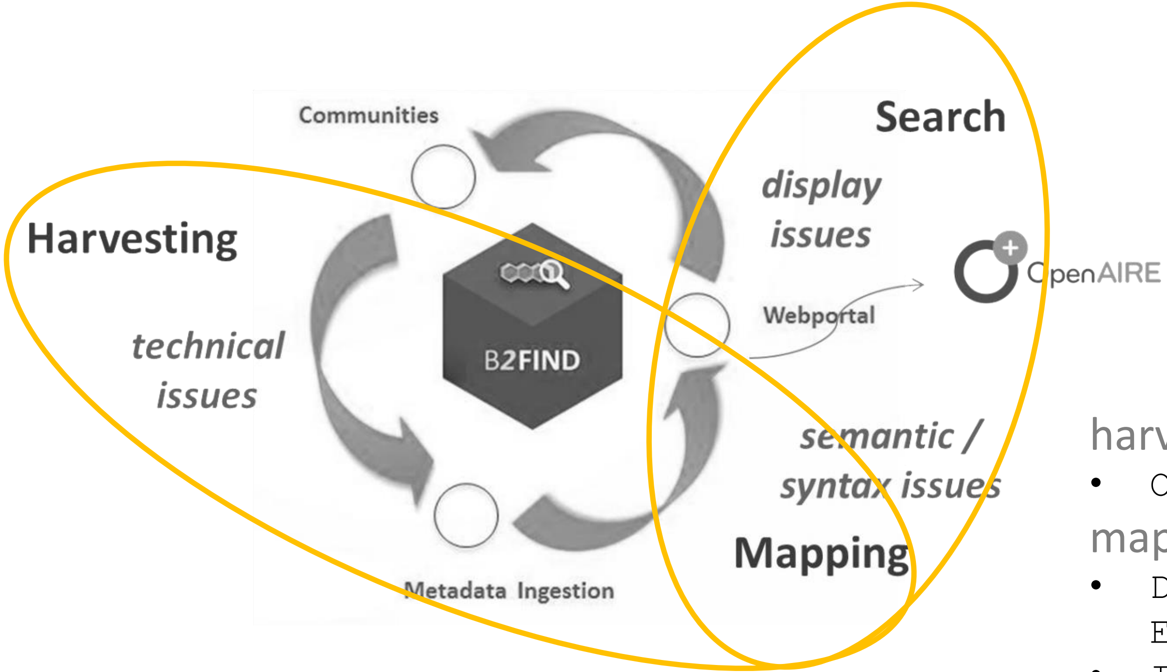
Metadata exchange issues – when standard meets reality Part II

Lessons learned from B2FIND

Claudia Martens, DKRZ

Metadata Issues: Part II



discovery

- search options
  (how to search)
- metadata ingestion
  (which information can be made searchable)

harvesting

- `OAI-PMH, CSW, RestAPI, [SPaRQL]`

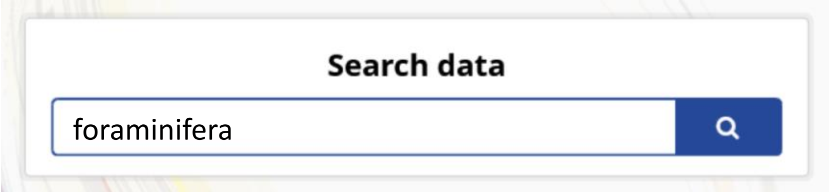
mapping

- `DublinCore, Datacite, OpenAire, EUDAT Core`
- `ISO 19115/19139 [INSPIRE], DDI 2.5`
- `community specific`

