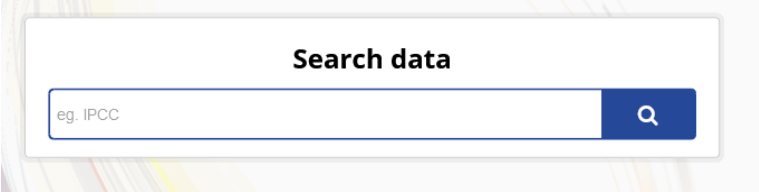
Metadata exchange issues - when standard meets reality

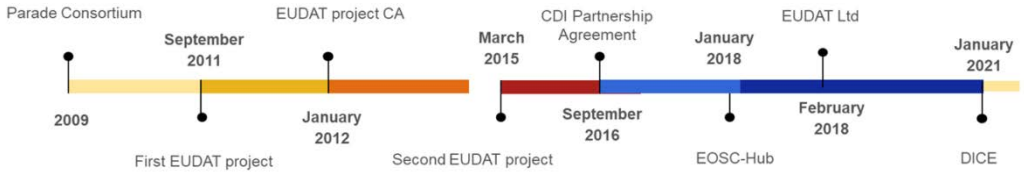
Lessons learned from B2FIND

B2FIND



http://b2find.eudat.eu/

- **already existing** simple and user-friendly **discovery portal** for **research data**
- evolving system, focus on **Community specific needs**, pracitcal approach
- up to now **46** Communities on productive machine, 16 on test-instance; **~ 1,5 mio records** searchable
- joint metadata catalogue for records from very **divergent research areas** or **Infrastructures**
- central indexing tool for **EOSC-hub,** discovery service in **EOSC**



B2FIND

Harvesting

technical issues

Search

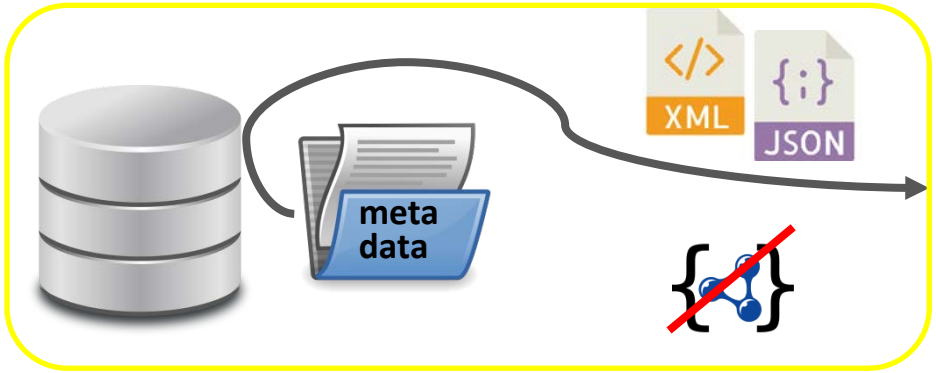
display issues

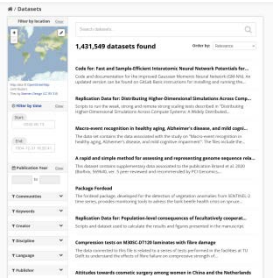
OpenAIRE

semantic / syntax issues

Mapping

technical issues Harvesting





supported

- OAI-PMH
  [Open Archive Initiative – Protocol for Metadata Harvesting]
- CSW
  [Catalogue Service for the Web / Open Geospatial Consortium]
- RestAPI

not (yet) supported

- SPaRQL
  [SPARQL Protocol and RDF Query Language]
  thus no RDF or JSON LD harvesting

