B2FIND and Metadata Quality

3rd EUDAT Conference

25 September 2014

Heinrich Widmann

and B2FIND team

Outline

- B2FIND
  - the EUDAT Metadata Service
- Semantic Mapping of Metadata
- Quality of Metadata
- Summary and outlook

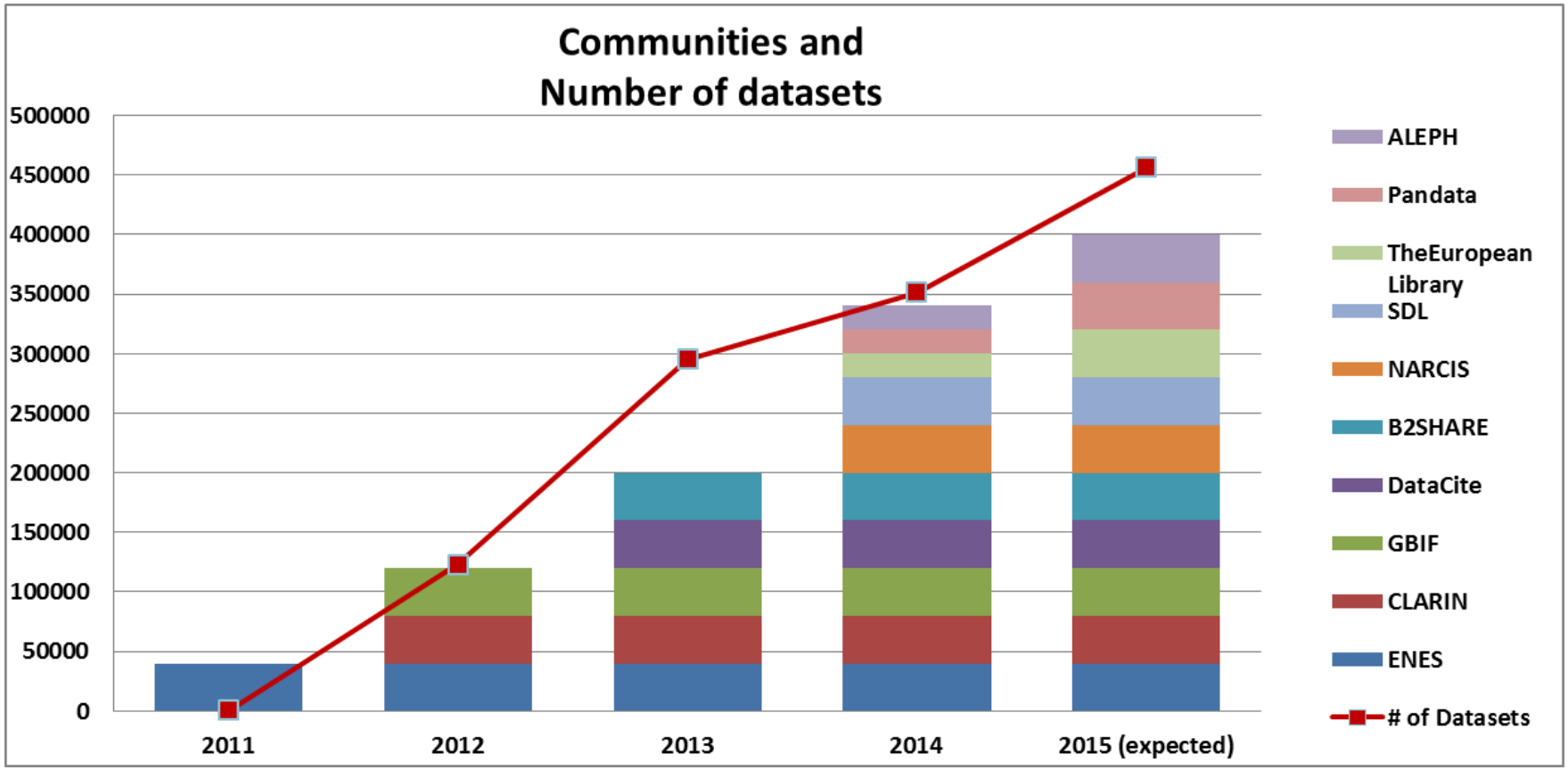
The EUDAT Metadata Service B2FIND

B2FIND consists of

- a comprehensive metadata catalogue of research data, that spans a large number of multi formatted datasets
  - harvested from various research communities
  - stored through the EUDAT service B2SHARE
  - covering a wide range of disciplines of high diversity
  - mapped to a common vocabulary
- an open search portal allowing researchers
  - to find easily collections of scientific resources
  - to access those resources through the given references in the metadata.

