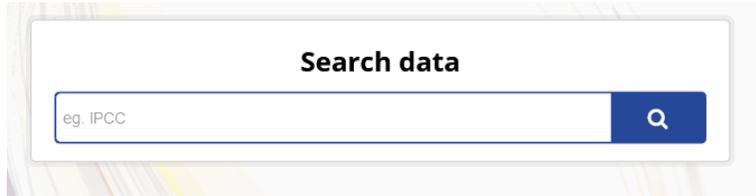


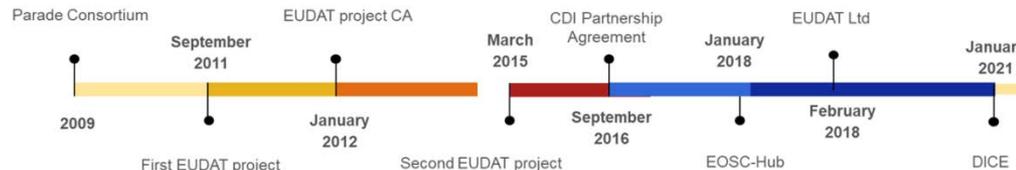
# Metadata exchange issues - when standard meets reality

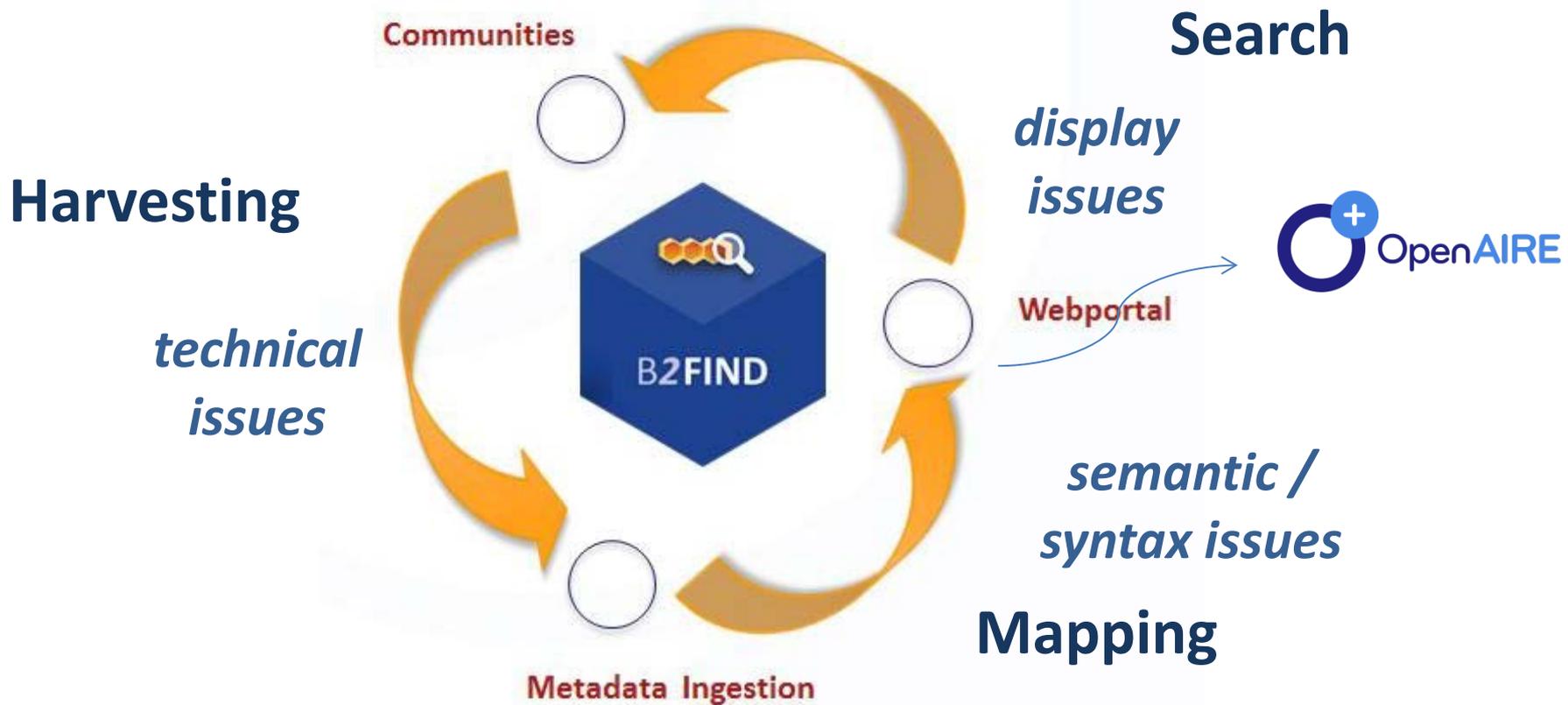
Lessons learned from B2FIND

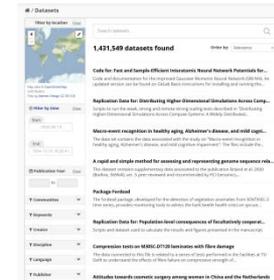
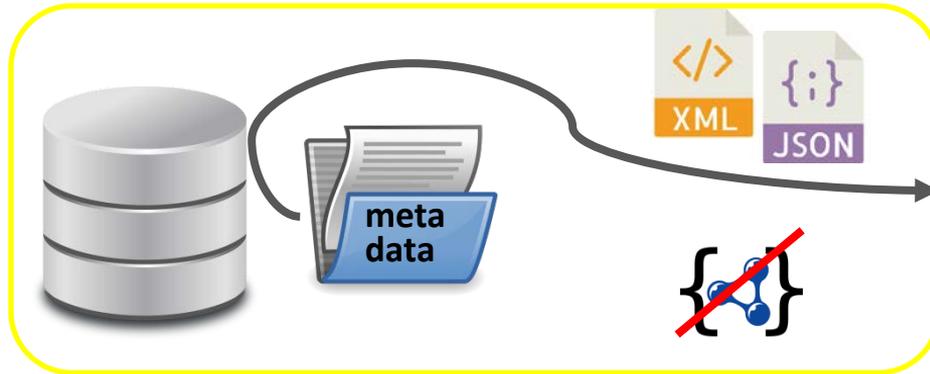


<http://b2find.eudat.eu/>

- ❖ **already existing** simple and user-friendly **discovery portal** for **research data**
