Scientific Data Curation

Luigia Cristiano, HMC Hub Matter

Helmholtz Institute Berlin

www.helmholtz-metadata.de

Agenda

Practices for data curation:

Data description and staging

- **Write a clear documentation for the data interpretation and reuse**
- **Organize and name the files to improve the findability and reproducibility of the acquisition and analysis process**
- **Use community standards for data and metadata when available**
- Use repositories for staging the data
- Automatize the data processing and curation
- Assign PIDs to datasets, software and cite them in your publication

Why this training

Familiarize end users with basics of scientific data curation

- How to extract, format and standardize data description
- How to stage the data locally and in institutional repository

- Overall we want to show the benefits of the compliance to the FAIR data guidelines of your reseach data

Why do we need scientific data curation

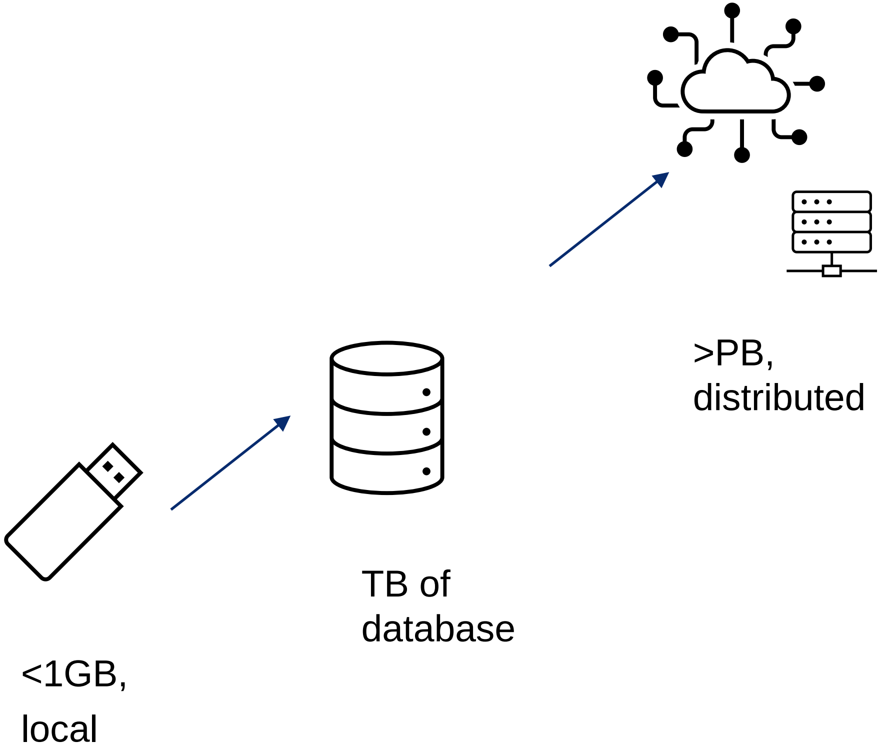
Evaluate the data quality, completeness

Provide full information for the data interpretation and data reuse

Enable the data discovery

Speed up the processes and test the robustness of the staging and processing solutions

- Independently of data size and format
- Independently of the data analysis method
- Independently of the data storage solution



<1GB, local

TB of database

>PB, distributed

Why do we need scientific data curation

Where did I store that file ?

How can I control the effect of the analyis parameters on the model resolution ?

Which parameter did I use ?

When and how did I generate this plot ?

How can I speed up the process ?

Why do we need scientific data curation

Where did I store that file ?

How can I control the effect of the analyis parameters on the model resolution ?

Protocol of the analysis in the form of e-labbook, README, tags, parameter files

Data structures optmized for specific analysis processes

Which parameter did I use ?

How can I speed up the process ?

Data structures representing the procedures implemented

When and how did I generate this plot ?

Systematic file naming

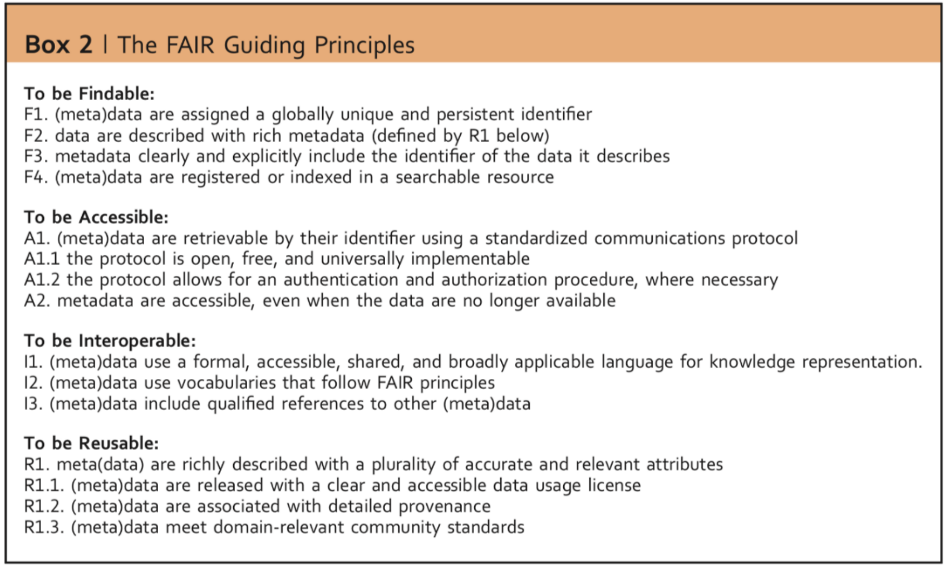
FAIR

- FAIR is not Open Access
- Different solutions and different paces of implementation are suggested for the different communities

Guidelines for promoting data visibility, reuse and ensure data quality

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

HELMHOLTZ
METADATA
COLLABORATION

01 **Scientific data documentation**

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

Who ?

What ?

When ?

Why ?

How ?

Where ?

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

- **Proposal**
- Acquisition
- Reduction
- Analysis
- Publication

A Research Data Management Plan is available

- **A Research Data Management Plan is submitted**
- Short description of the experiment
- PIs and co
- Instruments reservation
- Beam time/Lab reservation
- Samples description
- Start time and end time are indicated

Research Data Management Plan

EU Projects guidelines

e.g. Horizon 2020

- **RDMP as live document**
- Must be kept up to date
- It represents a documentation of the pre-measurement phase
- It contains information of the potential size of the database and the proposed staging solutions and conditions for the data access
- It should be edited in collaboration with the RDM group at the institute

- https://ec.europa.eu/futurium/sites/futurium/files/d6.1_693849_data_management_plan.pdf

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

- Proposal
- **Acquisition**
- Reduction
- Analysis
- Publication

- Control software (author, license, version)
- Instrument type and acquisition setting (possibly ID)
- **Parameter file**
- **Sample preparation workflow**
- **File naming convention**
- **Data structure**
- Log file
- Location
- Creator
- Lab-book

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

- Proposal
- Acquisition
- **Reduction**
- Analysis
- Publication

- Software (author, license, version)
- **Parameter file**
- Log file (step by step procedure to ensure reproducibility)
- Location
- Creator
- Connecting derivative and raw data

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

- Proposal
- Acquisition
- Reduction
- **Analysis**
- Publication

- Software (author, license, version)
- **Parameter file**
- Log file (step by step procedure to ensure reproducibility)
- Location
- Creator
- Connecting derivative and raw data
- Connect the post analysis with pre analysis data set
- File naming convention for the graphical output

Data life cycle and data curation

Each phase of the life cycle is interested by data curation activity

At each phase it is essential to implement FAIR principles

- Proposal
- Acquisition
- Reduction
- Analysis
- **Publication**

- Software (author, license, version)
- **Data workflow**
- Log file (step by step procedure to ensure reproducibility)
- Location
- Contributors and authors
- Make data accessible in a standard-open format

What are the key information to preserve

1. How are the key parameters identified ?

2. How are these informations captured ?

- **Proposal**
- Acquisition
- Reduction
- Analysis
- Publication

Who ?