Search Options in B2FIND

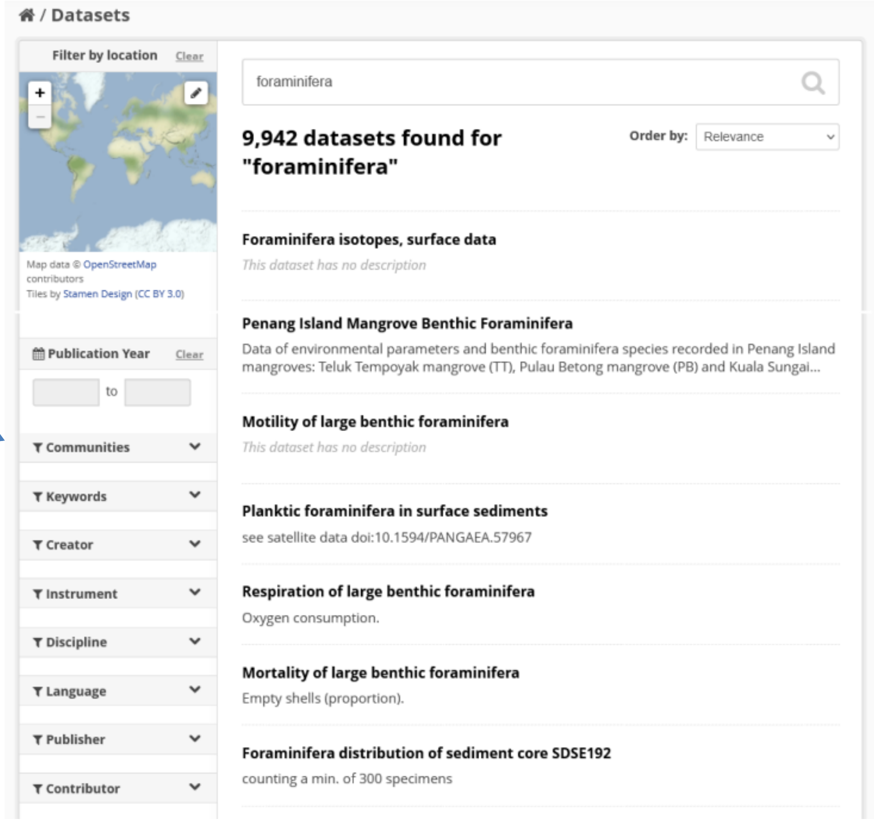
narrow down results using the facets

```
Spatial/TemporalCoverage,
PublicationYear,
Communities, Keywords,
Creator, Instrument,
Discipline, Language,
Publisher, Contributor,
ResourceType, OpenAccess
```



Search data

foraminifera



/ Datasets

Filter by location Clear

Map data © OpenStreetMap contributors
Tiles by Stamen Design (CC BY 3.0)

Publication Year Clear
to

Communities
Keywords
Creator
Instrument
Discipline
Language
Publisher
Contributor

foraminifera

9,942 datasets found for "foraminifera"

Order by: Relevance

Foraminifera isotopes, surface data
This dataset has no description

Penang Island Mangrove Benthic Foraminifera
Data of environmental parameters and benthic foraminifera species recorded in Penang Island mangroves: Teluk Tempoyak mangrove (TT), Pulau Betong mangrove (PB) and Kuala Sungai...

Motility of large benthic foraminifera
This dataset has no description

Planktic foraminifera in surface sediments
see satellite data doi:10.1594/PANGAEA.57967

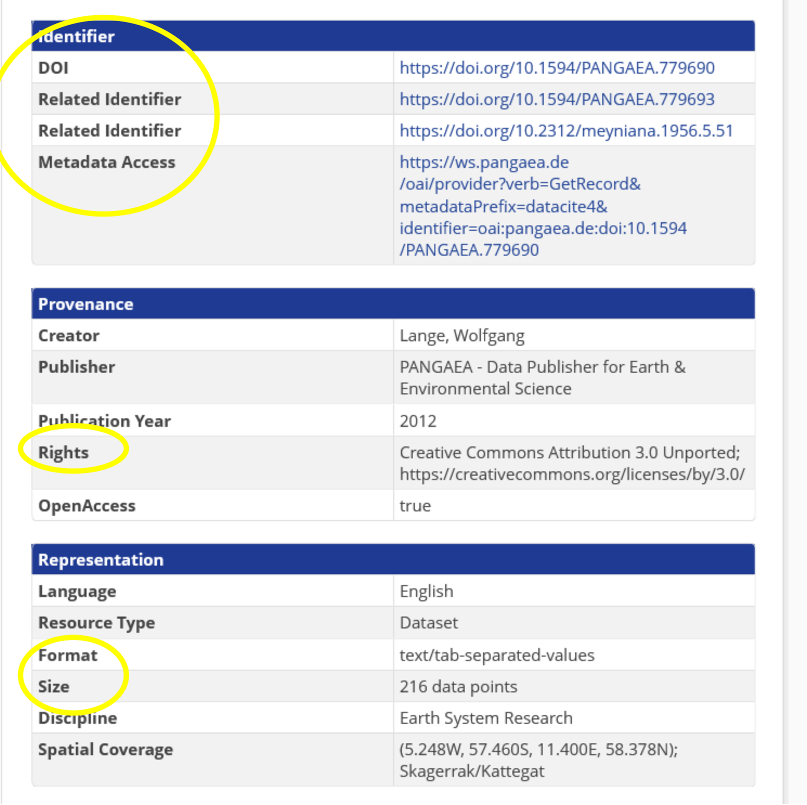
Respiration of large benthic foraminifera
Oxygen consumption.

