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

EUDAT B2FIND

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 …

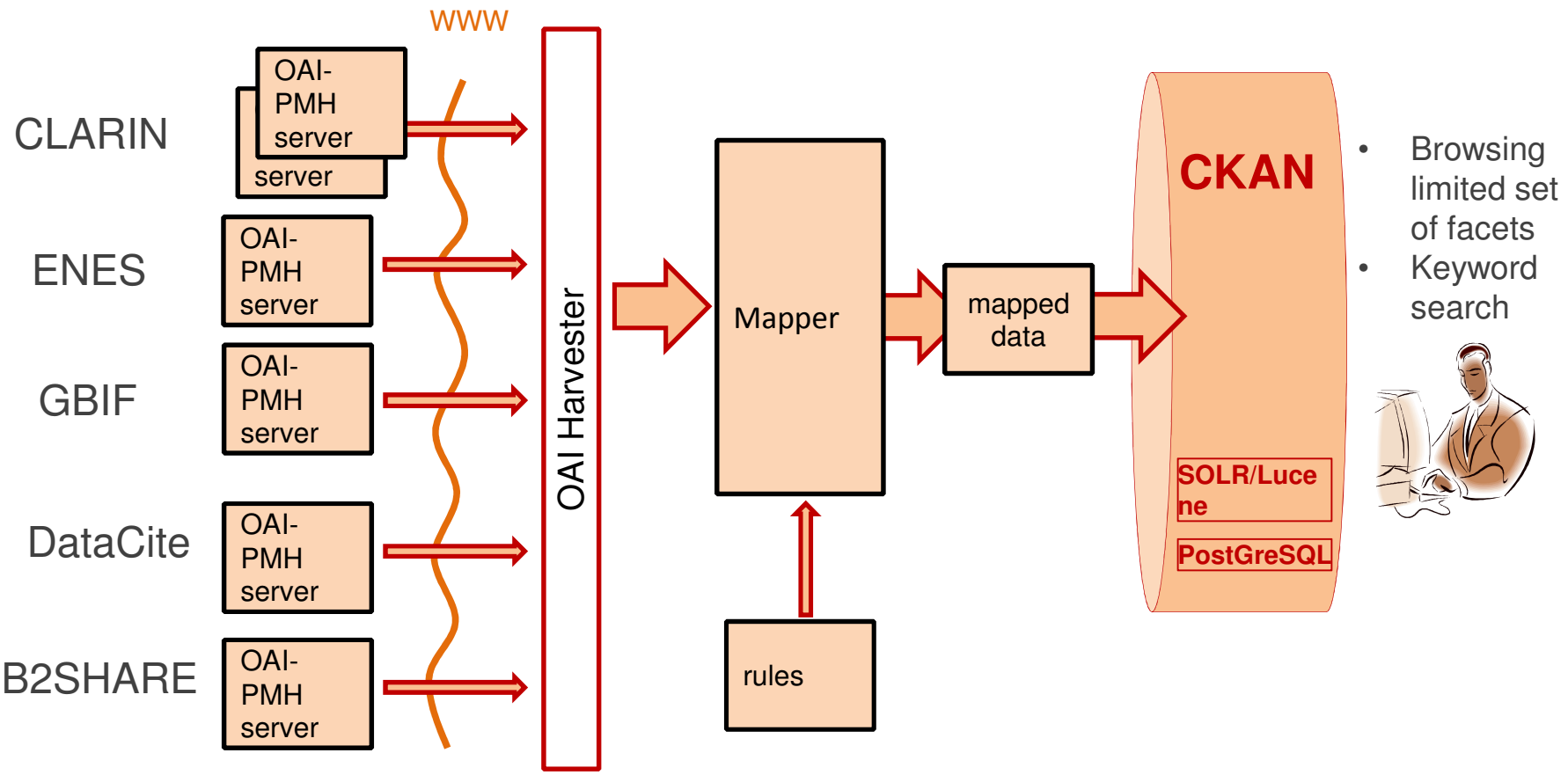
EUDAT B2FIND

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

EUDAT B2FIND

B2FIND Architecture



EUDAT B2FIND

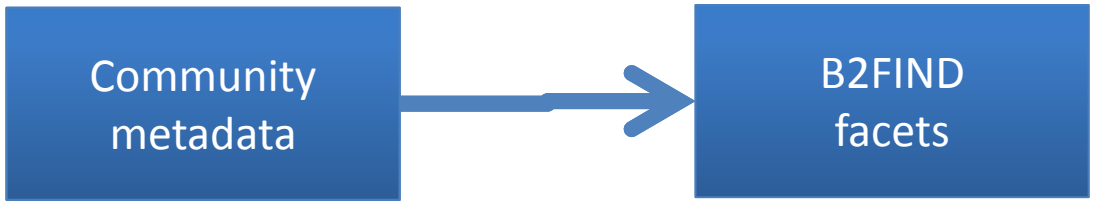
B2FIND Faceted Browser

- 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

EUDAT B2FIND

Faceted browsing

- Most faceted browsing implementations use SOLR/Lucene



- Requires translation of information like:

… <Creator>Tom Mueler</Creator>

into

… facetname=Author, value="Tom Mueler"

EUDAT B2FIND

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

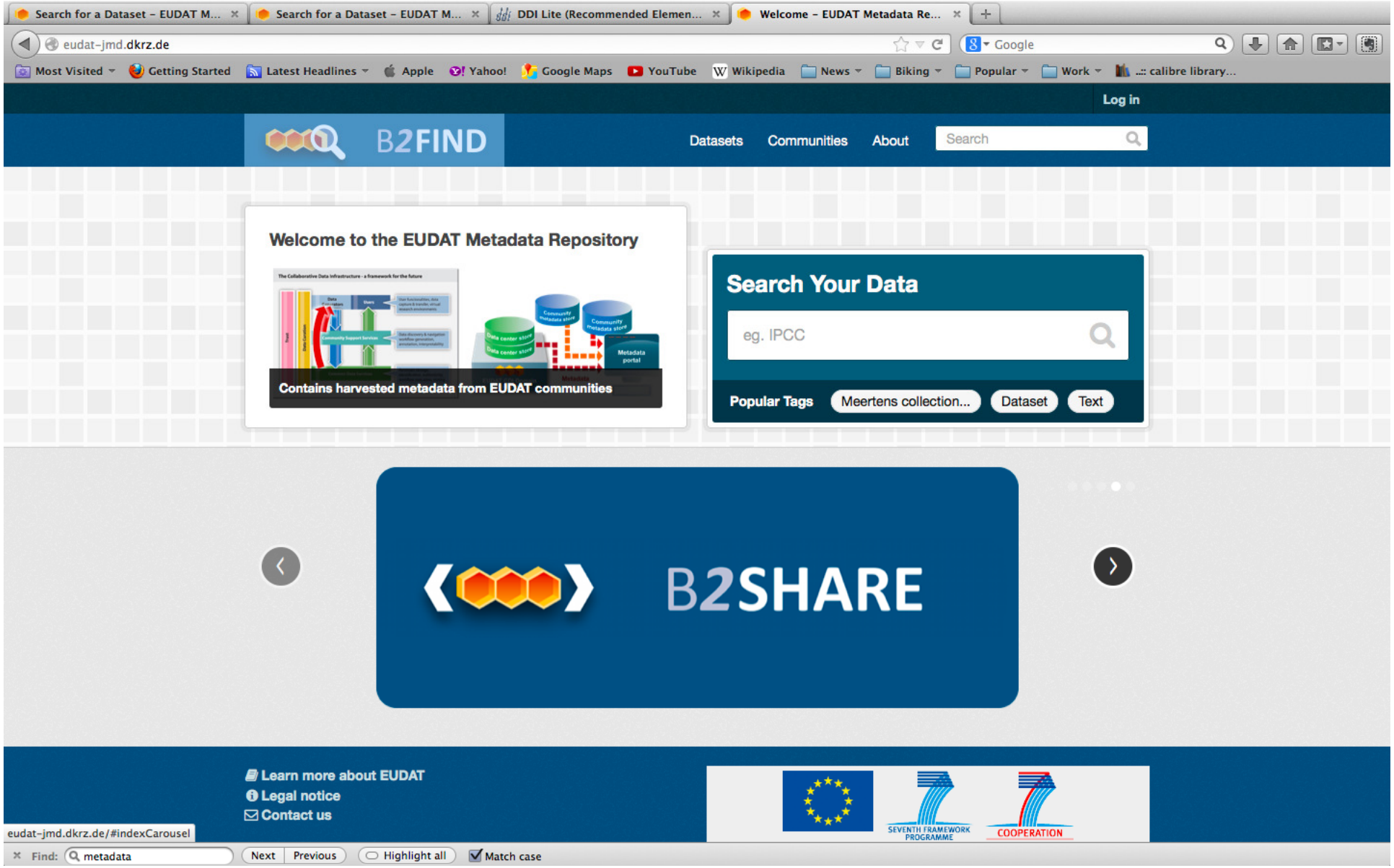
- Objectives
  - Extensible
  - None of mapping semantics is “hardcoded”
  - Editing does not require advanced programming skills
- Implementation
  - Java based engine
  - Mappings defined by simple XML files
  - Mainly based on XPath expressions
  - Evaluated in a chain: try matching until a non-empty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