Status

- **10 communities** are integrated and in total > **350k datasets** are uploaded



- Portal and search functionality is installed based on open source software CKAN

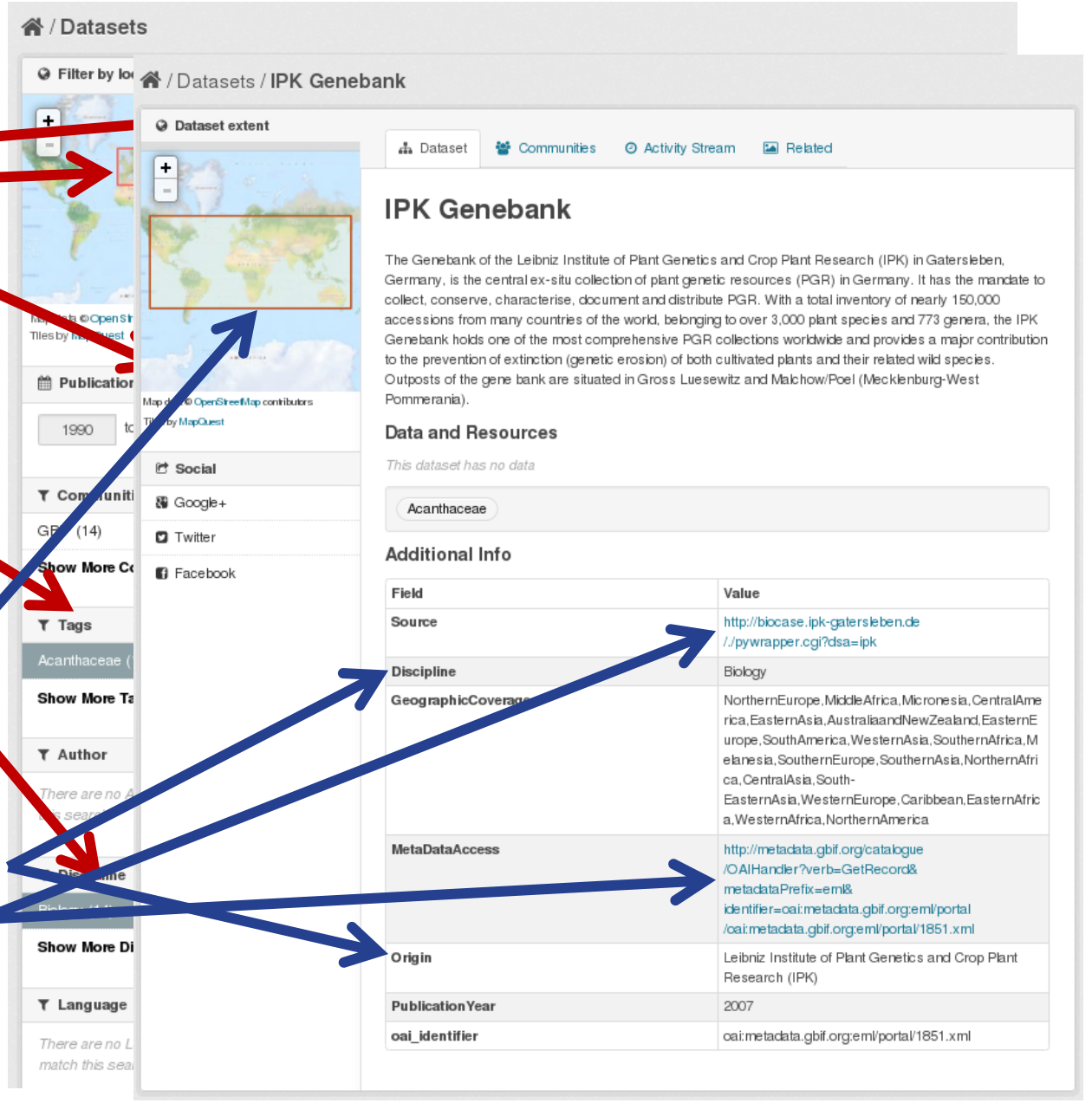
Functionalities

B2FIND provides 'faceted' search functionalities as

- Free text search,
- Geo spatial
- Time line search
- Search for facets as
  - 'Author'
  - 'Tags'
  - 'Discipline' etc.