Mortality of large benthic foraminifera
Empty shells (proportion).

Foraminifera distribution of sediment core SDSE192
counting a min. of 300 specimens

Search Options in B2FIND

single result page

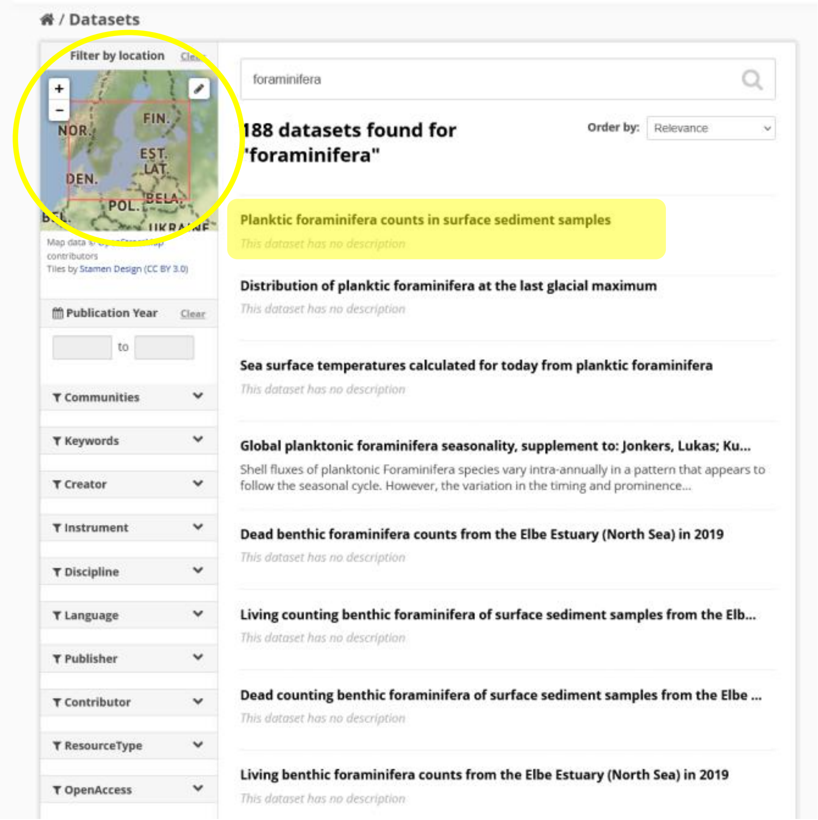


Identifier	
DOI	https://doi.org/10.1594/PANGAEA.779690
Related Identifier	https://doi.org/10.1594/PANGAEA.779693
Related Identifier	https://doi.org/10.2312/meyniana.1956.5.51
Metadata Access	https://ws.pangaea.de/oai/provider?verb=GetRecord&metadataPrefix=datacite4&identifier=oai:pangaea.de:doi:10.1594/PANGAEA.779690

Provenance	
Creator	Lange, Wolfgang
Publisher	PANGAEA - Data Publisher for Earth & Environmental Science
Publication Year	2012
Rights	Creative Commons Attribution 3.0 Unported; https://creativecommons.org/licenses/by/3.0/
OpenAccess	true

Representation	
Language	English
Resource Type	Dataset
Format	text/tab-separated-values
Size	216 data points
Discipline	Earth System Research
Spatial Coverage	(5.248W, 57.460S, 11.400E, 58.378N); Skagerrak/Kattegat

search result page



foraminifera

188 datasets found for "foraminifera"

Planktic foraminifera counts in surface sediment samples
This dataset has no description

Distribution of planktic foraminifera at the last glacial maximum
This dataset has no description

Sea surface temperatures calculated for today from planktic foraminifera
This dataset has no description

Global planktonic foraminifera seasonality, supplement to: Jonkers, Lukas; Ku...
Shell fluxes of planktonic Foraminifera species vary intra-annually in a pattern that appears to follow the seasonal cycle. However, the variation in the timing and prominence...

