as much as possible
 - Scalability testing is still in progress 2M records seems ok for searching, *but not for importing!*
 - EUDAT will still be investigating other catalog technologies
- Working on adapting CKAN to our needs:
 - Better GUI: accurate temporal search specification, taxonomies, ...
- Priorities:
 - Increase user experience -> metadata quality + ...
 - Include more communities



B2FIND Architecture





B2FIND Faceted Browser

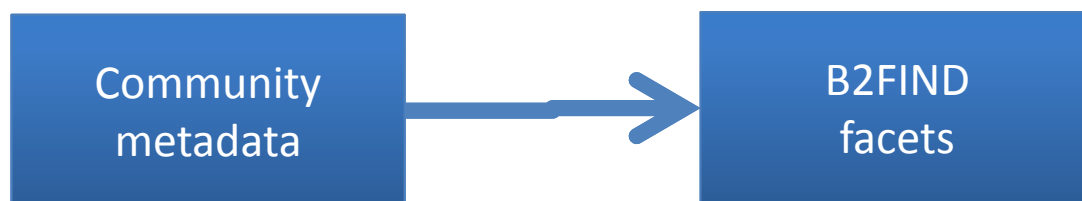
- Facets:
 - title, author, discipline, organization, publication year, format, language
- Geospatial search interface
- Full text search on whole metadata record
- Current Communities:
 - B2SHARE: EUDAT simple store
 - CLARIN: linguistics
 - ENES: Climatology
 - GBIF: Bio Diversity
 - DataCite: registry for DOI identified data



B2FIND

Faceted browsing

- Most faceted browsing implementations use SOLR/Lucene



- Requires translation of information like:
... <Creator>Tom Mueller</Creator>
into
... facetname=Author, value="Tom Mueller"



Metadata Quality

- Problematic quality
 - encoding of values even within one community is not always coherent e.g. even clarin->language
- No single static mapping will give a good user experience
 - Sparsely filled in records
 - Facets need to be filled or records become invisible
 - e.g. “Author” in CLARIN metadata is difficult to fill and needs to derive from actor information in different roles
- Therefore **if;then;else** constructs are tried



Flexible Mapping in JMD

- Objectives
 - Extensible
 - None of mapping semantics is “hardcoded”
 - Editing does not require advanced programming skills
- Implementation
 - Java based engine
 - Mappings defined by simple XML files
 - Mainly based on XPath expressions
 - Evaluated in a chain: try matching until a non-empty result is achieved



Example Mapping Types

- Most mappings simply extract an element
 - Empty if element is undefined, so proceed to next
- Complex join operations
 - e.g. to generate value of author facet, join values of “author” and “originator” in the source
 - same person(s) may be listed in both, so remove duplicates
- Conditional operations
 - For example, used to skip unneeded values like “Unspecified” in source

B2Find EUDAT Metadata Portal

The screenshot shows the B2Find EUDAT Metadata Portal website. The browser address bar displays `eudat-jmd.dkrz.de`. The page features a dark blue header with the B2FIND logo and navigation links for Datasets, Communities, and About. A search bar is present in the header. The main content area includes a welcome message, a search box with the example text "eg. IPCC", and a carousel for B2SHARE. The footer contains links for learning more about EUDAT, legal notices, and contact information, along with logos for the European Union and the Seventh Framework Programme.

Search for a Dataset - EUDAT M... x Search for a Dataset - EUDAT M... x DDI Lite (Recommended Elemen... x Welcome - EUDAT Metadata Re... x +

eudat-jmd.dkrz.de

Most Visited Getting Started Latest Headlines Apple Yahoo! Google Maps YouTube Wikipedia News Biking Popular Work calibre library...

Log in

B2FIND Datasets Communities About Search

Welcome to the EUDAT Metadata Repository

The Collaborative Data Infrastructure - a framework for the future

Contains harvested metadata from EUDAT communities

Search Your Data

eg. IPCC

Popular Tags Meertens collection... Dataset Text

B2SHARE

Learn more about EUDAT
Legal notice
Contact us

eudat-jmd.dkrz.de/#indexCarousel

Find: metadata Next Previous Highlight all Match case

B2Find communities

Search for a Dataset - EUDAT M... x Search for a Dataset - EUDAT M... x [ddi: DDI Lite \(Recommended Elemen...](#) x Communities - EUDAT Metadat... x +

eudat-jmd.dkrz.de/group

Most Visited Getting Started Latest Headlines Apple Yahoo! Google Maps YouTube Wikipedia News Biking Popular Work calibre library...

Log in

B2FIND Datasets **Communities** About Search

Communities

EUDAT represents a unique partnership between research communities and data centers.

It brings together data service providers and users who are directly involved with the design of data services.

Search communities...