What ?

When ?

Why ?

How ?

Where ?

What are the key information to preserve

1. How are the key parameters identified ?

- **Proposal**
- Acquisition
- Reduction
- Analysis
- Publication

- Sample characteristics
- **PIs**
- Instruments
- Location
- Contributors and authors
- Experiment settings
- Experiment begin and end time
- File convention and total size
- Abstract
- Project description

What are the key information to preserve

2. How are these informations captured ?

- **Proposal**
- Acquisition
- Reduction
- Analysis
- Publication

- e.g. GATE system (https://www.helmholtz-berlin.de/pubbin/hzbgate)
- **Sample Formula and tracking system or sample ID (https://www.igsn.org/)**
- Instruments ID (https://www.helmholtz-berlin.de/pubbin/igama_output?modus=einzel&sprache=en&gid=2127&typoid=75136)

- Location
- Institution and researcher ID (we will talk about this in the coming slides)
- Community standard for File naming and data structure (we will talk about this in the coming slides)

What are the key information to preserve

1. How are the key parameters identified ?

- Proposal
- **Acquisition**
- **Reduction**
- **Analysis**
- Publication

- Software, scripts versions
- **Parameter file**
- File convention and total size

What are the key information to preserve

2. How are these informations captured ?

- Proposal
- **Acquisition**
- **Reduction**
- **Analysis**
- Publication

- Log files
- Input parameter files
- Output files with automatic naming
- Output data structure
- Workflow track system
- Elog books
- Elab books
- File headers
- codebooks

Data description, tabular

- Readme for the datafile
- Title for the dataset
- Name/institution/email/ORCID
- Contributors and authors, contact person
- Acquisition time
- Geographic location
- Language information
- Keywords to describe the topic of the study
- Information about funding program and agencies
- Versioning
- Link to related data collection
- Information about data type, variables definition, unit of the measured quantities
- Explanation on acquisition settings
- Uncertainty associated, gaps
- Short abstract

Tools

- Elab book
- Parameter files
- Activity reports
- Files header
- Data processing workflow and logfiles
- Codebooks
- Data acquisition software
- Data management software

Automated generation of metadata by data staging workflows (from measurements to publication

How to preserve informations

From README to Metadata:

- Proposal
- Acquisition
- Reduction
- Analysis
- Publication

Let’s create a README for a specific experiment workflow :

Hand on session: README file

Hand on session: Elab Book

How to capture informations with elab-books

From Hand notes to electronic lab-books:

- **icat**
- ElabFTW
- Chemotion
- Nomad-labbook

Let’s watch a demo!

Federated access to digital tools

AAI

Let’s watch a demo!

How to preserve information in a standardized manner

Persistent Identifiers and Metadata:

- Person
- Institute
- Sample
- Instrument
- Publication: software and datasets

02 **Metadata**

www.helmholtz-metadata.de

Data about data

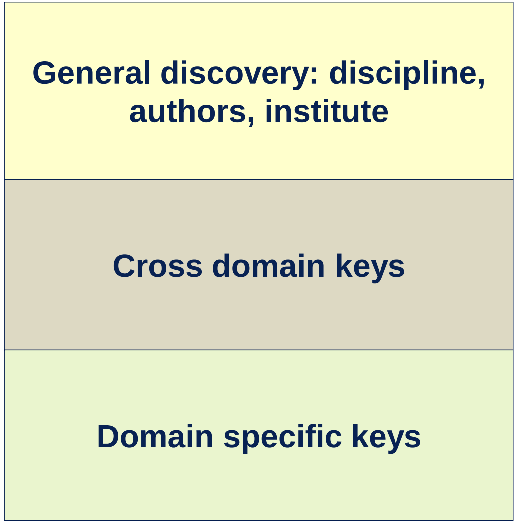
How to ensure your data are understable and usable by others

- Machine and human redable data documentation: structured information
- Metadata organized in keys and values are descriptors to data discovery, interpretation and analysis
- Metadata keep info on data provenance, copyrights and access restrictions
- Metadata can be generic or discipline specific

Metadata are remaining even when the data are gone

[1]

Metadata: different levels of informations



General discovery: discipline, authors, institute

Cross domain keys

Domain specific keys

- Minimum set of data should allow the discovery and citation- Mandatory

Metadata categories

Administrative	Authors ID, dataset id, time frame, embargo, versioning
Descriptive	Methodological and geog. Info, data quality metrics
Structural	Discipline dependent
Ancillary informations	Info on the preprocessing, or derived metrics

Lib.ua.edu

[2]

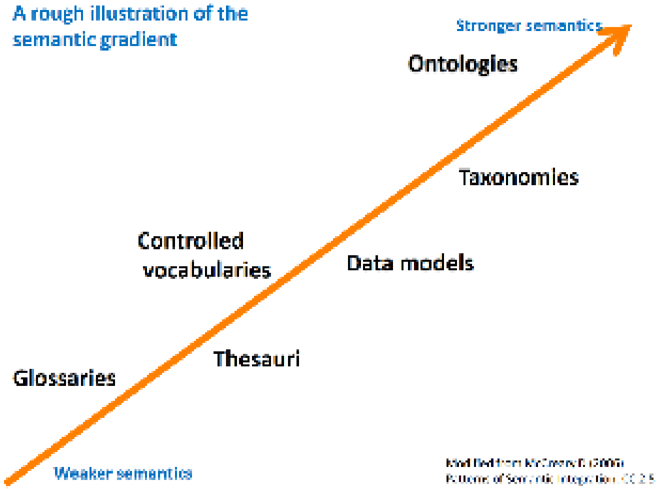
Metadata Structure

Thesaurus for the definition of categories

Dictionary to extract the components Ixml,json)

Hierarchical structure supported by a taxonomy

Metadata elements (categories/keys) and associated values Defined by ontologies



Modified from McCreary D. (2006)
Patterns of Semantic Integration. CC 2.5

The metadata can be distributed attached to the data as „header“ or as separated file. Both options offer advantages and disadvantages

[3]

Metadata Structure

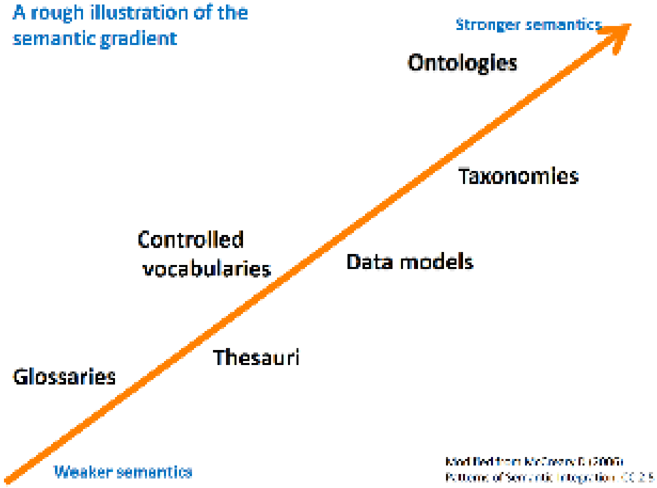
Thesaurus for the definition of categories

Dictionary to extract the components lxml,json)

Hierarchical structure supported by a taxonomy

Metadata elements (categories/keys) and associated values Defined by ontologies

Let‘s see some examples!