Dead benthic foraminifera counts from the Elbe Estuary (North Sea) in 2019
This dataset has no description

Living counting benthic foraminifera of surface sediment samples from the Elb...
This dataset has no description

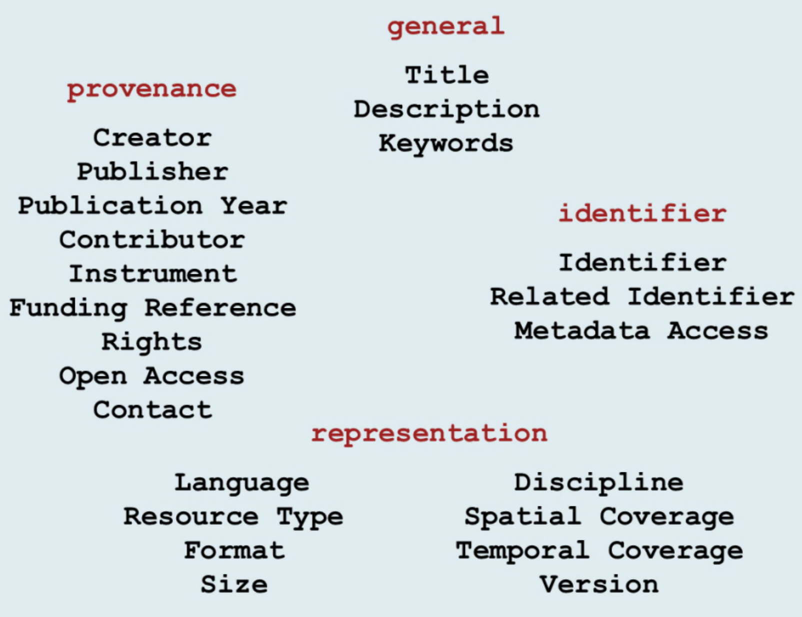
Dead counting benthic foraminifera of surface sediment samples from the Elbe ...
This dataset has no description

Living benthic foraminifera counts from the Elbe Estuary (North Sea) in 2019
This dataset has no description

B2FIND schema

A search portal may display only those information they get → schema

EUDAT Core Metadata Schema



general

Title
Description
Keywords

provenance

Creator
Publisher
Publication Year
Contributor
Instrument
Funding Reference
Rights
Open Access
Contact

identifier

Identifier
Related Identifier
Metadata Access

representation

Language
Resource Type
Format
Size

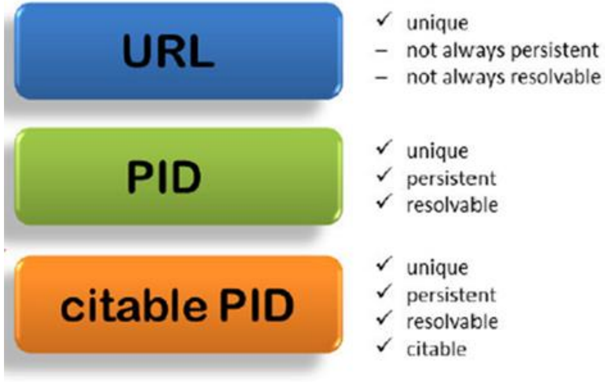
Discipline
Spatial Coverage
Temporal Coverage
Version

issues

- Identifier
- Discipline
- OpenAccess
- Temporal Coverage [demo]

Identifier

- persistent identification of (research) data is crucial
- search portals that aggregate metadata need something to link to -> Landing Page
- ways of using <identifier> changed over the years -> more and more DOIs…
- B2FIND was developed with a 'low barrier' approach: we take what we get



URL
✓ unique
– not always persistent
– not always resolvable

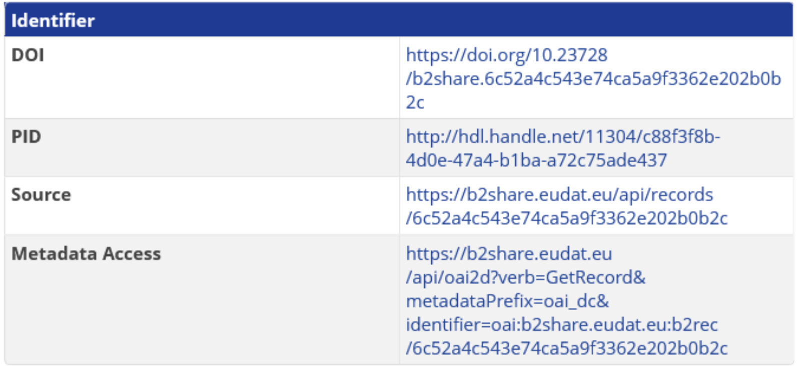
PID
✓ unique
✓ persistent
✓ resolvable

citable PID
✓ unique
✓ persistent
✓ resolvable
✓ citable

Source
(any URL/URN)