5 communities found Order by: Name Ascending

CLARIN
The Common Language Resources and Technology Infrastructure (CLARIN) project...
35673 Datasets

DataCite
DataCite is a not-for-profit organisation formed in London on 1 December...
45789 Datasets

ENES
The European Network for Earth System modelling (ENES) provides information...
265 Datasets

EUDAT Simple Store
The European Data Infrastructure (EUDAT) will offer a Simple Store service...
0 Datasets

GBIF
The Global Biodiversity Information Facility (GBIF) was established by...
11090 Datasets

Find: metadata Next Previous Highlight all Match case

B2Find communities

The screenshot shows a web browser window with the URL `eudat-jmd.dkrz.de/dataset`. The page features a dark blue header with the B2FIND logo and navigation links for Datasets, Communities, and About. A search bar is present in the header. Below the header, the main content area is titled "Datasets" and includes a sidebar with filters for location (a world map) and communities (DataCite, CLARIN, GBIF, ENES). The main results area shows a search bar, a "92,817 datasets found" count, and an "Order by: Relevance" dropdown. The first three results are listed with their IDs and descriptions:

- Svan_folklore12**
This dataset has no description
- Van den oei, joei, joei**
This dataset has no description
- Kleine jongen / André Hazes. - [S.l.] : EMI [etc.], 2001. - 1 cd. - 724353145...**
This dataset has no description

Below these, two more results are partially visible:

- 52781d4c-01c9-5cb8-8548-099feaf24ef2**
This dataset has no description
- ef72f993-e310-5ae3-9e57-f8d5d02a1e35**
This dataset has no description

The footer of the browser window shows search controls: "Find: [input]", "Next", "Previous", "Highlight all", and "Match case" (checked).



B2FIND Future

- New communities
 - EPOS is a EUDAT core community, we are waiting on their OAI metadata provider
 - BBMRI (Bioinformatics) Sweden (considering using OAI), still refining their schema other BBMRI members are using other approaches.
 - CESSDA (Social Sciences) probably included in collaboration with DASISH project
- Commenting function
- CKAN GUI elements
 - Better more specific temporal search
 - Hierarchical taxonomy based search
- Ever better mapping rules, but there is a limit!



Thank you for your attention



B2FIND



```
file:///Users/la...-imdprofile.xml  
- <xpath>  
  string-join(distinct-values(//cmd:CMD/cmd:Components/cmd:Session/cmd:MDGroup/cmd:Actors  
/cmd:Actor[cmd:Role = 'Author' or cmd:Role = 'Collector' or cmd:Role = 'Researcher' or cmd:Role = 'Annotator' or  
cmd:Role = 'Filmer' or cmd:Role = 'Recorder']/cmd:Name/text()), ';')  
- <xpath>  
  substring-before(substring-after(//cmd:CMD/cmd:Components/cmd:Session/cmd:Resources... [1]  
/cmd:descriptions/cmd:Description[@LanguageId = 'ISO639:eng' or @LanguageId = '']/text()), 'Author: ')  
</xpath>  
- <xpath>  
  //cmd:CMD/cmd:Components/cmd:Song/cmd:Autnors/cmd:Author/cmd:Name/cmd:Original/text()  
</xpath>  
- <xpath>  
  //cmd:CMD/cmd:Components/cmd:SongScan/cmd:Authors/cmd:Author/cmd:Name/cmd:Original/text()  
</xpath>  
</field>  
- <!--  
  All other visible fields, shown in table 'Additional info'  
-->  
- <!--  
  The date field is unstructured. Heuristically, take the first  
  group of 4 digits we see. This will work for both dd-mm-yyyy and  
  yyyy-mm-dd.  
-->  
- <field name="Publication Year">  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Components/cmd:Session/cmd:Date/text(), '[12]d{3}')) then replace(//cmd:CMD  
/cmd:Components/cmd:Session/cmd:Date/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Header/cmd:MdCreationDate/text(), '[12]d{3}')) then replace(//cmd:CMD/cmd:Header  
/cmd:MdCreationDate/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
</field>  
- <field name="Language">  
- <xpath>
```

```
- <!--  
  The date field is unstructured. Heuristically, take the first  
  group of 4 digits we see. This will work for both dd-mm-yyyy and  
  yyyy-mm-dd.  
-->  
- <field name="Publication Year">  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Components/cmd:Session/cmd:Date/text(), '[12]d{3}')) then replace(//cmd:CMD  
/cmd:Components/cmd:Session/cmd:Date/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Header/cmd:MdCreationDate/text(), '[12]d{3}')) then replace(//cmd:CMD/cmd:Header  
/cmd:MdCreationDate/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
</field>  
- <field name="Language">  
- <xpath>
```

```
<!--  
  The date field is unstructured. Heuristically, take the first  
  group of 4 digits we see. This will work for both dd-mm-yyyy and  
  yyyy-mm-dd.  
-->  
- <field name="Publication Year">  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Components/cmd:Session/cmd:Date/text(), '[12]d{3}')) then replace(//cmd:CMD  
/cmd:Components/cmd:Session/cmd:Date/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
- <xpath>  
  if (matches(//cmd:CMD/cmd:Header/cmd:MdCreationDate/text(), '[12]d{3}')) then replace(//cmd:CMD/cmd:Header  
/cmd:MdCreationDate/text(), '.*([12]d{3}).*', '$1') else ""  
</xpath>  
</field>  
- <field name="Language">  
- <xpath>
```



B2FIND



Next Steps for mapping

- Improve mapping quality
 - We track coverage (i.e. percentage of all metadata records where a value is mapped for a specific facet)
 - Ranges from around 50% to 100% due to heterogeneity of sources
 - Target over 90% for every facet for every community:
 - No insurance for correctness
- Add other mapping types
 - Component-based metadata (e.g. CMDI) is not well suited to XPath based mappings
 - Concept registry based mapping type is planned



Who is responsible for metadata quality?

- In shared research infrastructures this is especially challenging: **center -> community infra -> EUDAT infra**
- Community metadata providers are first responsible
 - We get often VERY bad metadata
 - How to improve this?
- For fast progress no other course than do some curation at service provider (EUDAT) side
- For proper curation & mapping expertise is needed. Who is interested in doing this?
- Is there a business model possible to make this work sustainable