Modified from McCreary, D. (2006) Patterns of Semantic Integration. CC 2.5

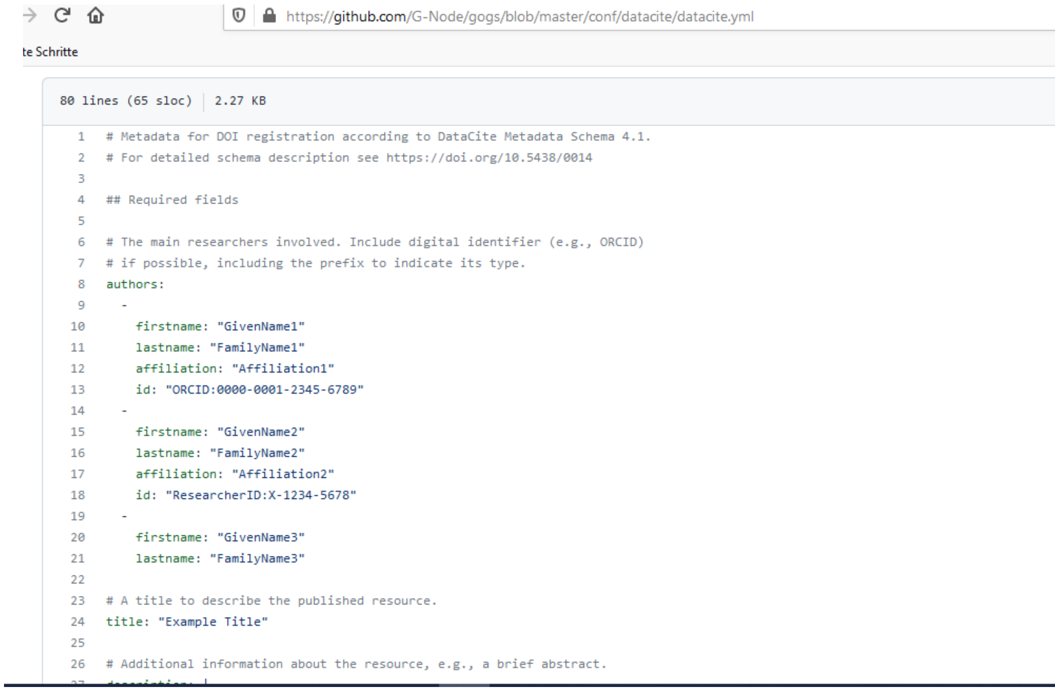
The metadata can be distributed attached to the data as „header“ or as separated file. Both options offer advantages and disadvantages

[3]

Formats converter

https://github.com/G-Node/gogs/blob/master/conf/datacite/datacite.yml

German Neuroinformatics Node, guidelines for DOI request


https://github.com/G-Node/gogs/blob/master/conf/datacite/datacite.yml
te Schritte
80 lines (65 sloc) | 2.27 KB

```
# Metadata for DOI registration according to DataCite Metadata Schema 4.1.
# For detailed schema description see https://doi.org/10.5438/0014

## Required fields

# The main researchers involved. Include digital identifier (e.g., ORCID)
# if possible, including the prefix to indicate its type.
authors:
  -
    firstname: "GivenName1"
    lastname: "FamilyName1"
    affiliation: "Affiliation1"
    id: "ORCID:0000-0001-2345-6789"
  -
    firstname: "GivenName2"
    lastname: "FamilyName2"
    affiliation: "Affiliation2"
    id: "ResearcherID:X-1234-5678"
  -
    firstname: "GivenName3"
    lastname: "FamilyName3"

# A title to describe the published resource.
title: "Example Title"

# Additional information about the resource, e.g., a brief abstract.
description: |
```

Several routines offer format conversions:

Good compromise between word/ASCII and json dictionary is the yaml format

It allows introducing a structure in free text environment.

Python functions/packages e.g.

[5]

Metadata schema

Generic

DublinCore, DataCite,

Schema.org, EUDAT (comm. Dep)

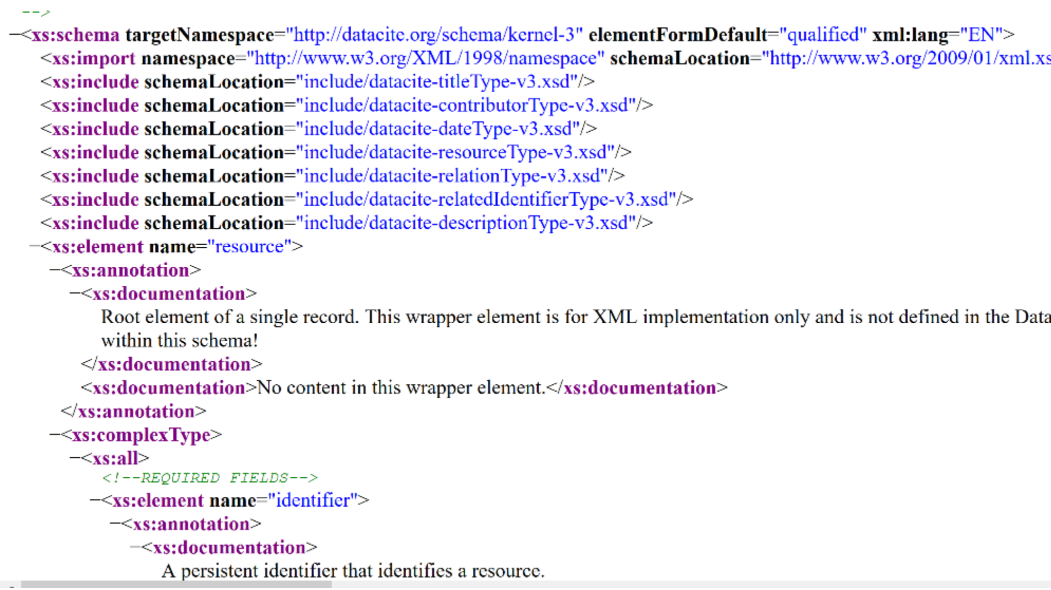
https://www.eudat.eu/

The Simple Dublin Core Metadata

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Formats

Xml, csv, rdf, html, json



```
 -->
−<xs:schema targetNamespace="http://datacite.org/schema/kernel-3" elementFormDefault="qualified" xml:lang="EN">
   <xs:import namespace="http://www.w3.org/XML/1998/namespace" schemaLocation="http://www.w3.org/2009/01/xml.xs
   <xs:include schemaLocation="include/datacite-titleType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-contributorType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-relationType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
  −<xs:element name="resource">
    −<xs:annotation>
      −<xs:documentation>
         Root element of a single record. This wrapper element is for XML implementation only and is not defined in the Data
         within this schema!
       </xs:documentation>
       <xs:documentation>No content in this wrapper element.</xs:documentation>
     </xs:annotation>
    −<xs:complexType>
      −<xs:all>
         <!--REQUIRED FIELDS-->
        −<xs:element name="identifier">
          −<xs:annotation>
            −<xs:documentation>
               A persistent identifier that identifies a resource.
```