- ❖ evolving system, focus on **Community specific needs**, practical 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**







## 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

- **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)

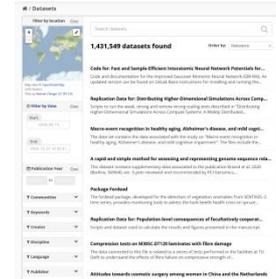
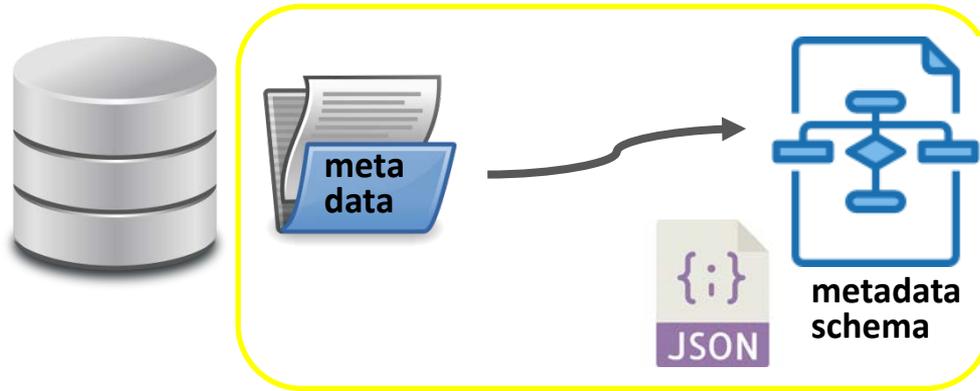
- **CSW**