EUDAT B2FIND

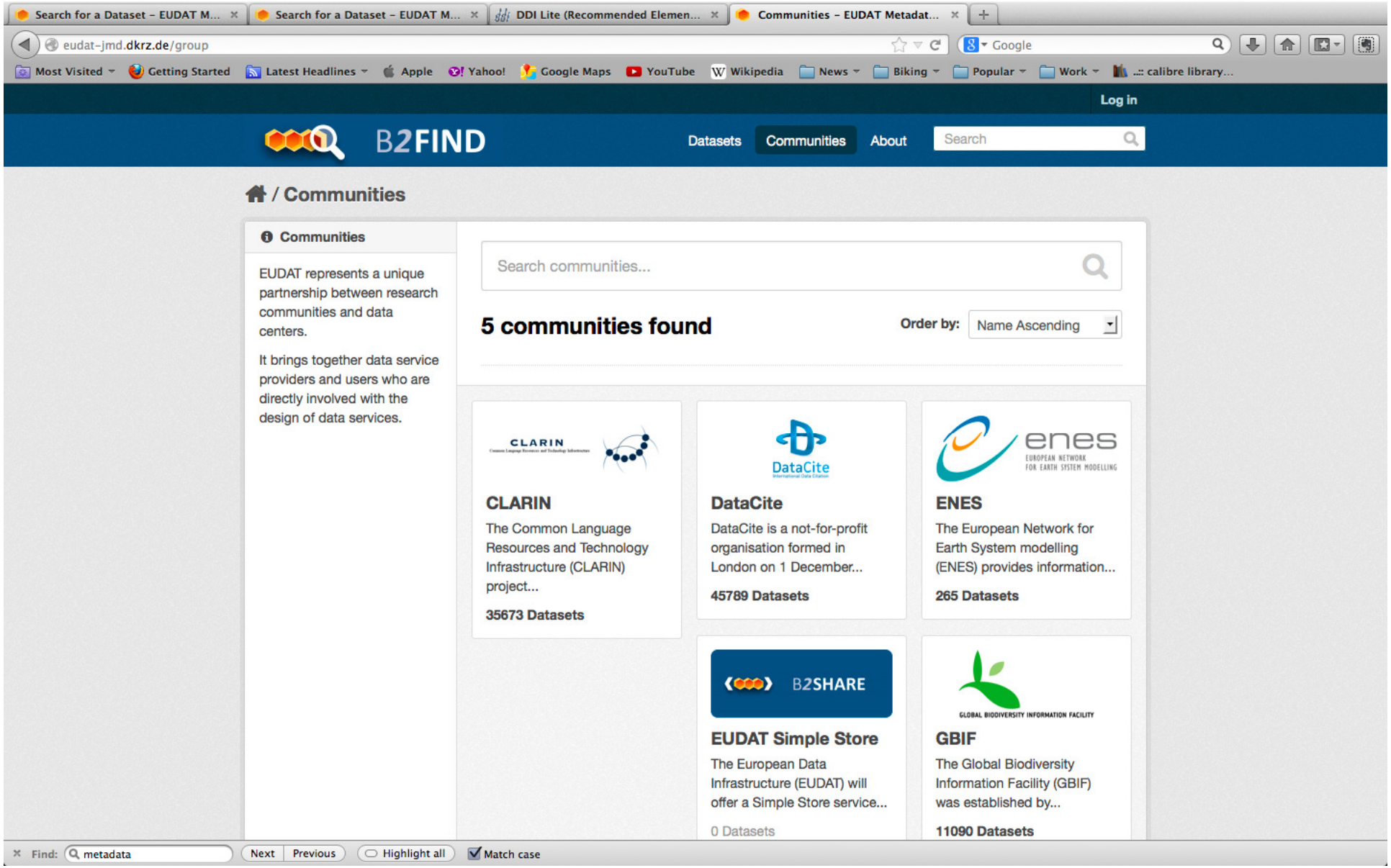
Example Mapping Types

- 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

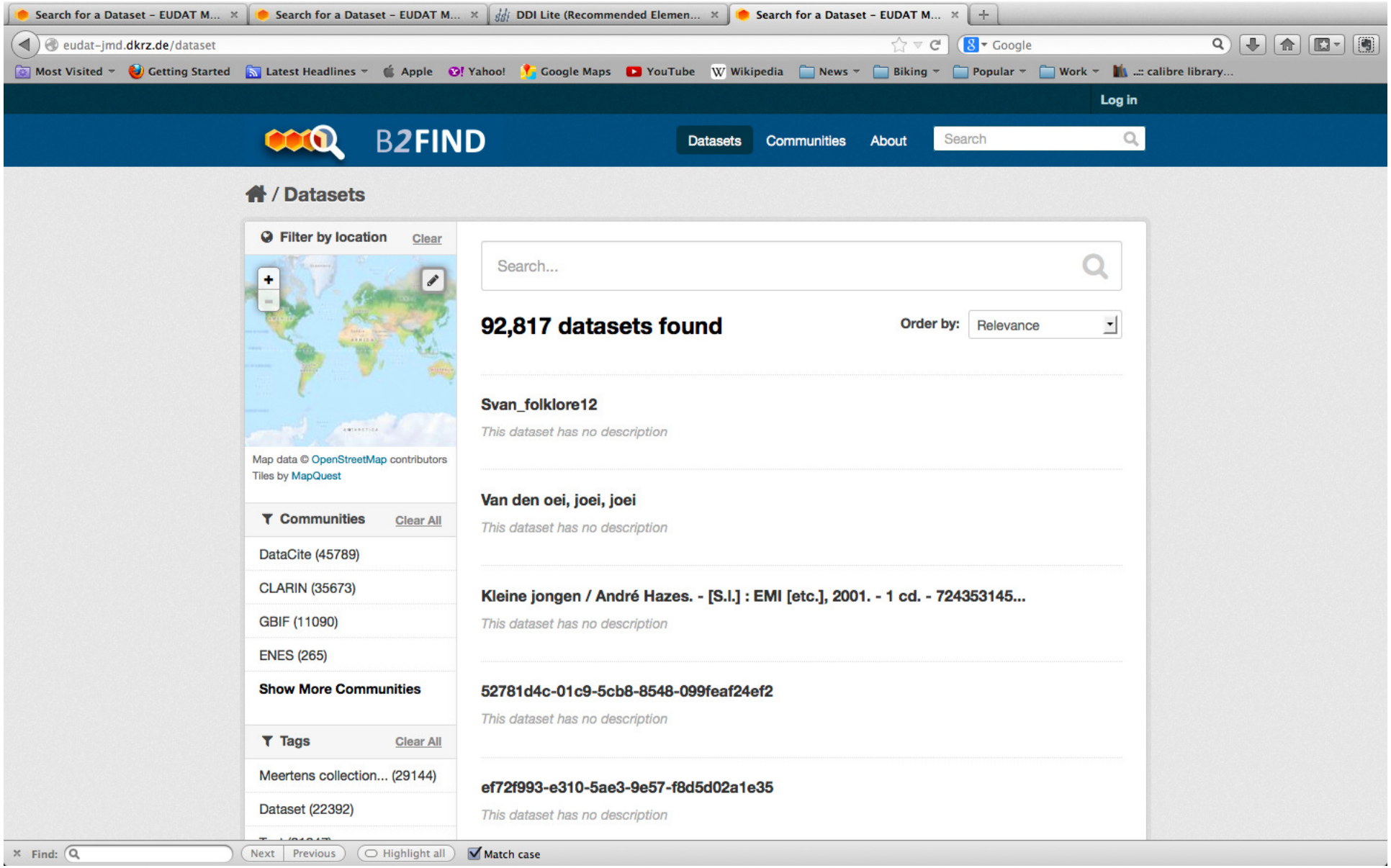
B2Find EUDAT Metadata Portal



B2Find communities



B2Find communities



B2FIND Future

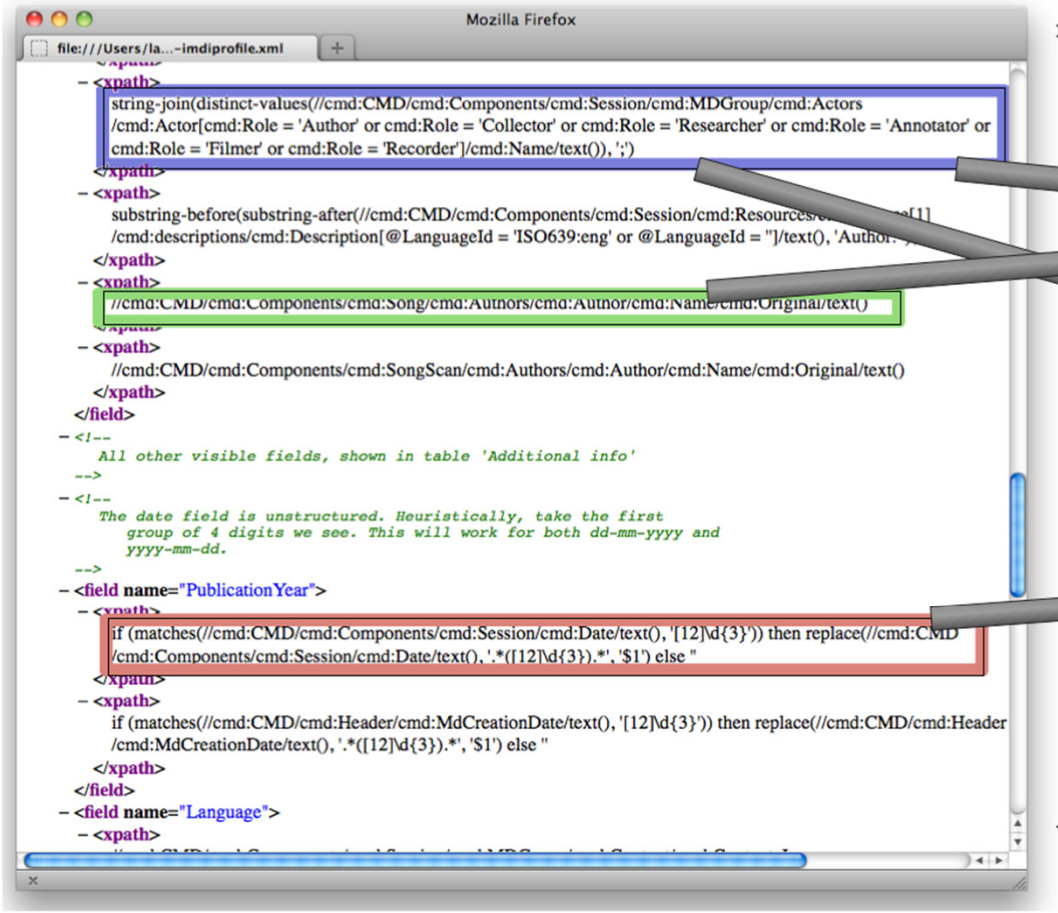
- New communities
  - EPOS is a EUDAT core community, we are waiting on their OAI metadata provider
  - BBMRI (Bioinformatics) Sweden (considering using OAI), still refining their schema other BBMRI members are using other approaches.
