



Data Discovery

How to find research data and make it
searchable

Heinrich Widmann / DKRZ

Agenda



- Discovery of Research Data (60′)
 - Some Principles, Concepts and Best Practices (15′)
 - Hands-on I : Search and Find Research Data ! (45′)
- Metadata Management (55′)
 - The Meta Data Life Cycle and Ingestion Workflow (15′)
 - Hands-on II : Harvest, Map and Index Metadata (40′)
 - Use case : Metadata from B2SHARE into B2FIND
- Other aspects, references and final reality use case 😊 (5′)

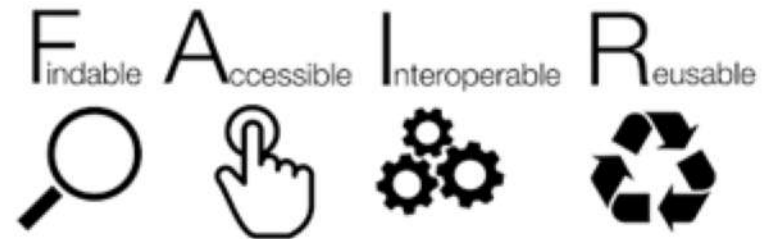
What is Data Discovery ?



... and Why are we looking for things (data) at all ?

- Data discovery is a user (re-Searcher(!)) oriented and iterative process for detecting digital objects and data resources
- Because Research Data are valuable it's crucial to know which data you (not) hold or need and where it is located
- The essential Goal is to Re-use and Compare the found data
- Data Discovery enables the 'F' and the 'A' and aims in the 'R' (and the 'I') of the 'FAIR' principles
- Benefit from discovered data in the following ways
 - Reuse data to save costs and time and to avoid re-inventing the wheel
 - Compare results, make replication studies and share your 'findings'
 - Enhance and assure Quality of (meta) data (FAIRsize (meta)data)

FAIR vs. Open Data



means not (necessarily)



F indability

Handle Values for: 11098/eudat.jnd_004e1b6d-166d-563a-aa80-955679bdaf60

Index Type	Timestamp	Data
1 URL	2018-04-05 16:42:45Z	http://b2find.eudat.eu/dataset/004e1b6d-166d-563a-aa80-955679bdaf60
2 CHECKSUM	2018-04-23 10:21:39Z	57982af67156d42bc29669bc

Metadata-PID (handle)

F1. Metadata tagged with permanent identifier (PID)



Metadata records handled by Metadata-PID

F2. Data described by rich metadata



Schema comprises 21 properties

F3. Metadata indexed in searchable resource



Metadata Catalogue and Discovery Service



- Discoverability is a feature of search tools
- Findability is a property of data
- Persistent Identifiers (PIDs) – for data and metadata - is a central concept of Data Management
- Discovery Portal with user-friendly GUI and faceted search simplifies search

A ccessibility

Identifier	
DOI	http://dx.doi.org/10.1094/PHYTO-0915-1501A
Metadata Access	http://dx.doi.org/10.1094/PHYTO-0915-1501A

B2FIND Identifier *DOI* and *MetadataAccess*

A1. Metadata retrievable by standard protocols



OAI-PMH and other standards used to harvest and disseminate Metadata

A2. Metadata accessible even when data are not



Sustainable storage of Metadata in Open Repository



Metadata display with DOI and Landing page



- Accessibility of data is supported by
- Following standards for Harvesting, Resolution of links (URIs) to data
 - Providing identifiers, which redirects to landing page or directly to referenced data collection
 - Sustainable storage of and access to Metadata

I nteroperability

For B2FIND Metadata Schema see at
<http://b2find.eudat.eu/guidelines/mapping.html>

I1. Metadata use a **shared language** for knowledge



Metadata schema is based on **DataCite 4.1**

I2. Metadata use FAIR **vocabularies**



E.g. Taxonomy of **Research Disciplines**



EUDAT-B2FIND follows common standards :

- Metadata Schema is based on DataCite's Schema
- Closed vocabs are used, e.g.
 - ‚Classification of Research Area‘ for facet Discipline is developed with re3data
 - Languages are mapped according iso363