Discipline dependent

Nxdl

https://github.com/nexusformat/definitions

https://github.com/nexusformat/definitions/blob/main/nxdl.xsd

[4]

https://schema.datacite.org/meta/kernel-4.1/example/datacite-example-full-v4.1.xml
https://www.fdsn.org/xml/station/fdsn-station-1.1.xsd

Metadata schema

Generic

DublinCore, DataCite,

Schema.org, EUDAT (comm. Dep)

https://www.eudat.eu/

Let‘s look at the schema features!

Discipline dependent

Nxdl

https://github.com/nexusformat/definitions

https://github.com/nexusformat/definitions/blob/main/nxdl.xsd

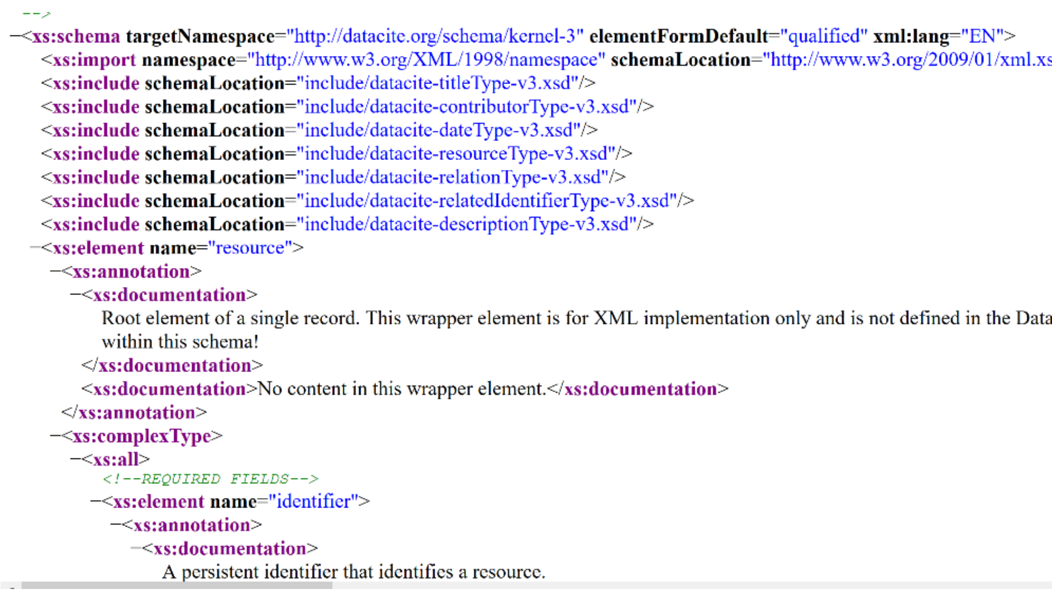
[4]

The Simple Dublin Core Metadata

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Formats

Xml, csv, rdf, html, json



```
 -->
-<xs:schema targetNamespace="http://datacite.org/schema/kernel-3" elementFormDefault="qualified" xml:lang="EN">
   <xs:import namespace="http://www.w3.org/XML/1998/namespace" schemaLocation="http://www.w3.org/2009/01/xml.xs
   <xs:include schemaLocation="include/datacite-titleType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-contributorType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-relationType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.xsd"/>
   <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
  -<xs:element name="resource">
    -<xs:annotation>
      -<xs:documentation>
         Root element of a single record. This wrapper element is for XML implementation only and is not defined in the Data
         within this schema!
       </xs:documentation>
       <xs:documentation>No content in this wrapper element.</xs:documentation>
     </xs:annotation>
    -<xs:complexType>
      -<xs:all>
         <!--REQUIRED FIELDS-->
        -<xs:element name="identifier">
          -<xs:annotation>
            -<xs:documentation>
               A persistent identifier that identifies a resource.
```

https://schema.datacite.org/meta/kernel-4.1/example/datacite-example-full-v4.1.xml

https://www.fdsn.org/xml/station/fdsn-station-1.1.xsd

Hands-on session: YAML, XML, JSON

HELMHOLTZ
METADATA
COLLABORATION

03 **Scientific data structures**

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Data formats and data structures

Formats

- Prefer open
- Community standards
- Look at the repository recommendation
- Machine readable (metadata)

Structures

- Keep the versioning in the filenaming
- File versioning system
- Codes for format conversion
- Associate readme, codebook, data dictionary and data list

[12]

Data documentation workflow

- **Attach-embed-link** to the data all the information relative to the data production and data analysis.
- **Log and parameter** files help the tracking of the analysis settings and facilitate the reproducibility of the data processing
- Keep **updated** the documentation with an history of the data processing
- Use of scripting to **automatize** the update and associate quality parameters to the data analysis
- Use templates or define one. Use structured doc

Data organization

- Optimize the data exploration and the data processing
- Which structure is optimal for your data set depends on the analysis tools, the data type and the data processing you need to perfom
- The access protocol and also the access terms to the data (raw or processed) plays also a key role on deciding which database structure is the most suitable for your research work

- **SQL versus noSQL databases**

(training module on this topic is planned, register to our mailing list if interested)

- **Svn vs Git repository**

(training module on this topic is planned, register to our mailing list if interested, Git training is offered by HIFIS https://events.hifis.net/category/4/)

Data organization

General guidelines for the data organization:

Raw data---optimized for the access and back-up

Processed data:

- Clone of the raw data structure

- Folder naming associated to the analysis method

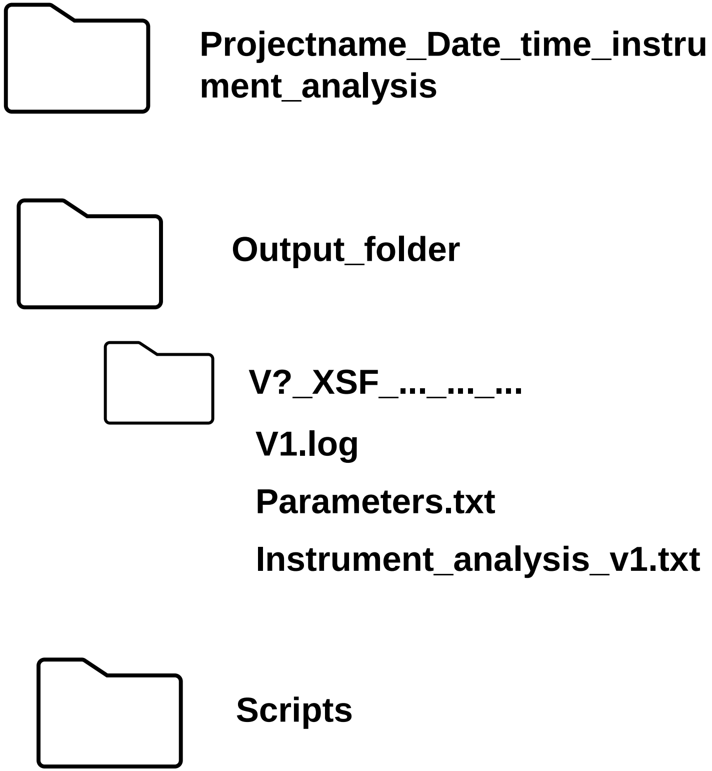
- Versioning of the file, parameter and log files saved in the folder

- Readme and vocabulary for abbreviations

- Keep the syntax machine readable

Defining a file name convention facilitates the data discovery

A complete data description should be available in each folder



Projectname_Date_time_instrument_analysis

Output_folder

V?_XSF_..._..._...