Dataset view provides display of metadata :

- Spatial extent
- Table of field-value pairs
- Links to data resources

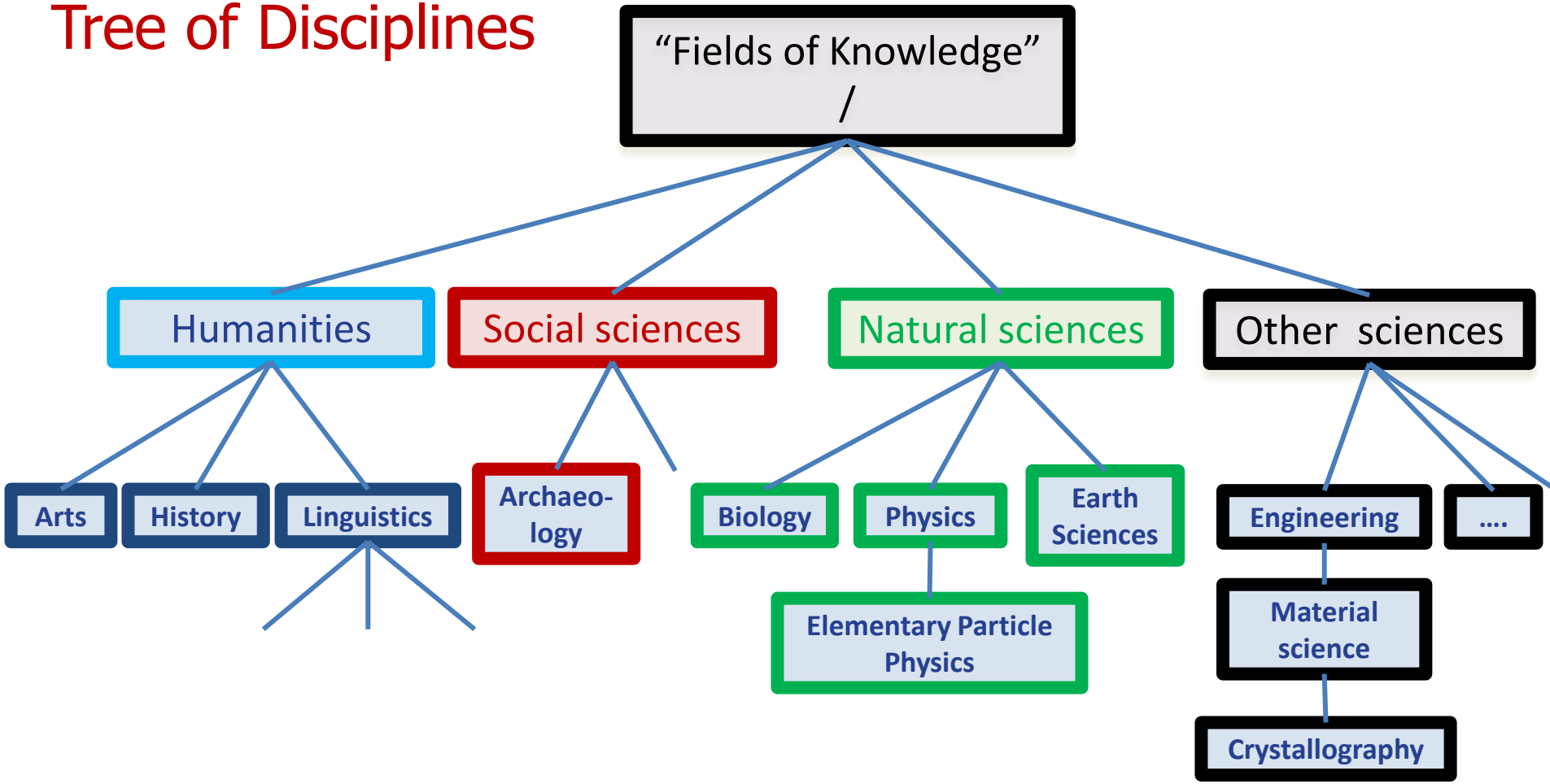


EUDAT

B2FIND Common Vocabulary (extract)