R eusability

Publisher	Max-Planck-Institut fuer Meteorologie, Deutsches Klimarechenzentrum (DKRZ)
Publication Year	2011
Rights	For Scientific Use only

Provenance Information

R1. Rich metadata with relevant attributes



Licences (R1.1)
and
Provenance (R1.2)

R2. Domain-relevant standards



Support of
(m)any **specific metadata formats**
and harvesting **APIs**



Reusability is enabled by providing metadata on

- Access and Usage Licences
- Provenance

and by support of

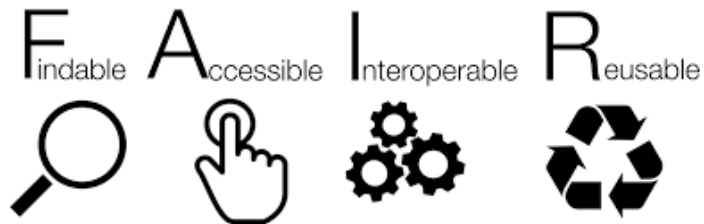
- of various specific metadata formats
- by different APIs to harvest metadata (OAI-PMH, JSON-APIs, CSW and others)

The Process of Data Discovery



- You can structure your search according to the following (iterative) steps (based on www.cessda.eu/DMEG)
 - Develop a clear picture of the research data you need, specify the 'search target' and build the search request
 - Choose and locate appropriate search interfaces and/or data sources
 - Set up and submit the search request
 - Check the response (found items) and select data candidates
 - Analyse how far they fit to your search criteria and evaluate data quality
 - Adjust search query in case not suitable data were found and go to 1.

EUDAT-B2FIND



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



- = MD Catalogue + Search Index + Discovery Portal
- Follows FAIR principles
- Interdisciplinary
- Covering wide range of Research Domains
- Research Communities can make Research Data visible and findable
- Uptake follows Low Barrier Approach
- End-users can search, browse, find and access data on a cross-domain level

B2FIND Discovery Portal



B2FIND provides 'faceted' search for

- Free text
- Geospatial
- Temporal coverage
- Publication year
- Textual facets as
 - Tags
 - Creator
 - Discipline etc.

Dataset view provides display of metadata :

- Spatial extent
- Links to data resources

The screenshot shows the B2FIND Discovery Portal interface. At the top, there is a search bar with the text "Search datasets..." and a magnifying glass icon. Below the search bar, it indicates "862 datasets found" and an "Order by: Relevance" dropdown menu. The main content area displays the "Collection of Hymenoptera" dataset. On the left, there is a "Dataset extent" map showing Europe and Africa. Below the map are social media sharing options for Social, Google+, Twitter, and Facebook. The right side of the page features a "Collection of Hymenoptera" title, a description of the collection, and a table of "Additional Info".