V1.log

Parameters.txt

Instrument_analysis_v1.txt

Scripts

Hands-on session: file naming and data exploration

How prepare data to FAIRness

- Identifier to the data, repository, metadata, licences
- Standard or not proprietary formats
- Keep data documented and versioned
- Project or community repository
- Cross references to related data

Data organization

From Files Collections to databases:

- Proposal
- Acquisition
- Reduction
- Analysis
- Publication

03 **Helmholtz Metadata Collaboration**

www.helmholtz-metadata.de

HMC role

- **The FAIR principles will be guiding** the development of data management services

HMC promotes planning on data curation and the implementation of standardized and automate procedure for the data staging.

Focus is the future reuse of the own data by developing tools from our consultancy with researchers.

We support researchers to understand what to share with data and which data are to be shared.

We offer solutions for the data FAIRification

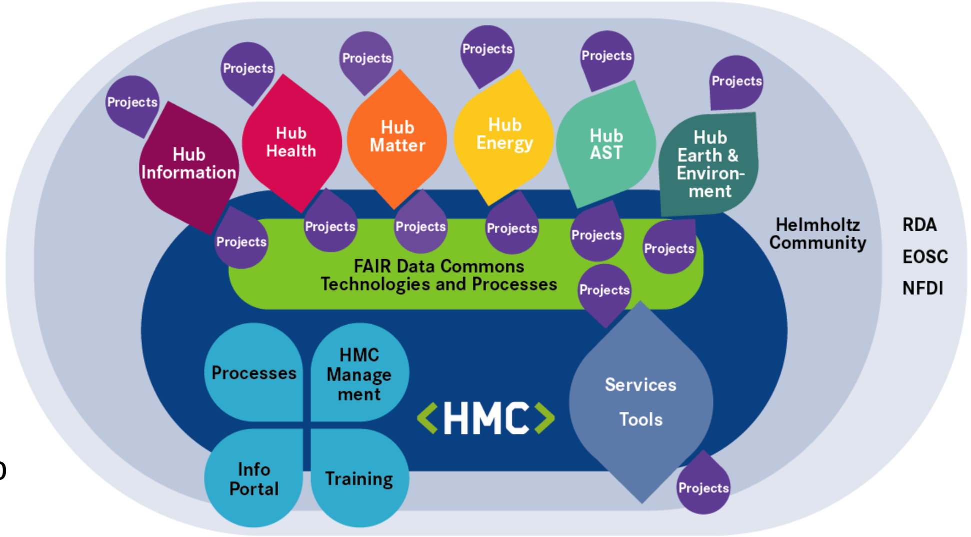
HMC role

- **The FAIR principles will be guiding** the development of data management services

The mission of the Helmholtz Metadata Collaboration (HMC) is to facilitate the discovery, access, machine readability, and reuse of research data of the Helmholtz Association.

HMC promotes a coordination with the scientific community for the services development in order to establish widely accepted practices in the handling of research data.

Implementation of standards for the data description to favour the data interoperability and discovery are also goals of our initiative



[7]

HMC activities as FAIR initiatives

- **GO BUILD** services for FAIR technology of data management
- **GO CHANGE** promote the research management policies and incentives the FAIR implementation
- **GO TRAIN** transfer of skills and FAIR awareness

Our training is promoting the compliance of good data management practices in different researchers communities for a Bottom-Up implementation of global FAIR research database

04 **Community standards**

NeXus Format

Metadata and data format: Nexus

- Nexus for rich and easy to access metadata.

  Why NeXus ? https://www.nexusformat.org/
- HDF5/NeXus used as institutional standard at neutron, x-ray and muon facilities
- Each facility diversify the dictionary limiting the immediate re-usability.
- NeXus files may help to improve the situation.
- HDF5 format and a tree structure for metadata representative of the complexity of PaN data
- Built in Vocabulary for research community interoperability
- Geometry of the beamline, sample stages, orientation and description of detectors, exposure time, beamline calibration info, scan description
- To store multiple related data set create more entries

[7]

Hierarchy in Nexus

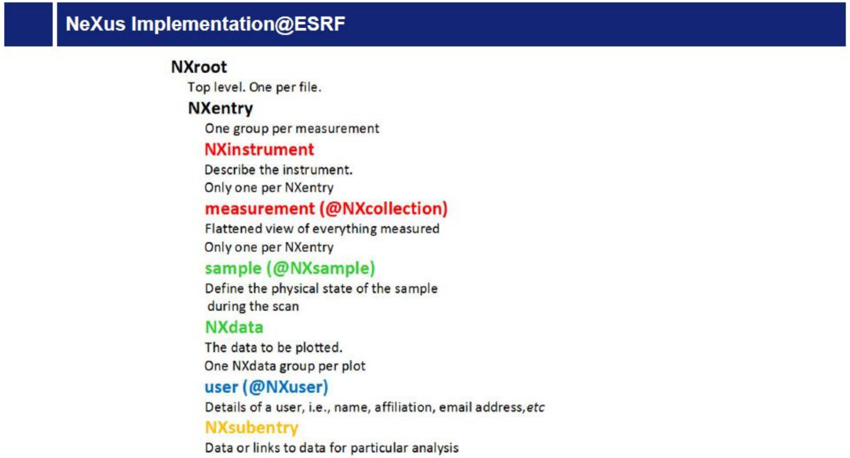
Classes (dictionary)

Groups

Levels

Attribute

MultiD array and scalars



NeXus Implementation@ESRF

NXroot
Top level. One per file.