Lessons Learned I

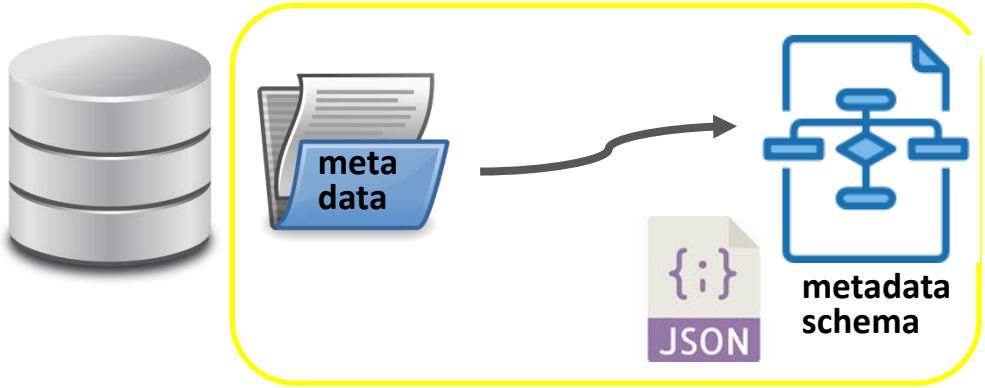
- **OAI-PMH**
  - ✓ was built for harvesting metadata
  - ✓ allows to create `OAI-Sets`, enables incremental harvesting with `fromDate`, multiple metadata prefixes are possible (default is DublinCore)
  - – not supported anymore *but* it is still stable and it works
  - – a bit tricky to set-up (Tomcat issues)
- **CSW**
  - ✓ specificly develop for and used in 'Geo'Communities
  - – made for (and restricted to) Iso19115, Iso19139 = INSPIRE
- **RestApi**
  - ✓ 'individual' solution (full control of metadata exposure for Data provider)
  - – 'individual' solution (configuration of API requests for each endpoint on B2FIND side)

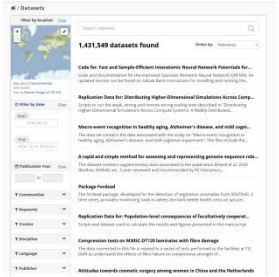
Lessons Learned II

- relational Database *vs.* Triplestore
  - Linked Data were made for the Semantic Web; mainly developed and used in Academia (University Libraries, some research Communities / Infrastructures)
  - however relational databases still broadly used (especially in data intensive research)
  - there is no 'one-fits-all' solution, need to combine different methods (harvesting triples enabled in beta version)
- Future: Graph DBs, Discovery Graphs

  (no triples but nodes -> edges)

syntax/semantic issues **Mapping**





supported metadata standards

generic	*thematic*	*specific*
Datacite	**Iso19139/**	**FF** [Danish
Duclin Core	**Iso19115**	Archeology]
OpenAire	[= **INSPIRE**]	(**CMDI** in old
EUDAT Core	**FGDC**	version)
(**MarcXML** in old version)		

syntax/semantic issues **Mapping**

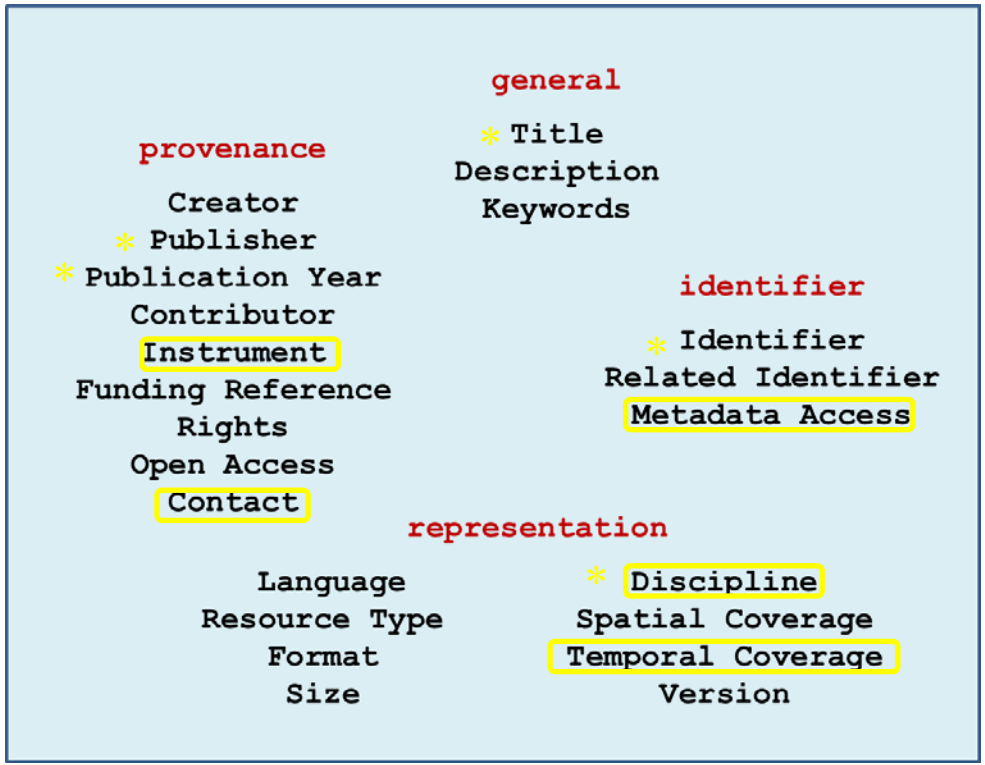
syntax = structure, relation and labeling of metadata elements

- [meta]data model
- format of metadata (XML, key-value-pairs, triples)
- ➢ generic vs. discipline-specific metadata schemas, where is the equilibrium? Spoiler: ongoing process…

semantic = terms within the metadata schema

- crosswalks
- including external references (closed vocabs, thesauri)
- ➢ how to transfer content?