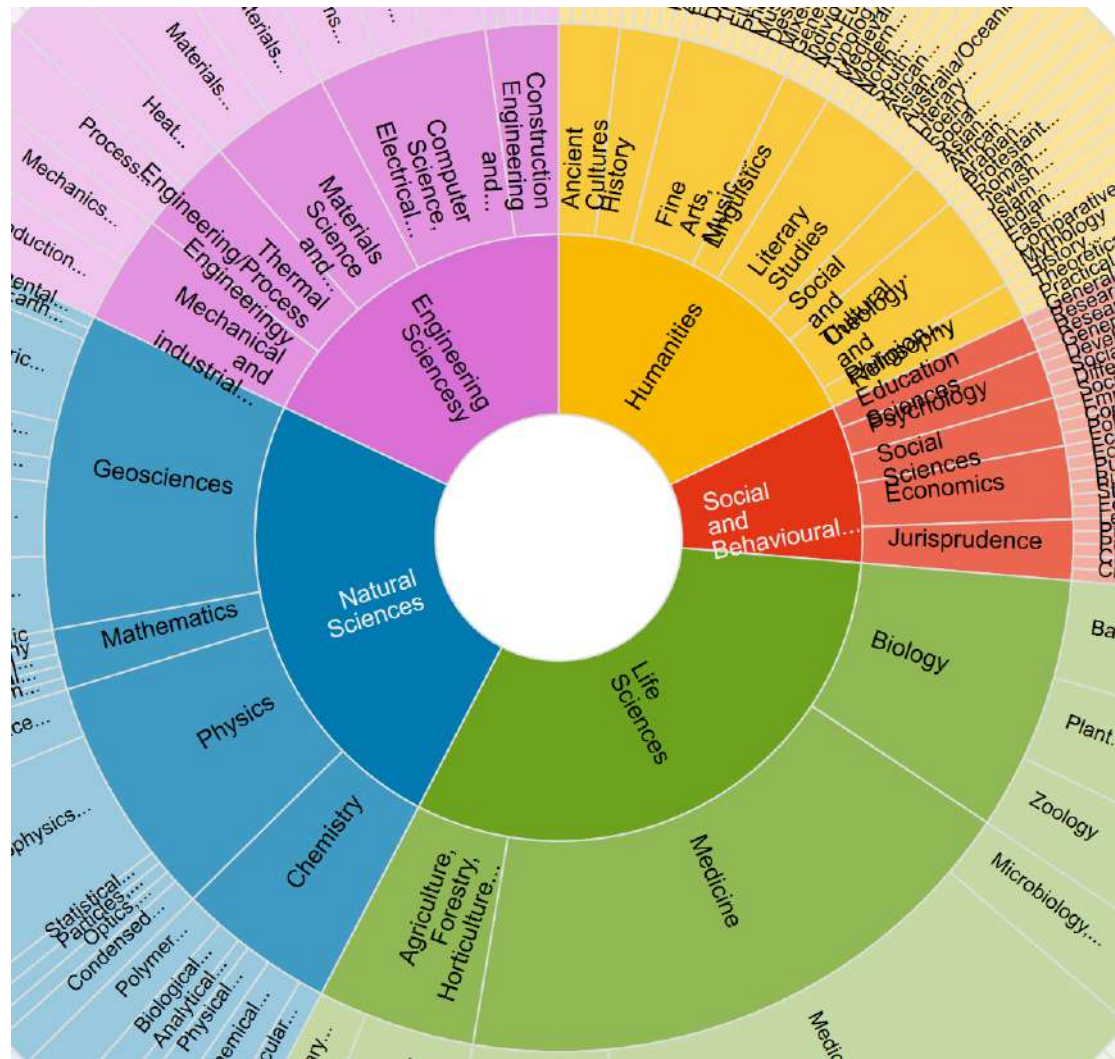
Field	Value
Source	http://212.87.9.194/tapir/tapir.php/uwr-mnhw-hymenoptera
Discipline	Biology
GeographicCoverage	NorthernEurope,SouthernEurope,EasternAsia,SouthernAsia,AustraliaandNewZealand,NorthernAfrica,CentralAsia,EasternEurope,WesternEurope,SouthAmerica,WesternAsia
MetadataAccess	http://metadata.gbif.org/catalogue/OAIHandler?verb=GetRecord&metadataPrefix=eml&identifier=oai:metadata.gbif.org:eml/portal/oai:metadata.gbif.org:eml/portal/1453.xml
Origin	Wroclaw University, Museum of Natural History
PublicationYear	2007

Categoration of Research Areas

re3data and B2FIND are developing in the project [clarascience](http://clarascience.org) a 'Classification of Research Areas'

Graphical Interface to browse through 'Research Areas' → <http://eudat7-ingest.dkrz.de/statistics/disciplines.html>

(work in progress)



Extract of the GUI for browsing through B2FIND's research disciplines

Hands-On I : Find data !