NXentry
One group per measurement

NXinstrument
Describe the instrument.
Only one per NXentry

measurement (@NXcollection)
Flattened view of everything measured
Only one per NXentry

sample (@NXsample)
Define the physical state of the sample
during the scan

NXdata
The data to be plotted.
One NXdata group per plot

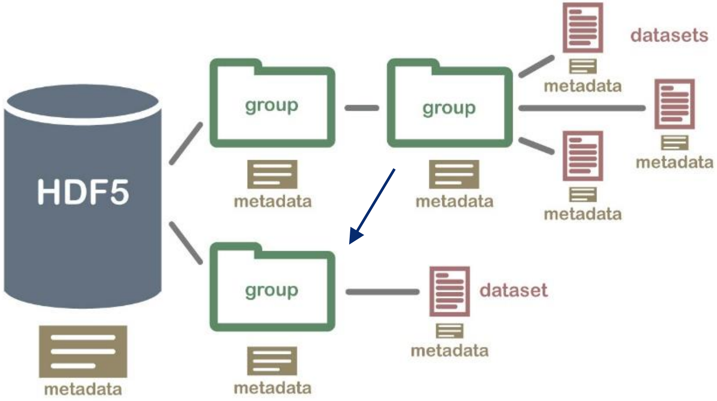
user (@NXuser)
Details of a user, i.e., name, affiliation, email address, *etc*

NXsubentry
Data or links to data for particular analysis

NeXus structure allows links and pointing to data stored in other parts of the group

HDF5 Format

- HDF5 format and a tree structure for metadata representative of the complexity of PaN data
- Allows chunked storage and slices reading
- Metadata can be attached
- The I/O can be faster than contiguous data files
- Compression
- Can be prefixed or open database size
- Heterogeneous database with links
- Platform independent
- Suitable for massive databases with a datatype and dataspace definition per dataset



HDF5 structure allows links and pointing to data stored in other parts of the group

[8]

NeXus Writer and Icat data ingestion

- Ingestion workflow
- Access to the tools by virtual machine

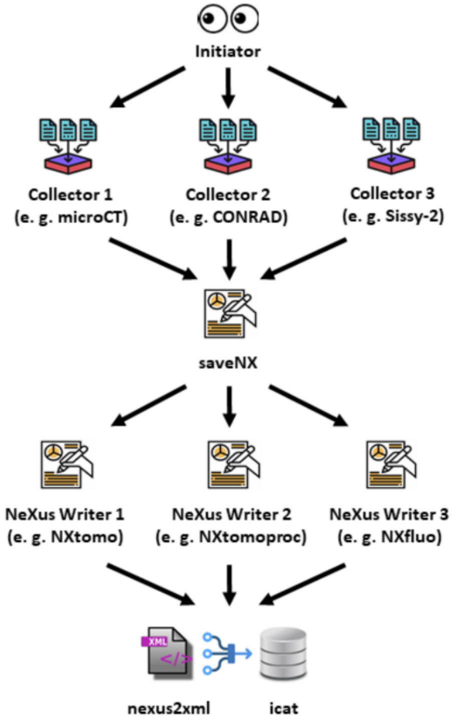
- 1. data collection
- 2. identification of the instrument dictionary
- 3. sample data info collection
- 4. parameters attribution
- 5. local data saving
- 6. icat repository ingestion

Contact person: Gerrit Günther, HMC

https://gitlab.helmholtz-berlin.de/jaf/nexuswriter/-/blob/master/nexusCore/icatRepo/nexus2xml.py

Nexus application for describing XAS

https://github.com/nexusformat/definitions/blob/main/applications/NXxas.nxdl.xml


Initiator
Collector 1 (e. g. microCT)
Collector 2 (e. g. CONRAD)
Collector 3 (e. g. Sissy-2)
saveNX
NeXus Writer 1 (e. g. NXtomo)
NeXus Writer 2 (e. g. NXtomoproc)
NeXus Writer 3 (e. g. NXfluo)
XML
nexus2xml
icat

Initiator: The entry point that starts a collector. There are various initiators ranging from command-line interface to GUI.

Collector: An experiment specific module that collects data from different sources and assigns values to a python dictionary {…}

saveNX: A distributor that starts the appropriate NeXus Writer according to the NeXus definition schema of an entry; it may start different writer routines for a file.

NeXus Writer: A module that reads the python dictionary {…} and writes its content to the NeXus file according to a specific NeXus schema.

nexus2xml: An interface to HZB's icat to satisfy its demands: structures the produced files in folders, reads their content and writes a summary of searchable terms to a xml file before starting the ingest.

Data analysis with Nexus files

- https://manual.nexusformat.org/utilities.html#data-analysis

- A number of Python routines to process X-ray photons emission data in hdf5
- XRF spectroscopy PyMCA
- https://gitlab.elettra.eu/panosc/xrffitvis/
- Xrayutilities for conversions spec hdf5

- IGOR pro can upload HDF5 www.wavemetric.com
- ORIGIN lab (+HDF5Browser App)
- DAWN
- Matlab
- Spec2hdf5 available tools (silx.org)
- Spec2nexus (https://spec2nexus.readthedocs.io)
- PyMCA (http://pymca.sourceforge.net/)
- NeXpy (http://nexpy.github.io/nexpy/)
- Mantid (Mantid Project — MantidProject landing page documentation)

HELMHOLTZ
METADATA
COLLABORATION

NeXus structure and playground

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HELMHOLTZ
METADATA
COLLABORATION

05 Resource Identification PIDs

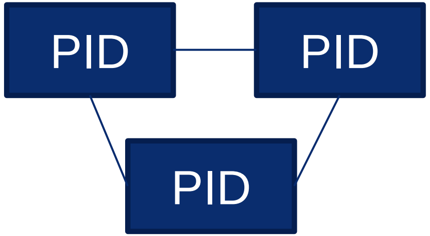
www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

PID

- **Persisten identifiers** are **long lasting unique** digital reference to an object, contributor and organization.
- They act as pointers/entry points between the reference to the object and its actual location
- They are character strings globally unique

Labeling a dataset i.e. with a PID support the data findability: independent on the data physical location



[19] V. Bunakov, R. Krahl, B. Matthews, N. Vizcaino, A. Vukolov(2022) Advanced infrastructure for PIDs in Photon and Neutron Ris. ExPaNDS Deliverable 2.5. https://doi.org/10.5281/zenodo.5905351

PID

It is managed and updated with new location in the registry

and resolvable

PID attribution (ARK, arXiv, DOI, ePIC, Handle, URN) to people and digital object.

Object for **Instruments**, **Institutes**, **research products** (workflows, labs procedures, software) and **samples**

Different protocols are used for the implementation of PID, a global handling system of the different services guarantee that a PID generated at national agencies is valide worldwide

[19]

ORCID

- **ORCID (Open Researcher and Contributor ID)** registry for researchers. Provides an ID to associate publications, grants and employments to individuals and desambiguate.
- It consist of a 16-digit alphanumerical code
- It acts as record of professional activities. Journals requires that at least the corresponding authors has an ORCID

  https://orcid.org/

[20]

DOI

- **DOI** is built on the top of the Handle system and maintains a registry of metadata related to the object, It has local Registration agencies (DataCite for **datasets**, CrossRef for **publications** (chapters, proceedings, EIDR for audio objects). It is widely used in the publication registry and has specific metadata schemata, business model and assignment practices
- Associate the ID number to the URL hosting the data

[20]

ROR & grant ID

- **ROR research organization** Registry. It stores metadata about the organization hosting the project and the project framework of the research object.
- **GRID Global Research Identifier Database (www.grid.ac)**

- **Crossref Funder Registry (crossref.org)** enables the link between grants, projects and research products.
- Funding agency identifier are in the form of a DOI and listed in the Funder Registry (e.g. 10.13039/501100001659 for DFG)
- This tool facilitates the visualization of the projects results and the collections of products into “researchers portfolio” ready to be used e.g. when submitting new proposals to funding agencies.