  - CESSDA (Social Sciences) probably included in collaboration with 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!

EUDAT B2FIND

Thank you for your attention

EUDAT

B2FIND


Mozilla Firefox
file:///Users/la...-imdiprofile.xml
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()), ';')
substring-before(substring-after(//cmd:CMD/cmd:Components/cmd:Session/cmd:Resources/
/cmd:descriptions/cmd:Description[@LanguageId = 'ISO639:eng' or @LanguageId = '']/text(), 'Author:
//cmd:CMD/cmd:Components/cmd:Song/cmd:Authors/cmd:Author/cmd:Name/cmd:Original/text()
//cmd:CMD/cmd:Components/cmd:SongScan/cmd:Authors/cmd:Author/cmd:Name/cmd:Original/text()
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="PublicationYear">
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 ''
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 ''
<field name="Language">
<cmd:CMD xmlns:cmd="http://www.clarin.eu/cmd/">
<cmd:ArtworkProfile>
<cmd:CLARIN_PROFILE>clarin.eu:cr1:p_1345541703473</cmd:CLARIN_PROFILE>
<cmd:Title>Vivaio (Le chevalier au lion)</cmd:Title>
<cmd:Title>Perceval(le conte du Graal)</cmd:Title>
<cmd:Title>Chien</cmd:Title>
<cmd:ArtworkProfile>
<cmd:Component>
<cmd:CMD>
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

EUDAT B2FIND