B2FIND schema



- based on Datacite, practical approach
- developed to enable **discoverability** of research output – simple structure, *we take what we get!*
- evolved over time, low barrier approach
- enable **interdisciplinary** discoverability
- = EUDAT Core Metadata Schema

Generic Metadata Standards

EUDAT Core	Datacite	OpenAire
Community	[prefix]	Community
Identifier either DOI / PID / Source	only DOI	Identifier types DOI, Handle, PURL, URL, URN, ARK
Discipline	Subject	Subject
Instrument	-	-
Contact	within <contributorType = ContactPerson>	within <contributorType = ContactPerson>
Temporal Coverage	Date	Date
Spatial Coverage	GeoLocation	GeoLocation

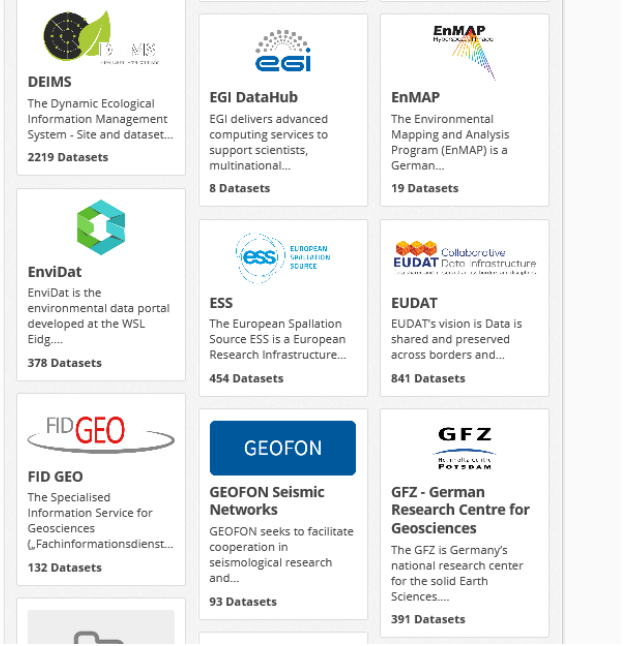
Examples

Community

OAI-PMH Sets for displaying multiple Communities: GFZ

```
class GeofonDatacite(Community):
    NAME = 'geofon'
    IDENTIFIER = NAME
    URL = 'http://doidb.wdc-terra.org/oaip/oai'
    SCHEMA = SchemaType.DataCite
    SERVICE_TYPE = ServiceType.OAI
    OAI_METADATA_PREFIX = 'oai_datacite'
    OAI_SET = 'DOIDB.SEISNET'
    PRODUCTIVE = True

    def update(self, doc):
        doc.discipline = self.discipline(doc, 'Seismology')
```



Set
setName Reference quality citations only.
setSpec REFQUALITY Identifiers Records

Set
setName DOI Database
setSpec DOIDB Identifiers Records

Set
setName DOI Database (reference quality citations only)
setSpec DOIDB.REFQUALITY Identifiers Records

Set
setName CRC1211DB CRC 1211 Database
setSpec DOIDB.CRC1211 Identifiers Records

Set
setName CRC1211DB CRC 1211 Database (reference quality citations only)
setSpec DOIDB.CRC1211.REFQUALITY Identifiers Records

Set
setName EnMAP
setSpec DOIDB.ENMAP Identifiers Records

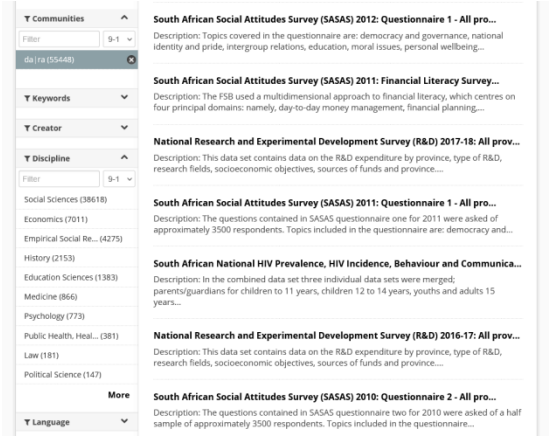
Set
setName EnMAP (reference quality citations only)
setSpec DOIDB.ENMAP.REFQUALITY Identifiers Records