- ✓ specifically 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)

- 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)



## supported metadata standards

*generic*

**Datacite**

**Duclin Core**

**OpenAire**

**EUDAT Core**

**(MarcXML in old version)**

*thematic*

**Iso19139/**

**Iso19115**

**[= INSPIRE]**

**FGDC**

*specific*

**FF [Danish**

**Archeology]**

**(CMDI in old**

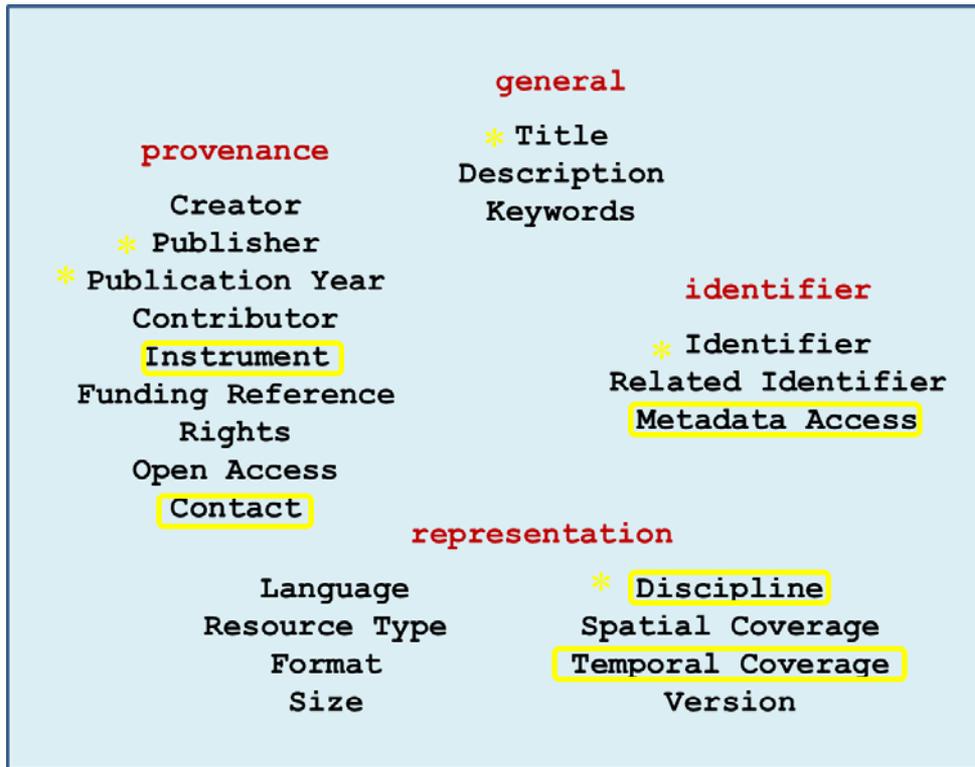
**version)**

**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?



- 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

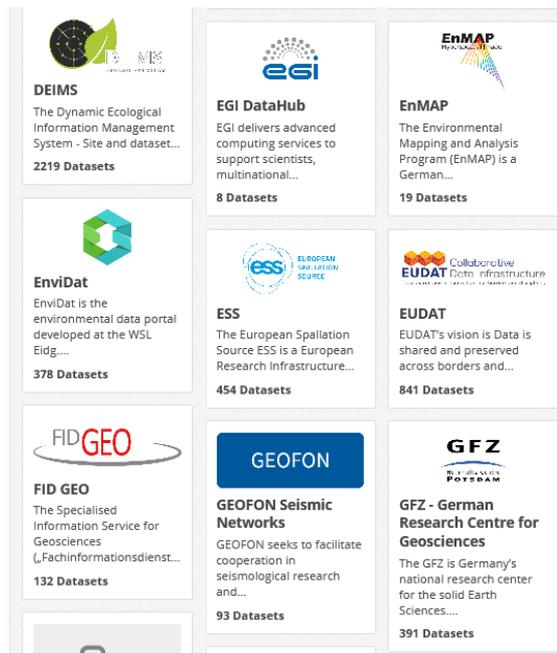
EUDAT Core	Datacite	OpenAire
Community	[prefix]	Community
Identifier either DOI / PID / Source	only DOI	Identifier types DOI, Handle, PURL, URL, URN, ARK
Discipline	Subject	Subject
Instrument	-	-
Contact	within <contributorType = ContactPerson>	within <contributorType = ContactPerson>
Temporal Coverage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

## Community

### OAI-PMH Sets for displaying multiple Communities: GFZ

```
class GeofonDatacite(Community):
    NAME = 'geofon'
    IDENTIFIER = NAME
    URL = 'http://doidb.wdc-terra.org/oaip/oi'
    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')
```



 <p><b>DEIMS</b> The Dynamic Ecological Information Management System - Site and dataset... 2219 Datasets</p>	 <p><b>EGI DataHub</b> EGI delivers advanced computing services to support scientists, multinational... 8 Datasets</p>	 <p><b>EnMAP</b> The Environmental Mapping and Analysis Program (EnMAP) is a German... 19 Datasets</p>
 <p><b>EnviDat</b> EnviDat is the environmental data portal developed at the WSL Eidg... 378 Datasets</p>	 <p><b>ESS</b> The European Spallation Source ESS is a European Research infrastructure... 454 Datasets</p>	 <p><b>EUDAT</b> EUDAT's vision is Data is shared and preserved across borders and... 841 Datasets</p>
 <p><b>FID GEO</b> The Specialised Information Service for Geosciences (Fachinformationsdienst... 132 Datasets</p>	 <p><b>GEOFON Seismic Networks</b> GEOFON seeks to facilitate cooperation in seismological research and... 93 Datasets</p>	 <p><b>GFZ - German Research Centre for Geosciences</b> The GFZ is Germany's national research center for the solid Earth Sciences... 391 Datasets</p>