PID
(e.g. Handle)

DOI



Identifier	
DOI	https://doi.org/10.23728/b2share.6c52a4c543e74ca5a9f3362e202b0b2c
PID	http://hdl.handle.net/11304/c88f3f8b-4d0e-47a4-b1ba-a72c75ade437
Source	https://b2share.eudat.eu/api/records/6c52a4c543e74ca5a9f3362e202b0b2c
Metadata Access	https://b2share.eudat.eu/api/oai2d?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:b2share.eudat.eu:b2rec/6c52a4c543e74ca5a9f3362e202b0b2c

Identifier

	Datacite	OpenAire	EUDAT Core
`<identifier>`	DOI	ARK, DOI, Handle, PURL, URN, URL	DOI, Handle, PURL, URN, URL (ARK, arXiv, bibcode, EAN13, EISSN, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PMID, UPC, w3id)
`<alternate Identifier>`	Free text (but type mandatory)	Free text (but type mandatory)	- (multiple <identifier>)
`<related Identifier>`	ARK, arXiv, bibcode, DOI, EAN13, EISSN, Handle, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PMID, PURL, UPC, URL, URN, w3id	ARK, arXiv, bibcode, DOI, EAN13, EISSN, Handle, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PMID, PURL, UPC, URL, URN, w3id	Free text

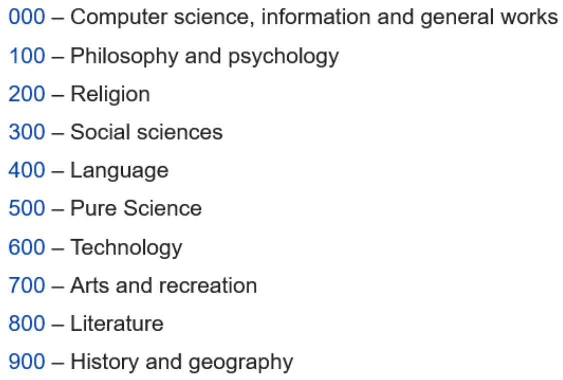
Discipline

- starting point: a vocabulary that defines different 'Scientific Disciplines' or 'Research Areas' → Wikipedia list for "Branches of Science"

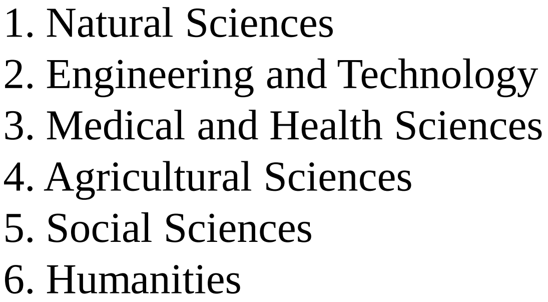
- evaluation of existing classifications: not really usable
  - either because too formal (Library Classification Systems, e.g. DDC or UDC)
  - or because too political (e.g. FOS by OECD)

- various vocabs for different purposes exist, but no "general classification" for all

- joint projects to develop a sustainable classification failed



000 – Computer science, information and general works
100 – Philosophy and psychology
200 – Religion
300 – Social sciences
400 – Language
500 – Pure Science
600 – Technology
700 – Arts and recreation
800 – Literature
900 – History and geography

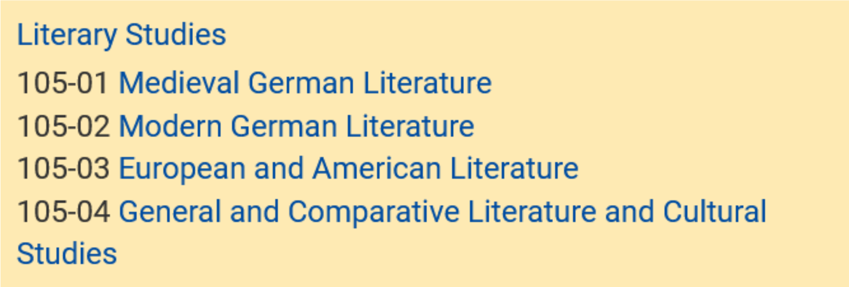


1. Natural Sciences
2. Engineering and Technology
3. Medical and Health Sciences
4. Agricultural Sciences
5. Social Sciences
6. Humanities