Work in groups and follow the instructions in

https://gitlab.eudat.eu/eudat-prace-2019/instructions_for_datadiscovery_handson/
Hands-On I : Discovery of Research Data / Excercise I. Discover data.md

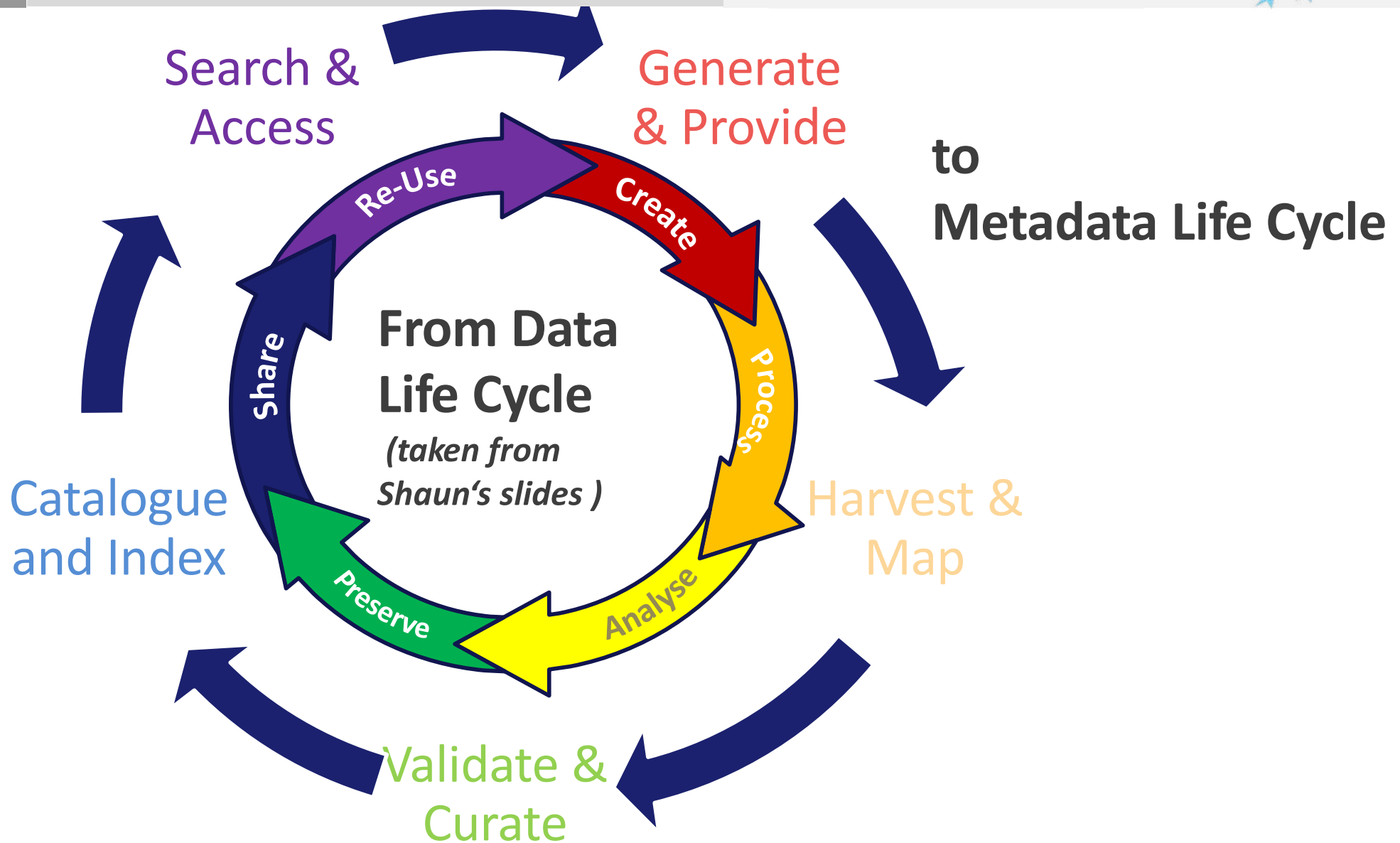
- Each group should agree on 1-3 'use/search case' from tab 'Search cases' or add your own one
- Go ahead ! : Choose portals, submit search requests, refine and make your user experience during your search voyage
- But don't forget to discuss and document user experience and search paths
- Finally summarize, present and exchange your findings about 'finding' 😊 afterwards