[20]

Notes: corrections are needed in the PID slides from Expands

PID for Samples

- IGSN **research organization** Registry.
- It stores metadata about the organization
- Guaranteed to be unique with Handle identifier (handle.net)
- It has a sample metadata profile (alligned to datasite) at federated IGSN
- IGSN has a landing page with sample descritpion accessible by QR code
- Mandatory metadata are bibliographic but metadata structure include domain specific elements
- It is possible to link samples to data and publication
- ~10 millions samples registered worldwide

- Logs of all the operations (sample preparation) on the sample embed the sample history.
- This log file is preserved and attached to the metadata
- Potential assignement of a PID ?

Considerations : sample charactr are changing while treated, measurements and sample creation can be really distinguished? What is the sample preservation after the measurements ?

[20]

Making research workflow FAIR:
03 **Repositories** and **Licenses**

How to publish database and software ?
How to identify a suitable repository

- **Offer permanent archives and access to the data, provide persistent Identifiers (FAIR)**
- **Look for an institutional repository, project repository**
- **Generalist/ community/institutional repository:** zenodo, figshare, eudat, dryad home/ CXIDB, EMPIAR, EMDataResource, NOMAD (https://nomad-lab.eu/services/repo-arch)
- **FAIRsharing** registry of repositories https://fairsharing.org/ structured overview
- RatSWD https://www.konsortswd.de/datenzentren/alle-datenzentren/

  Collection of german data centers
- Re3data.org international registry of research data repository

https://www.re3data.org/search?query=&subjects%5B%5D=30701%20Experimental%20Condensed%20Matter%20Physics

- Measure of trustworthy of repository given (e.g. certificate, PIDs, metadata)

License

- **Terms of use of your data**
- **Data proper attribution**
- **Data and databases for the licence**

https://eu-datenschutz.org/

Open Data Commons group

3 standard licenses and additional community norms

PDDL: Public domain dedication

ODC-By: free user under the provision of referencing to the source

ODC-Odbl: any use of the database must refer to the source, any derivative product is to be distributed under same terms

https://creativecommons.org/licenses/

Creative Commons

CC-BY-SA attribution share alike and same licencing of related work

CC-BY- Attribution only

CC0 : Public domain dedication

PDM: Public Domain mark – free of know restrictions

[23]

Terms of use of your data

- **HZB data policy**

https://www.helmholtz-berlin.de/pubbin/vademecumdatei?did=326

https://www.helmholtz-berlin.de/pubbin/vademecumdatei?did=131

- Raw data :

  The raw data is under embargo for 5 years and hosted for a least 10 ys at HZB

  The embargo can be extended under request and can be shortened according to project needs

  The raw data are licences with CC0

  Results data are not affected

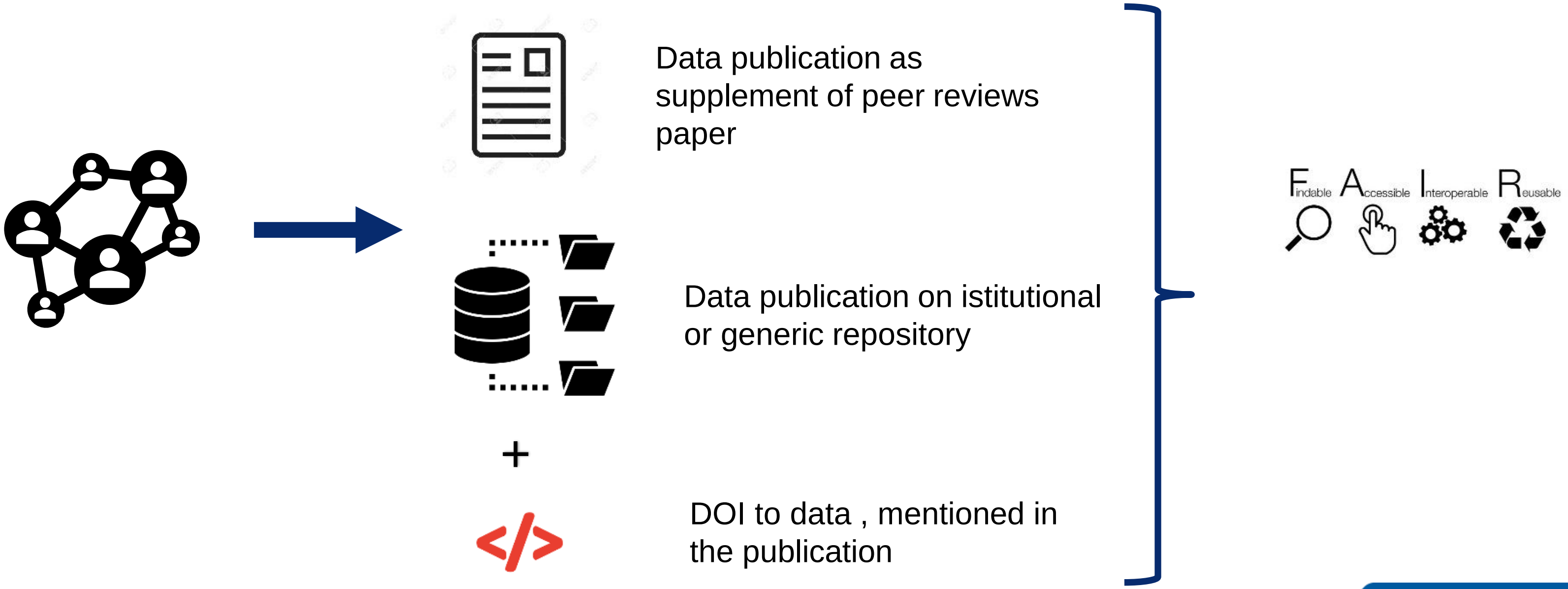
paN-data Europe common policy framework (february 2011) ---→ extended to FAIR by Pansoc

- Open access to raw data and metadata
- Curation of raw data supported by the facility
- Data catalogue to make data accessible
- Embargo on the raw data

[24]

Data publication guidelines

- Check the repository guidelines
- Check the publication service guidelines
- Check previous slides on data curation and metadata for publication



Data publication as supplement of peer reviews paper

Data publication on istitutional or generic repository

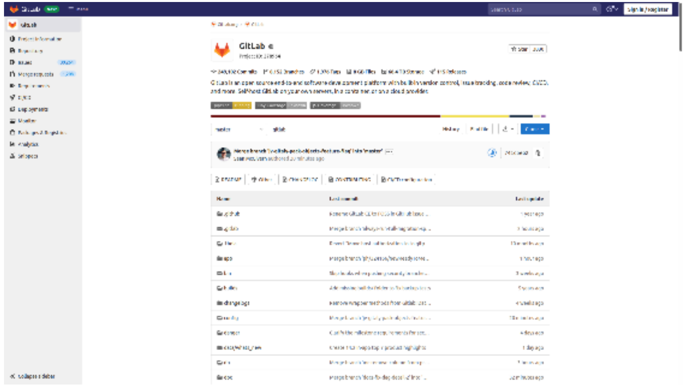
DOI to data , mentioned in the publication