#### 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

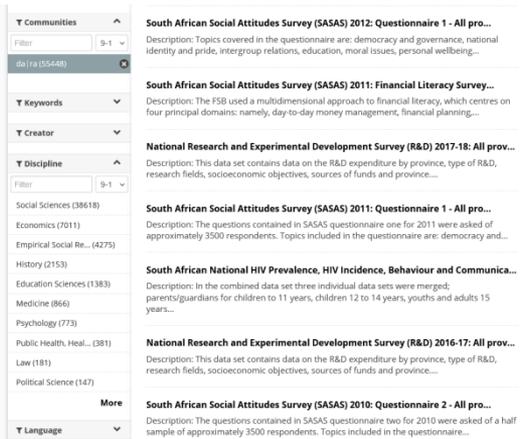
#### Set

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

#### Set

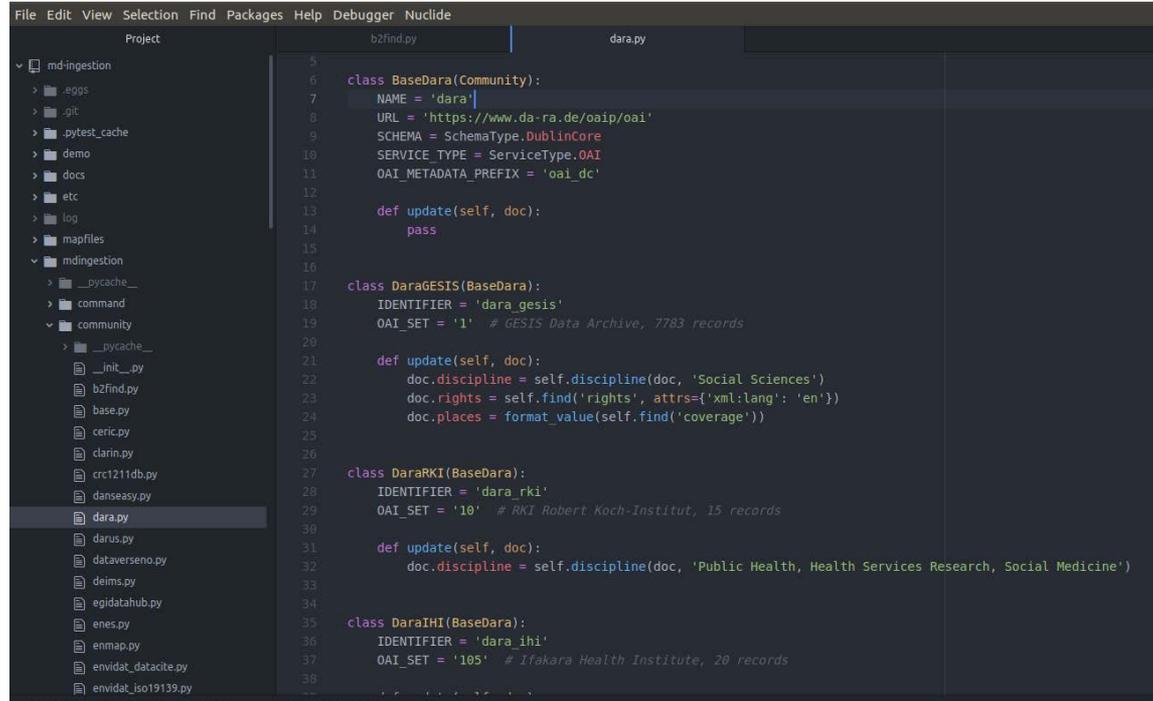
setName FID GEO  
setSpec DOIDB.FID Identifiers Records

## Community OAI-PMH Sets for specific mapping



The screenshot shows a search results page for 'Communities'. The left sidebar has filters for Communities (da/ra (55448)), Keywords, Creator, and Discipline (Social Sciences (38618), Economics (7011), Empirical Social Res... (4275), History (2153), Education Sciences (1383), Medicine (866), Psychology (773), Public Health, Heal... (381), Law (181), Political Science (147)). The main content area lists several survey datasets with their descriptions:

- South African Social Attitudes Survey (SASAS) 2012: Questionnaire 1 - All pro...**  
Description: Topics covered in the questionnaire are: democracy and governance, national identity and pride, intergroup relations, education, moral issues, personal wellbeing...
- South African Social Attitudes Survey (SASAS) 2011: Financial Literacy Survey...**  
Description: The FSLB used a multidimensional approach to financial literacy, which centres on four principal domains: namely, day-to-day money management, financial planning...
- National Research and Experimental Development Survey (R&D) 2017-18: All prov...**  
Description: This data set contains data on the R&D expenditure by province, type of R&D, research fields, socioeconomic objectives, sources of funds and province...
- South African Social Attitudes Survey (SASAS) 2011: Questionnaire 1 - All pro...**  
Description: The questions contained in SASAS questionnaire one for 2011 were asked of approximately 3500 respondents. Topics included in the questionnaire are: democracy and...
- South African National HIV Prevalence, HIV Incidence, Behaviour and Communica...**  
Description: In the combined data set three individual data sets were merged: parents/guardians for children to 11 years, children 12 to 14 years, youths and adults 15 years...
- National Research and Experimental Development Survey (R&D) 2016-17: All prov...**  
Description: This data set contains data on the R&D expenditure by province, type of R&D, research fields, socioeconomic objectives, sources of funds and province...
- South African Social Attitudes Survey (SASAS) 2010: Questionnaire 2 - All pro...**  
Description: The questions contained in SASAS questionnaire two for 2010 were asked of a half sample of approximately 3500 respondents. Topics included in the questionnaire...



```

File Edit View Selection Find Packages Help Debugger Nuclide
Project b2find.py dara.py
└─ md-ingestion
  └─ .eggs
  └─ .git
  └─ .pytest_cache
  └─ demo
  └─ docs
  └─ etc
  └─ log
  └─ mapfiles
  └─ mdingestion
    └─ __pycache__
    └─ command
    └─ community
      └─ __pycache__
        └─ __init__.py
        └─ b2find.py
        └─ base.py
        └─ ceric.py
        └─ clarin.py
        └─ crc1211db.py
        └─ danseasy.py
        └─ dara.py
        └─ darus.py
        └─ dataverseno.py
        └─ deims.py
        └─ egidatagub.py
        └─ enes.py
        └─ enmap.py
        └─ envidat_datacite.py
        └─ envidat_iso19139.py
  5
  6 class BaseDara(Community):
  7     NAME = 'dara'
  8     URL = 'https://www.da-ra.de/oaip/oaip'
  9     SCHEMA = SchemaType.DublinCore
 10     SERVICE_TYPE = ServiceType.OAI
 11     OAI_METADATA_PREFIX = 'oai_dc'
 12
 13     def update(self, doc):
 14         pass
 15
 16
 17 class DaraGESIS(BaseDara):
 18     IDENTIFIER = 'dara_ghesis'
 19     OAI_SET = '1' # GESIS Data Archive, 7783 records
 20
 21     def update(self, doc):
 22         doc.discipline = self.discipline(doc, 'Social Sciences')
 23         doc.rights = self.find('rights', attrs={'xml:lang': 'en'})
 24         doc.places = format_value(self.find('coverage'))
 25
 26
 27 class DaraRKI(BaseDara):
 28     IDENTIFIER = 'dara_rki'
 29     OAI_SET = '10' # RKI Robert Koch-Institut, 15 records
 30
 31     def update(self, doc):
 32         doc.discipline = self.discipline(doc, 'Public Health, Health Services Research, Social Medicine')
 33
 34
 35 class DaraIHI(BaseDara):
 36     IDENTIFIER = 'dara_ihi'
 37     OAI_SET = '105' # Ifakara Health Institute, 20 records
 38
  
```

## Identifier

internal ranking for representation of <identifiers>

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

<b>URL</b>	<ul style="list-style-type: none"> <li>✓ unique</li> <li>– not always persistent</li> <li>– not always resolvable</li> </ul>
<b>PID</b>	<ul style="list-style-type: none"> <li>✓ unique</li> <li>✓ persistent</li> <li>✓ resolvable</li> </ul>
<b>citable PID</b>	<ul style="list-style-type: none"> <li>✓ unique</li> <li>✓ persistent</li> <li>✓ resolvable</li> <li>✓ citable</li> </ul>

🏠 / Datasets / Respostas a um ...

**Social**

- 🐦 Twitter
- 📘 Facebook

👤 Dataset
👥 Communities

### Respostas a um questionário para aferição de disseminação e aceitação de conhecimento tácito

📄 DOI
📄 PID

Lista das 192 respostas a um questionário que pretendeu aferir a disseminação e aceitação de 41 peças de conhecimento tácito no âmbito do combate e prevenção de fogos rurais.

🔍 Incêndios Fogo Incê...