Discipline

Current status

- using a formal classification from the German National Funding Agency, modified to our needs
- internal mapping from `<subject>` or added fix in mapfile for Communities



Literary Studies

105-01 Medieval German Literature

105-02 Modern German Literature

105-03 European and American Literature

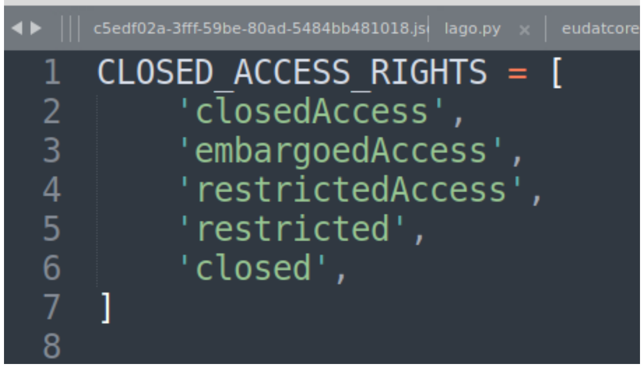
105-04 General and Comparative Literature and Cultural Studies

Future work

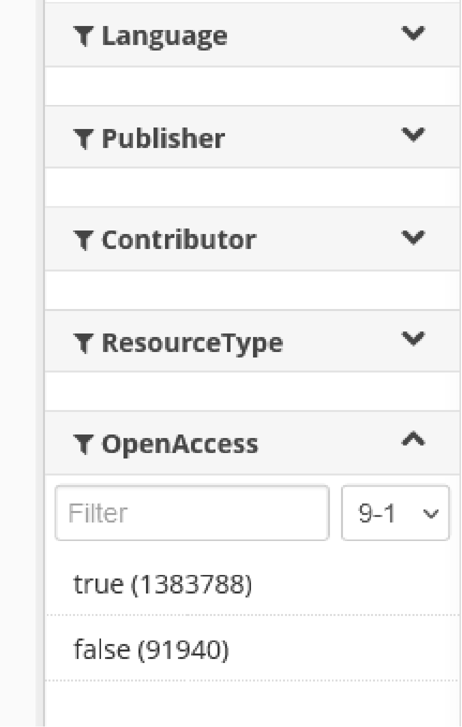
- either finding a suitable existing classification to be used in B2FIND (and EUDAT)
- or developing something that is reusable and sustainable
  - different concepts → classification vs. thesaurus vs. ontology vs. closed vocabulary
  - scientific methods allow to 'examine' anything → e.g. the molecular structure of proteins on an old painting or within a mammal, how to classify the result?
  - assumption: difficult to define 'borders' for and within research areas

OpenAccess

- good example for the development of user needs
  - growing demand to know whether a dataset is openly accessible or not
- problem: definition of "Open"...
- internal mapping: boolean operator, default is `'True'` if not specified differently within `<rights>`



```
CLOSED_ACCESS_RIGHTS = [
        'closedAccess',
        'embargoedAccess',
        'restrictedAccess',
        'restricted',
        'closed',
]
```



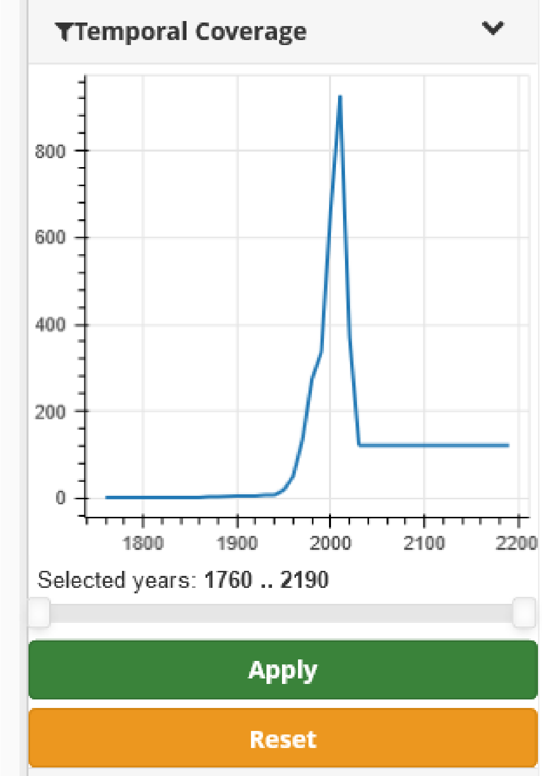
true (1383788)

false (91940)