Metadata Management

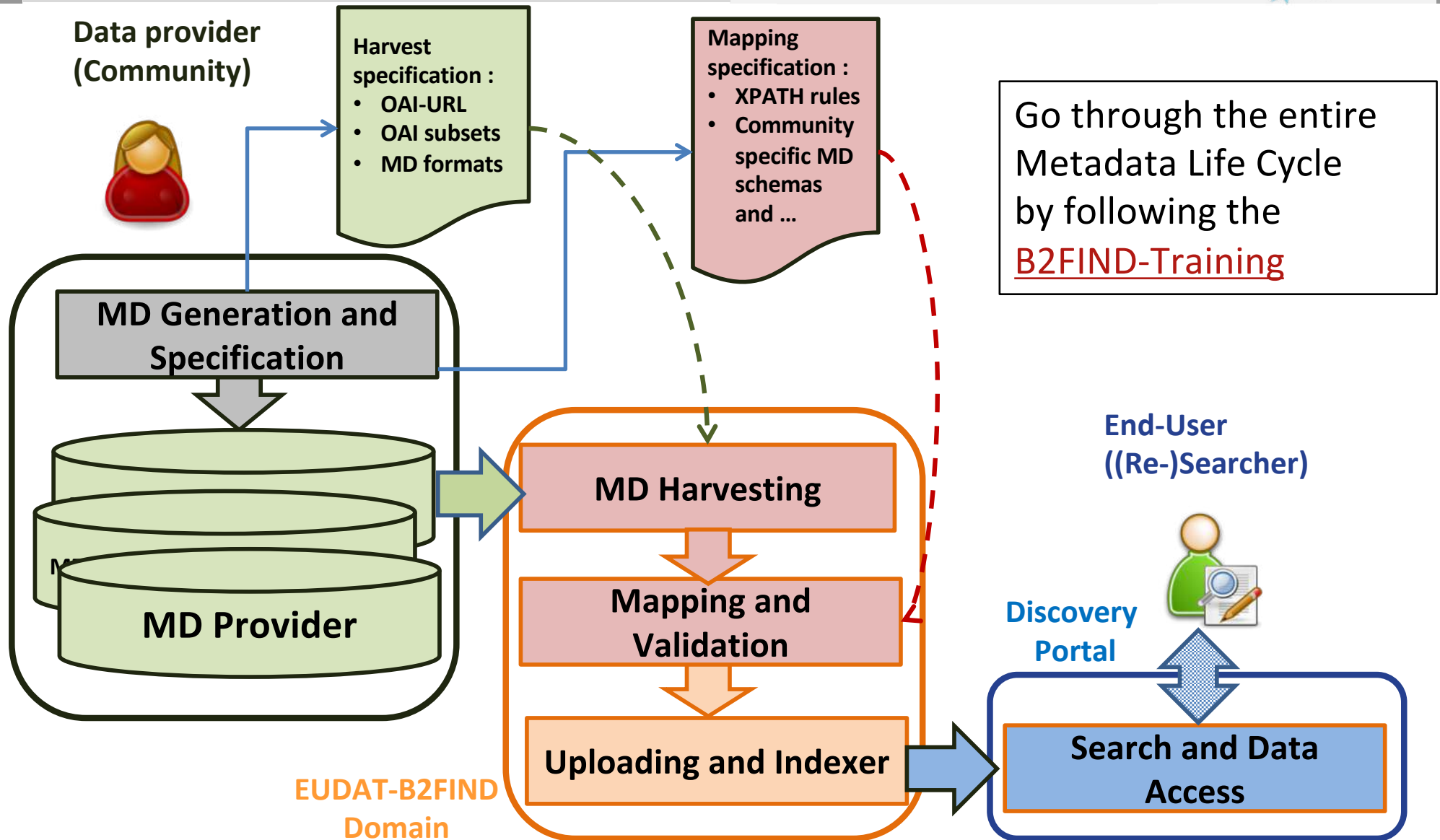


- (Good) Metadata (Management) is crucial for Data Discovery !
- We don't talk about 'What is Metadata' in this session (see Shaun's session and in the link list)
- Metadata Management comprises the process of 'Making data searchable and accessible by means of metadata'
- This process comprises essentially collecting, mapping, indexing and presenting Metadata
- Most archives, repositories and generic discovery services implement this by a Metadata Catalogue, a Search Index and a Discovery Portal (GUI for search)
- But search portals differ in architecture, design, scope, addressed audience, granularity, etc.pp.