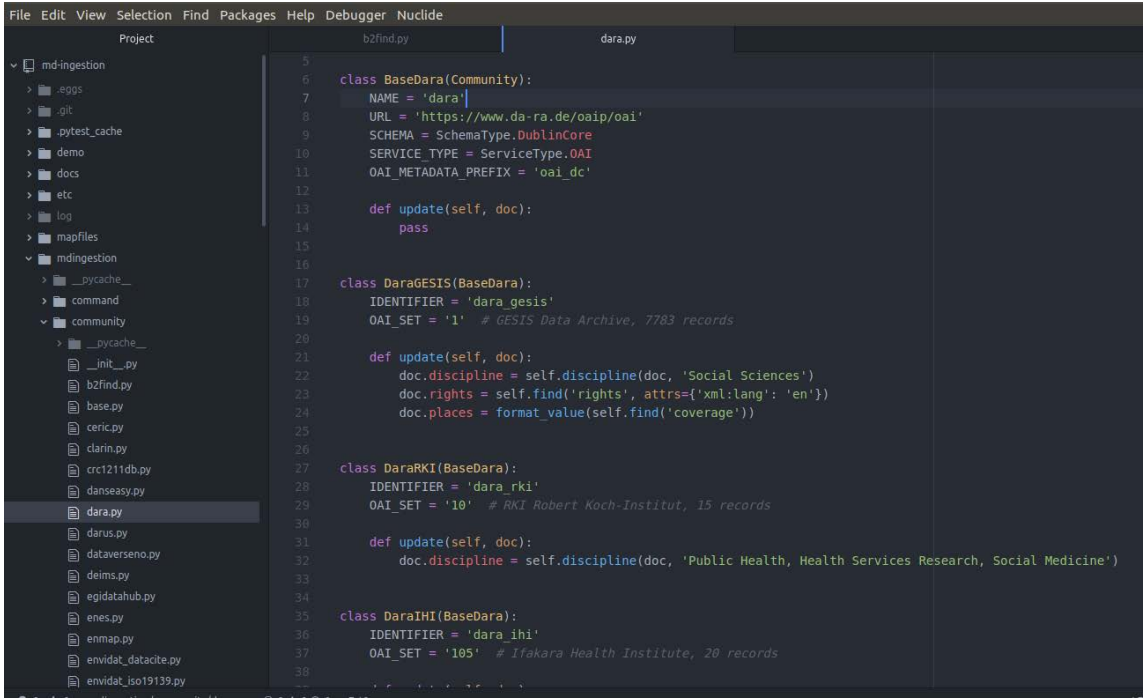
Set
setName FID GEO
setSpec DOIDB.FID Identifiers Records

Examples

Community

OAI-PMH Sets for specific mapping

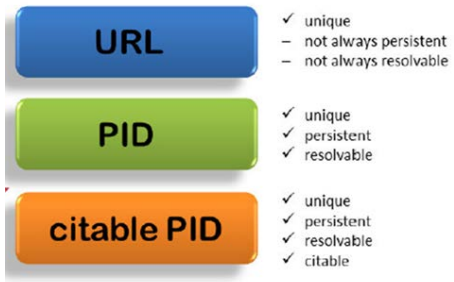




Examples

Identifier
internal ranking for representation of
`<identifiers>`

* DOI
* PID (Handle, ARK)
* Source (URL/URN)



Examples

Discipline

automated mapping from
`<subject>` to B2FIND
internal closed vocab of
`<disciplines>`

Example for INRAE

	A	B	C
1	**Subject**	**Discipline**	**Discipline Id**
2	Agricultural Sciences	Agriculture, Forestry, Horticulture	3.3.1
3	Arts and Humanities	Humanities	1
4	Astronomy and Astrophysics	Astrophysics and Astronomy	4.2.5
5	Business and Management	Business and Management	2.4.6
6	Chemistry	Chemistry	3,2
7	Computer and Information Science	Computer Sciencey	5.4.3
8	Earth and Environmental Sciences	Geosciences	4,4
9	Engineering	Engineering Sciencesy	5
10	Law	Jurisprudence	2,5
11	Mathematical Sciences	Mathematics	4,3
12	Medicine	Medicine	3,2
13	Health and Life Sciences	Life Sciences	3
14	Physics	Physics	4,2
15	Social Sciences	Social Sciences	2,3