Identifier	
<b>DOI</b>	<a href="https://doi.org/10.23728/b2share.d8cf4efcd2ac48e6b6eb755a62dd6b73">https://doi.org/10.23728/b2share.d8cf4efcd2ac48e6b6eb755a62dd6b73</a>
<b>PID</b>	<a href="http://hdl.handle.net/11304/9aa418ae-3429-4750-b33c-c53d6d5f413e">http://hdl.handle.net/11304/9aa418ae-3429-4750-b33c-c53d6d5f413e</a>
<b>Source</b>	<a href="https://b2share.eudat.eu/api/records/d8cf4efcd2ac48e6b6eb755a62dd6b73">https://b2share.eudat.eu/api/records/d8cf4efcd2ac48e6b6eb755a62dd6b73</a>
<b>Metadata Access</b>	<a href="https://b2share.eudat.eu/api/oai2d?verb=GetRecord&amp;metadataPrefix=oai_dc&amp;identifier=oai:b2share.eudat.eu:b2rec/d8cf4efcd2ac48e6b6eb755a62dd6b73">https://b2share.eudat.eu/api/oai2d?verb=GetRecord&amp;metadataPrefix=oai_dc&amp;identifier=oai:b2share.eudat.eu:b2rec/d8cf4efcd2ac48e6b6eb755a62dd6b73</a>

## Discipline

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

## Example for INRAE

	A	B	C
1	<b>Subject</b>	<b>Discipline</b>	<b>Discipline Id</b>
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

Dataset Communities

INRA:Wheat:36785



Abstract:DI08011 is a Wheat accession from GnpIS.

Genetic Resource Health and Life Sci... Medicine

### Identifier

DOI	<a href="https://doi.org/10.15454/E8FP9Y">https://doi.org/10.15454/E8FP9Y</a>
Metadata Access	<a href="https://data.inrae.fr/oai?verb=GetRecord&amp;metadataPrefix=oai_datacite&amp;identifier=doi:10.15454/E8FP9Y">https://data.inrae.fr/oai?verb=GetRecord&amp;metadataPrefix=oai_datacite&amp;identifier=doi:10.15454/E8FP9Y</a>

### Provenance

Creator	GnpIS
Publisher	Portail Data INRAE
Contributor	Rinnova; URGI data; INRA
Publication Year	2018
Rights	CC BY 2.0; info:eu-repo/semantics/openAccess
OpenAccess	true
Contact	Rinnova (www.inrae.fr); URGI data (urgidata(at)inrae.fr)

### Representation

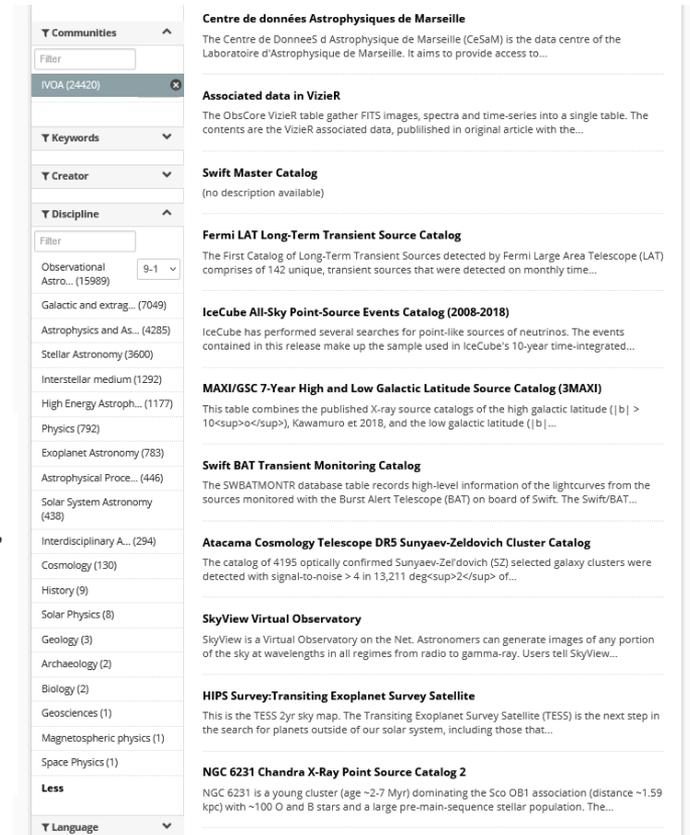
Resource Type	Physical Object; Dataset
Version	3.0
Discipline	Life Sciences; General Genetics

## 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",



The screenshot shows a search interface with a left sidebar and a main content area. The sidebar has a 'Discipline' section with a filter box and a list of disciplines. The main content area shows search results for 'Centre de données Astrophysiques de Marseille' and several other catalogs.