Metadata Life Cycle



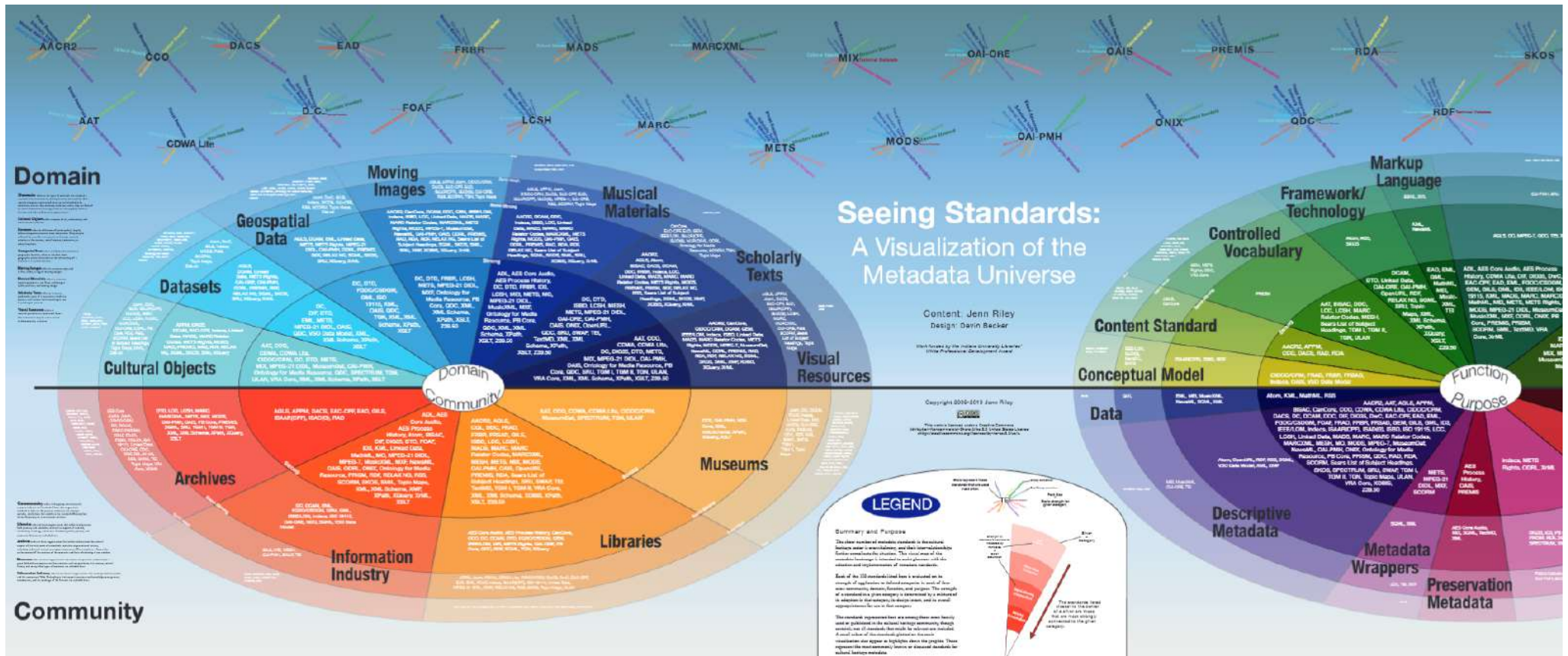
Metadata Ingestion Workflow



The Universe of Metadata Standards



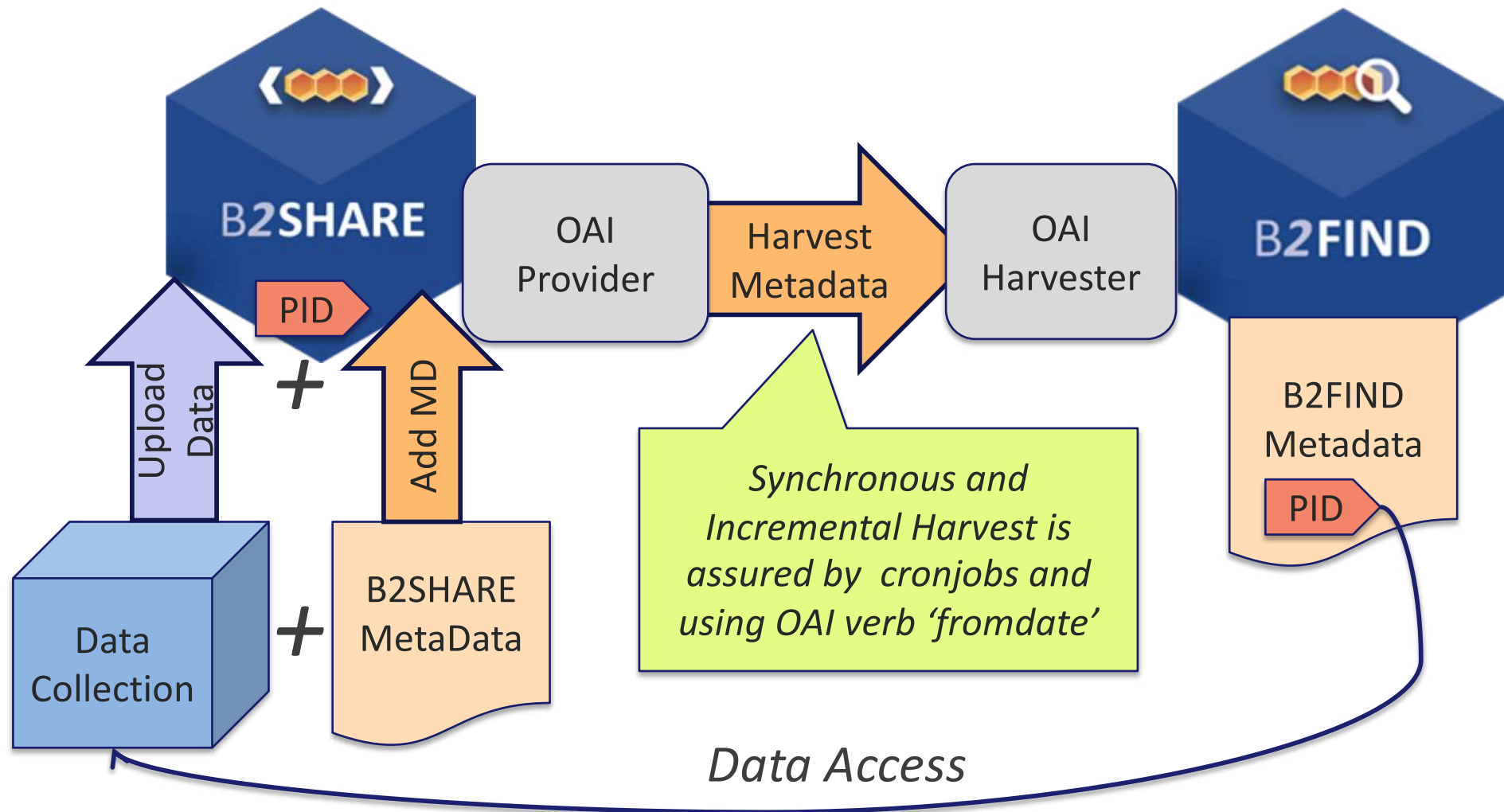
<http://jennriley.com/metadatamap/seeingstandards.pdf>



Copyright 2009-2010 Jenn Riley
This work is licensed under a CC-AN-SA 3.0 .