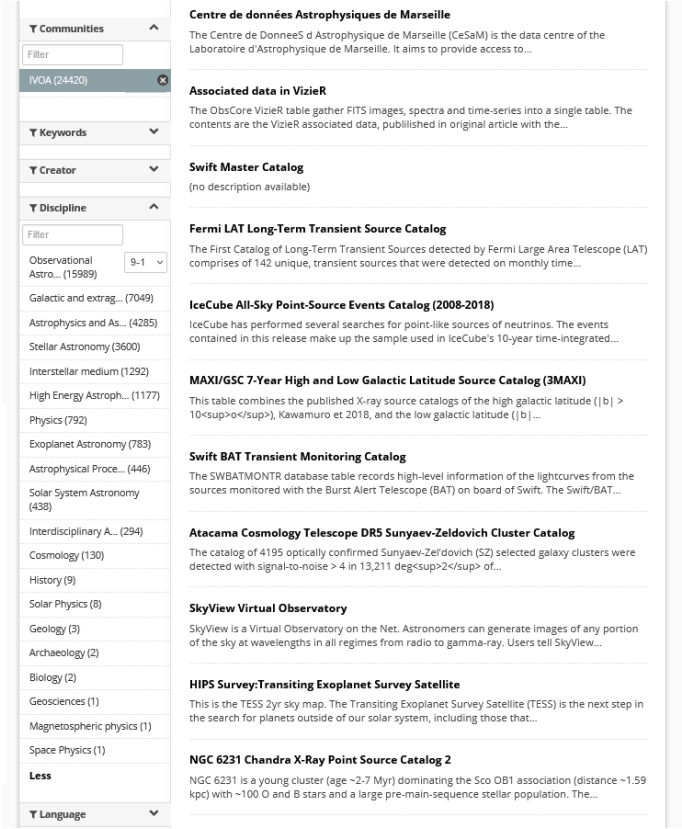
Examples

Discipline

integration of "Unified Astronomy Thesaurus" (https://astrothesaurus.org/) in B2FIND internal closed vocab of `<disciplines>`

Example for IVOA

```
"4.2.5.01#Astrophysics and Astronomy#Astrophysical Processes",
"4.2.5.02#Astrophysics and Astronomy#Cosmology",
"4.2.5.03#Astrophysics and Astronomy#Exoplanet Astronomy",
"4.2.5.04#Astrophysics and Astronomy#Galactic and extragalactic Astronomy",
"4.2.5.05#Astrophysics and Astronomy#High Energy Astrophysics",
"4.2.5.06#Astrophysics and Astronomy#Interdisciplinary Astronomy",
"4.2.5.07#Astrophysics and Astronomy#Interstellar medium",
"4.2.5.08#Astrophysics and Astronomy#Observational Astronomy",
"4.2.5.09#Astrophysics and Astronomy#Solar Physics",
"4.2.5.10#Astrophysics and Astronomy#Solar System Astronomy",
"4.2.5.11#Astrophysics and Astronomy#Stellar Astronomy",
```



Lessons Learned III

- Standards are great!
  As long as they are used…
- Standards are used when they there is a need (or benefit) for it – but not when they are enforced
- Standards are not static – they evolve over time, it´s a process
- there will (probably) never be a 'one-fits-all' standard for metadata exchange, focus on making existing ones interoperable -> requires (human) recources

- there will (probably) never be a 'one-fits-all' metadata schema
  a) the huge variety of data in extremly divergent research areas
  b) external restraints (national/european law)
- again: focus on interoperability -> requires (human) recources
- making everything machine readable is a splendid idea – but in practice interdisciplinary discoverability is not thinkable without human workforce

That´s it!

links

B2FIND portal
http://b2find.eudat.eu/
B2FIND Guidelines for data provider
http://b2find.eudat.eu/guidelines/introduction.html
B2FIND in GitHub
https://github.com/EUDAT-B2FIND/md-ingestion
B2FIND metadata schema concordance
http://b2find.eudat.eu/guidelines/mapping.html
B2FIND classification for disciplines
https://github.com/EUDAT-B2FIND/md-ingestion/blob/master/etc/b2find_disciplines.json

contact

EUDAT RT
for integration of new repositories
https://eudat.eu/contact-support-request
via email B2FIND
martens@dkrz.de
fluegel@dkrz.de