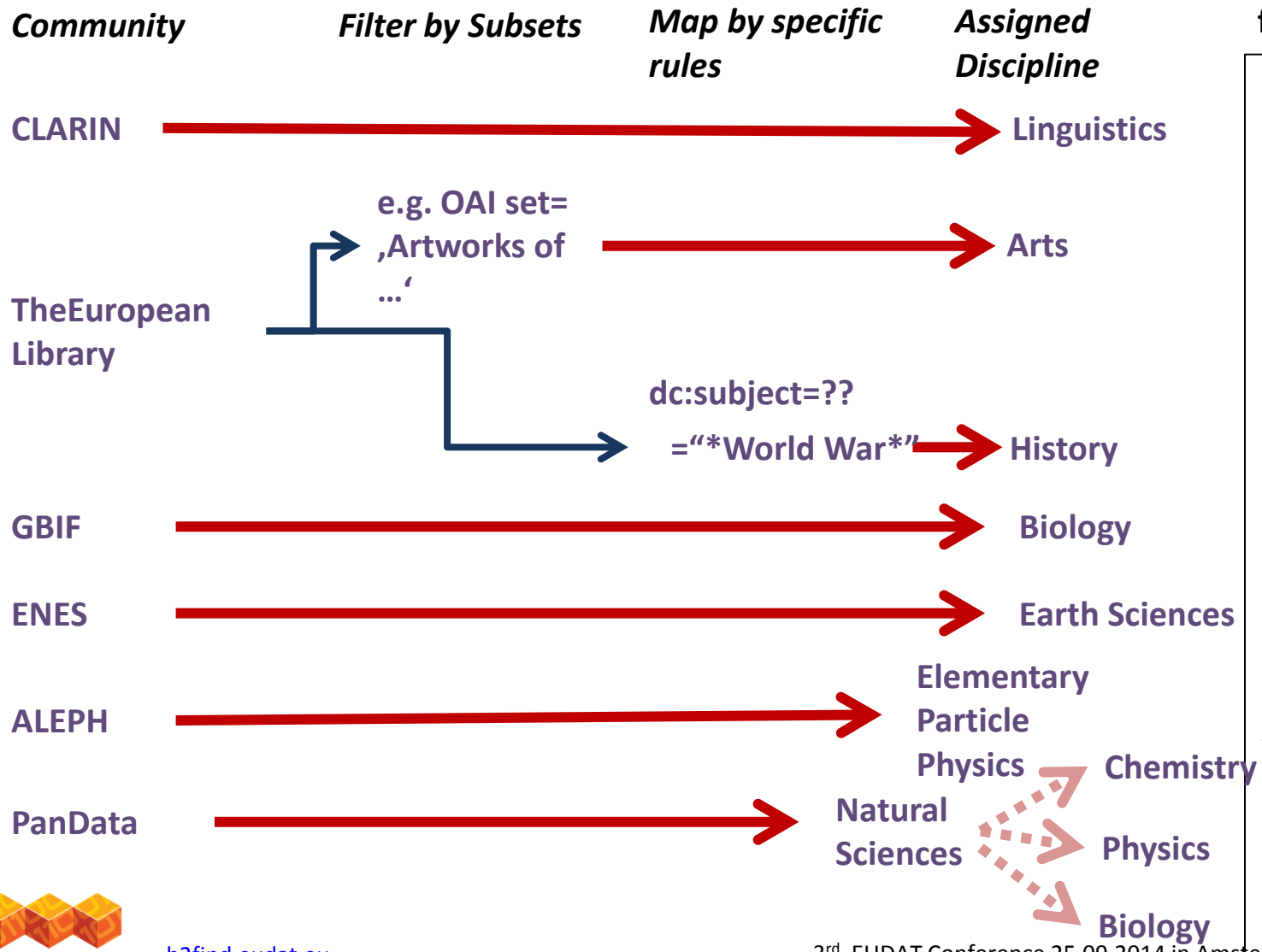
B2FIND Field name	Type and Display	Allowed values	Semantic definiton	Level of Obligation	Occu-rence
Title	Default field	Free text	A name or title a resource is known	M	1
Description	Default field	The format is open (but CKAN2.0 only supports plain text)	All additional information that does not fit in any of the other categories….	R	1
Author	B2FIND facet	Text field (list of citated names)	List of the main researchers involved in producing the publication, in priority order or …	R	1-n
Discipline	B2FIND facet	Text field (LOV)	Field of research ..	R	0-n
Source	Default field	Valid URL, DOI, PID, handle, … or any other URI	Identifiers of related resource(s).	M/R?	1-n
PublicationYear	B2FIND facet (from-to search)	YYYY	The year when the data was or will be made publicly	R	1
GeoSpatialCoverage	B2FIND facet (worldmap as search widget from which regions can be selected)	A spatial point or box specification, e.g. spatial={"type":"Polygon","coordinates":[[[minlat,minlon…]]}	The spatial limits of a place.	O	1
TemporalCoverage	B2FIND facet (from-to search widget in the portal)	Interval beetween two DateTimeStamps : [BeginDateTime , EndDateTime]	Relation to or Coverage of a specific interval in time.	O	1