**Communities** (Filter)

- IVOA (24420)

**Keywords**

**Creator**

**Discipline** (Filter)

- Observational Astro... (15989) [9-1]
- Galactic and extrag... (7049)
- Astrophysics and As... (4285)
- Stellar Astronomy (3600)
- Interstellar medium (1292)
- High Energy Astroph... (1177)
- Physics (792)
- Exoplanet Astronomy (783)
- Astrophysical Proce... (446)
- Solar System Astronomy (438)
- Interdisciplinary A... (294)
- Cosmology (130)
- History (9)
- Solar Physics (8)
- Geology (3)
- Archaeology (2)
- Biology (2)
- Geosciences (1)
- Magnetospheric physics (1)
- Space Physics (1)

**Centre de données Astrophysiques de Marseille**  
 The Centre de Données d'Astrophysique de Marseille (CeSAM) is the data centre of the Laboratoire d'Astrophysique de Marseille. It aims to provide access to...

**Associated data in VizieR**  
 The ObsCore VizieR table gather FITS images, spectra and time-series into a single table. The contents are the VizieR associated data, published in original article with the...

**Swift Master Catalog**  
 (no description available)

**Fermi LAT Long-Term Transient Source Catalog**  
 The First Catalog of Long-Term Transient Sources detected by Fermi Large Area Telescope (LAT) comprises of 142 unique, transient sources that were detected on monthly time...

**IceCube All-Sky Point-Source Events Catalog (2008-2018)**  
 IceCube has performed several searches for point-like sources of neutrinos. The events contained in this release make up the sample used in IceCube's 10-year time-integrated...

**MAXI/GSC 7-Year High and Low Galactic Latitude Source Catalog (3MAXI)**  
 This table combines the published X-ray source catalogs of the high galactic latitude ( $|b| > 10^\circ$ ), Kawamuro et al 2018, and the low galactic latitude ( $|b| < 10^\circ$ )...

**Swift BAT Transient Monitoring Catalog**  
 The SWBATMONTR database table records high-level information of the lightcurves from the sources monitored with the Burst Alert Telescope (BAT) on board of Swift. The Swift/BAT...

**Atacama Cosmology Telescope DR5 Sunyaev-Zeldovich Cluster Catalog**  
 The catalog of 4195 optically confirmed Sunyaev-Zeldovich (SZ) selected galaxy clusters were detected with signal-to-noise  $> 4$  in  $13,211 \text{ deg}^2$  of...

**SkyView Virtual Observatory**  
 SkyView is a Virtual Observatory on the Net. Astronomers can generate images of any portion of the sky at wavelengths in all regimes from radio to gamma-ray. Users tell SkyView...

**HIPS Survey: Transiting Exoplanet Survey Satellite**  
 This is the TESS 2yr sky map. The Transiting Exoplanet Survey Satellite (TESS) is the next step in the search for planets outside of our solar system, including those that...

**NGC 6231 Chandra X-Ray Point Source Catalog 2**  
 NGC 6231 is a young cluster (age ~2-7 Myr) dominating the Sco OB1 association (distance ~159 kpc) with ~100 O and B stars and a large pre-main-sequence stellar population. The...

- Standards are great!  
As long as they are used...
- Standards are used when 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) resources
- there will (probably) never be a 'one-fits-all' metadata schema
  - a) the huge variety of data in extremely divergent research areas
  - b) external restraints (national/european law)
- again: focus on interoperability -> requires (human) resources
- 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](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](mailto:martens@dkrz.de)

[fluegel@dkrz.de](mailto:fluegel@dkrz.de)