Metadata from B2SHARE to B2FIND



Hands-On II : Make data findable



Follow the instructions on

https://gitlab.eudat.eu/eudat-prace-2019/instructions_for_datadiscovery_handson/
Hands-On II : Make your Data Findable

1. Checkout the B2FIND-Training from <https://github.com/EUDAT-Training/B2FIND-Training> and load needed Python packages
2. Go through the 'Metadata Ingestion' workflow by publishing B2SHARE metadata in B2FIND :
 - a. [Generate XML, formatted as DublinCore, from comma separated list (netCDF example ?)]
 - b. Harvest XML from B2SHARE's OAI endpoints (via Browser / command line)
 - c. Map these XML records to JSON files following the B2FIND schema
 - d. Validate and assess these metadata records
 - e. Upload and index JSON records in the training instance of B2FIND

Other Aspects and Challenges



- Machine actionable searching for scientific knowledge
- Deep (Meta)Data Mining
- Linked Open Data and Graph databases (a wide field)
- Scalability, Granularity and Scalability
- (Meta) Data Curation and Quality (FAIRisizer)
- Trust : Don't trust data without (good and rich) metadata and without knowledge about data provider and producer (quality over quantity)
- Interoperability : Foster common formats, standards, protocols and tools to share excellent research in an excellent way
- User experience : Your feedback is highly appreciated and needed
!!!

Semantic Issues

'This is not pipe', but

- Metadata about a object called pipe ?
- No, just a visualisation of something which looks like a pipe ?
- No, it's an artwork from 'Magritte', called (in English) 'The Treasury of a Images'
- No, it's a picture of an artwork , which stresses (in French) that this is NOT a pipe...
- Or something completely different, damn, ... ???
- Your interpretation, please !



"La trahison des images (Ceci n'est pas une pipe), 1929" by pierrepaul43 is licensed under CC BY-NC-SA 2.0

Identify Scope, Type and Semantics of Referenced Object (by using Ontologies) !

Metadata of the artwork ,The Treachery of Images' as shown by wikipedia

https://en.wikipedia.org/wiki/The_Treachery_of_Images

The Treachery of Images



Artist	René Magritte
Year	1929
Medium	Oil on canvas
Dimensions	60.33 cm × 81.12 cm (23.75 in × 31.94 in)
Location	Los Angeles County Museum of Art ^[1]

Synonyms

Synonyms of described Object

A ‚Pipe‘ can be a lot of things → <https://en.wikipedia.org/wiki/Pipe>

Pipe

From Wikipedia, the free encyclopedia

For information on the use of "pipe links" on Wikipedia, see WP:PIPE.

Pipe may refer to:

Common uses [edit]

- **Pipe (fluid conveyance)**, a hollow cylinder following certain dimension rules
- **Piping**, the use of pipes in industry
- **Smoking pipe**

Places [edit]

- **Pipe, Wisconsin**, United States
- *Pipe*, the Hungarian name for Pipea village, **Nadeș Commune**, Mureș County, Romania


People [edit]

- **Jules Pipe** CBE, Mayor of the London Borough of Hackney, UK
- **Pipes** (surname)

Arts, entertainment, and media [edit]

Music [edit]

- **Pipe (instrument)**, a traditional perforated wind instrument
- **Bagpipe**, a class of musical instrument, aerophones using enclosed reeds
 - **Pipes and drums** or pipe bands, composed of musicians who play the Scottish and Irish bagpipes

 Look up *Pipe* or *pipe* in Wiktionary, the free dictionary.

Contents [hide]

- 1 Common uses
- 2 Places
- 3 People
- 4 Arts, entertainment, and media
 - 4.1 Music
 - 4.2 Other uses
- 5 Brands and enterprises
- 6 Technology
- 7 Other uses
- 8 See also

Hands-On 3 : Find the real pipe

Where is the Pipe ?

- There is a real pipe hidden on the CINECA area
- Search as well on the terrasse !
- Note : There is another meaning of the German translation 'Pfeife' !
 - You will find a lot of 'Pfeifen' around this place and in the whole world,
 - but we mean here a thing and not this kind of human beings (for whom actually no search engine is needed 😊)
- Good luck and good findings !

References



- B2FIND's Guidelines for Data Providers → <http://b2find.eudat.eu/guidelines>
- GO-FAIR Discovery IN →
- RDA Metadata Paradigm Interested Group →
- ...