Tree of Disciplines



taken from "List of Academic disciplines"
→ http://en.wikipedia.org/wiki/List_of_academic_disciplines_and_sub-disciplines and
„The Fields of Knowledge"
→ http://www.thingsmadethinkable.com/item/fields_of_knowledge.php?focus=natural_sciences

Mapping of the facet ‚Discipline'



B2FIND closed vocab for ‚Discipline'

1. Humanities
   - **1.1 History**
   - **1.2 Linguistics**
   - 1.3 Literature
   - **1.4 Arts**
     - 1.4.1 Performing arts
     - …
   - **1.5 Philosophy**
   - **1.6 Religion**
2. Social sciences
   - 2.1 Anthropology
   - **2.2 Archaeology**
   - ….
   - 2.7 Geography
3. Natural sciences
   - **3.1 Biology**
   - **3.2 Chemistry**
   - **3.3 Earth sciences**
   - **3.4 Physics**
   - …
4. Formal sciences
   - 4.1 Mathematics
   - 4.2 Computer sciences
5. Professions
   - **5.1 Agriculture**
   - ….
   - **5.6 Engineering**
     - 5.6.1 Chemical Eng.
   - 5.12 Library studies

Benefit

Metadata providers benefit from B2FIND by

- making their metadata visible and searchable
- allowing scientists to access their research data resources

- both in a wide European inter-disciplinary scope

Based on the example of the facet ‘Discipline’ the benefits are :

- the metadata are provided ‚re-arranged‘ and grouped by ‘field of research’
- Intra- and inter-disciplinary search is enabled
- an added value to the metadata is given by enabling better interoperability and reuse of data collections

MD Quality Criteria

- ✓ Visibility and easy searchability of metadata
  - Supported by B2FIND through easy-to-use discovery portal and facetted search functionality
- ✓ Completeness, coherence and comprehensiveness
  - B2FIND vocabulary covers the common intersection of community specific metadata properties
- ✓ Transparency and open data access (interoperability and reuse)
  - All metadata are free and open accessible in B2FIND
  - References to the resources are provided
- ✓ Timeliness and synchronisity
  - assured by frequent and incremental harvesting
- ? Correctness and accuracy of content
  - Open, no method for ‚content control' is available
  - Establish and follow minimal quality standards

Summary and outlook

- B2FIND provides a inter-disciplinary and broad-scoped metadata discovery service with powerful search functionalities
- Current developments
  - Time line search
    - Search for temporal coverage of datasets
  - Query-based Taxonomies
    - Hierarchical search in trees of ‘Disciplines’
- Plans for EUDAT II (towards an operational service)
  - Include metadata from B2SAFE and integrate further comm.’s
  - Establish content-related quality assurance
  - Direct end user feedback (Commenting functionality)
  - Use linked data (RDF’s) : Potential for semantic enrichment and integration with external reference materials (RDF vocabularies, ontologies, etc.)

Discussion and links

Open questions :

- How can B2FIND help metadata providers to improve quality of their metadata ?
- How can metadata quality be measured ?

B2FIND links :

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

Contact

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

Thank you for your attention !