Temporal Coverage

- timeline search on productive B2FIND not working properly
- new testmachine as foundation for next B2FIND version
- demo here

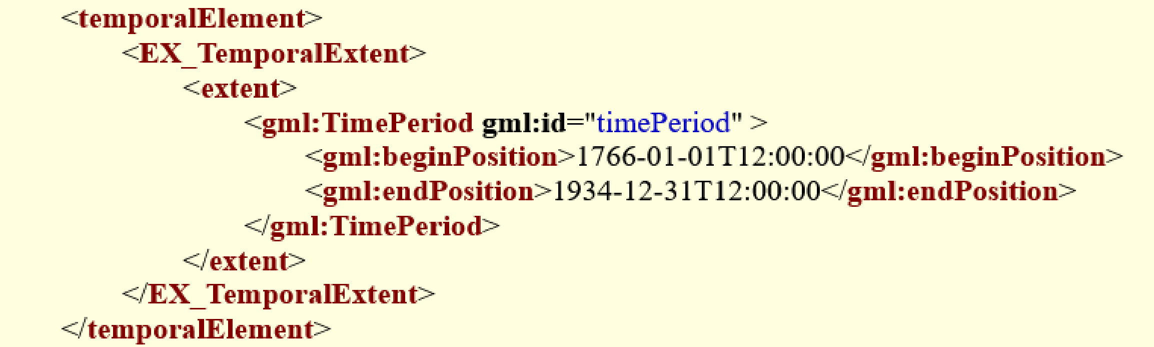
http://eudat9.cloud.dkrz.de/dataset



Selected years: 1760 .. 2190

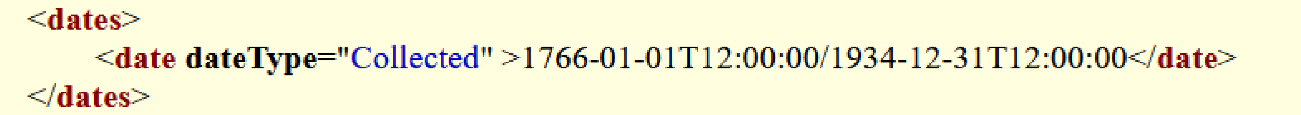
ISO 19115/19139 and Datacite

ISO 10115/19139



```xml
<temporalElement>
    <EX_TemporalExtent>
        <extent>
            <gml:TimePeriod gml:id="timePeriod" >
                <gml:beginPosition>1766-01-01T12:00:00</gml:beginPosition>
                <gml:endPosition>1934-12-31T12:00:00</gml:endPosition>
            </gml:TimePeriod>
        </extent>
    </EX_TemporalExtent>
</temporalElement>
```

Datacite 4



```xml
<dates>
    <date dateType="Collected" >1766-01-01T12:00:00/1934-12-31T12:00:00</date>
</dates>
```

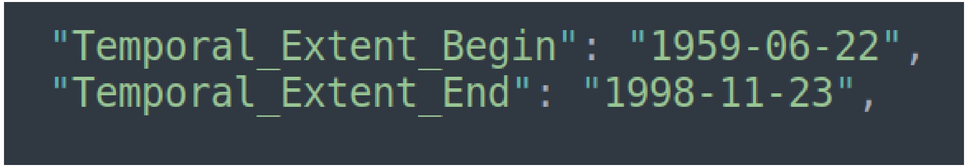
Example: PANGAEA

- exposing metadata with various standards
- transferring temporal information with ISO 19139 and Datacite

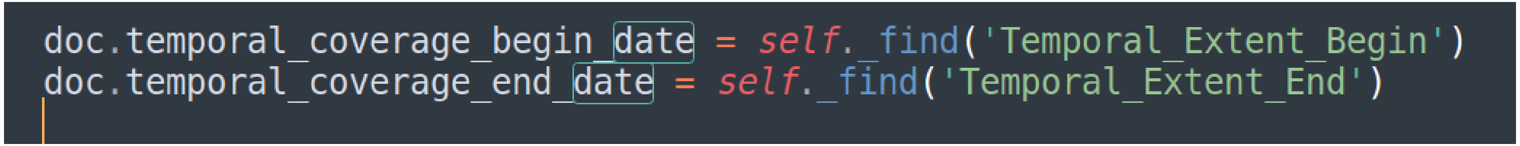
Community specific

Example: Blue-Cloud

- harvesting via JSON API
- specific mapping for given metadata elements
- information for temporal coverage within `'Temporal_Extent'`



```
"Temporal_Extent_Begin": "1959-06-22",
"Temporal_Extent_End": "1998-11-23",
```



```
doc.temporal_coverage_begin_date = self._find('Temporal_Extent_Begin')
doc.temporal_coverage_end_date = self._find('Temporal_Extent_End')
```

DDI 2.5

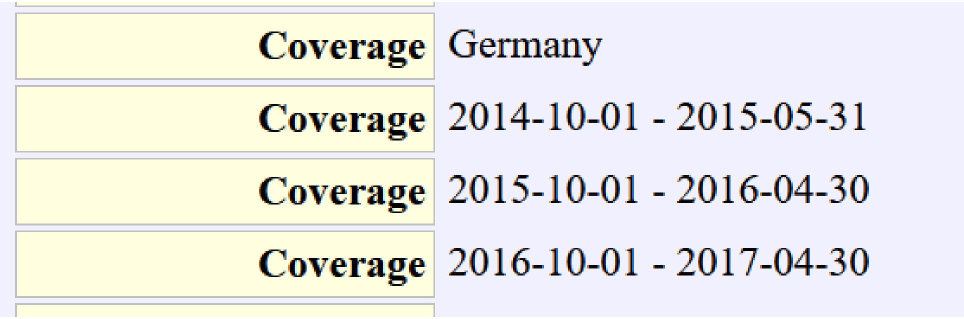
Example: CESSDA

- in principle DDI2.5 allows different ways to deal with information related to time
- for temporalCoverage fits `'timePrd'` with start / end date
- only the test system of CESSDA Data Catalogue, no practical example

DublinCore

Example: da|ra

- DublinCore has only one element `<coverage>` which is not specified
  - could be for spatial information (geo coordinates but also plain text) or for temporal information
- leads to mixed-up information



Coverage	Germany
Coverage	2014-10-01 - 2015-05-31
Coverage	2015-10-01 - 2016-04-30
Coverage	2016-10-01 - 2017-04-30

http://eudat9.cloud.dkrz.de/dataset/a0acd1e2-c784-5696-9567-71a7dba17cd3

Lessons Learned I

- **search interfaces change as user needs change**
  - display as much information as possible vs. display only important information – what is important? Where is the equilibrium? Ongoing discussion…
  - … which facets are useful?

  → different needs for different research areas (e.g. georeferenced data vs. medical data)

  → example: FundingReference

  → example: IVOA (=Astrophysics) have MOC (coordinates on sky) – how to display?

- **metadata schemas and standards evolve over time**
  - enabling discovery means: adaption, adoption, continous dvelopment
  - B2FIND enables the integration of many metadata standards as well as specific mappings for Community internal metadata schemas
    - ✓ good: more information = better precision and recall
    - ✓ bad: no way to make this happen without human workforce

  → but even standards may be 'misused'

  → community specific mapping requires effort

Lessons Learned II

- **Let machines do the work for us!**
  - nice idea, but not feasible now; even not (yet) google dataset search - all for SEO?
  - in reality most data are not exposed at all but 'hidden' in community specific repositories
  - those that are exposed use varying metadata schemas, even those who use 'standards' do this differently
  - making data FAIR is a good way – but it just started and it must be *done* by someone

- **B2FIND**
  - an entry point to search for research data
    → we can´t (and don´t intend to) replace existing search portals
  - given the flexible metadata ingestion, B2FIND is not only a metadata aggregator but also a metadata curator
    → make b2f specific mappings reusable by others!
  - consulting/advice is extremly important – communication is key!

That´s it

links

B2FIND search 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 classification for disciplines
https://github.com/EUDAT-B2FIND/md-ingestion/blob/master/etc/b2find_disciplines.json
DFG Classification
https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp
EUDAT Core Metadata Schema
https://gitlab.eudat.eu/eudat-metadata/eudat-core-schema/-/blob/master/eudat-core.xsd

contact

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