Software publication guidelines

- Prepare a README file, usage description and dep.
- Provide usage examples and test integration
- Prepare a MAKE file or a setup.py file to prepare the repository clone or the installation with all dependencies included
- Associate a usage licence
- Enable the installation on OSs and give the compatibilities with the dependencies versions
- Obtain a DOI for the software e.g. DataCite and a snapshot of the gitlab
- Versioning the software
- Offer help support and possibility to start the bug fixes
- Use when possible the Gitlab /GIThub with institutional maintenance to guarantee the long term access to the software repository

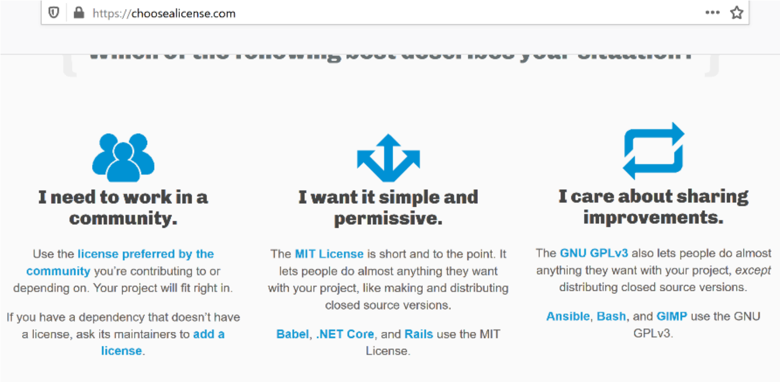
GitLab

GitHub


GitLab

https://github.com/

They both offer continuous integration and delivery and support the licences of your software


https://choosealicense.com

Which of the following best describes your situation?

I need to work in a community.

Use the license preferred by the community you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to add a license.

I want it simple and permissive.

The MIT License is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, .NET Core, and Rails use the MIT License.

I care about sharing improvements.

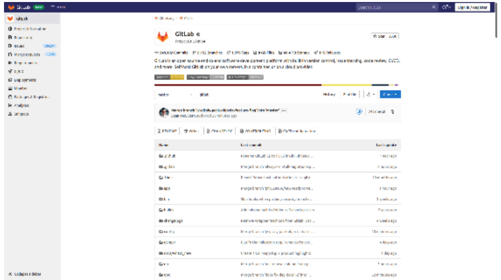
The GNU GPLv3 also lets people do almost anything they want with your project, *except* distributing closed source versions.

Ansible, Bash, and GIMP use the GNU GPLv3.

[26]

Software publication guidelines

- No widely accepted FAIR principles as for research data (status 2019)
- Software citation principles force11.org (https://doi.org/1025490/a97f-egyk)
- DataCite for software citation FORCE11.org principles mapped



GitLab

GitHub

https://github.com/

MIT Licence, Apache 2.0

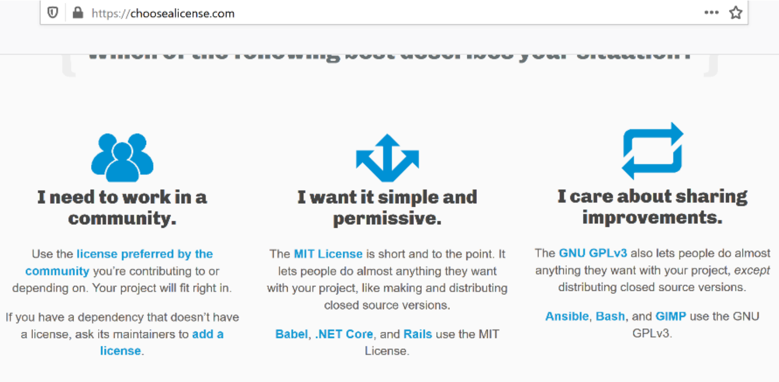
A short and simple permissive license requiring preservation of copyright and license notices.

https://github.com/G-Node/gogs/blob/master/LICENSE

Indication of modification respect the original version

GNU General Public License v3.0 for free software www.gnu.org

They both offer continuous integration and delivery and support the licences of your software



https://choosealicense.com

Which of the following best describes your situation?

I need to work in a community.

Use the license preferred by the community you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to add a license.

I want it simple and permissive.

The MIT License is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, .NET Core, and Rails use the MIT License.

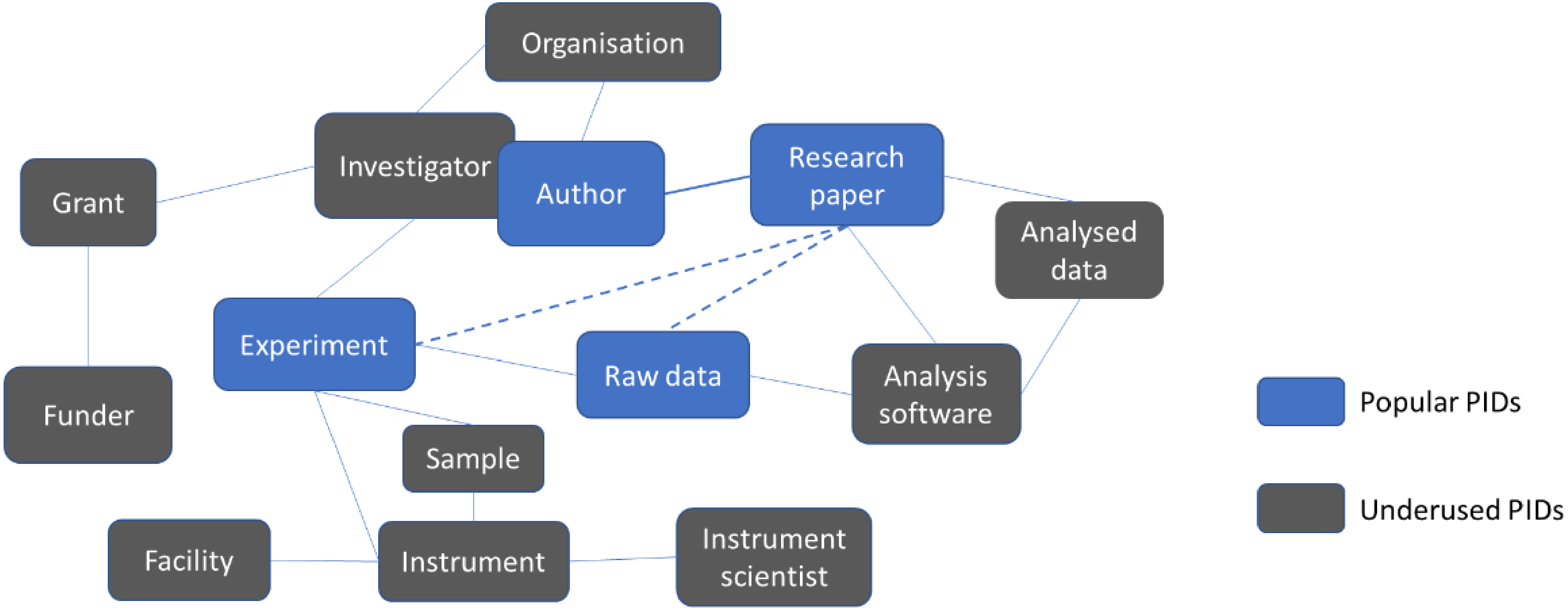
I care about sharing improvements.

The GNU GPLv3 also lets people do almost anything they want with your project, *except* distributing closed source versions.

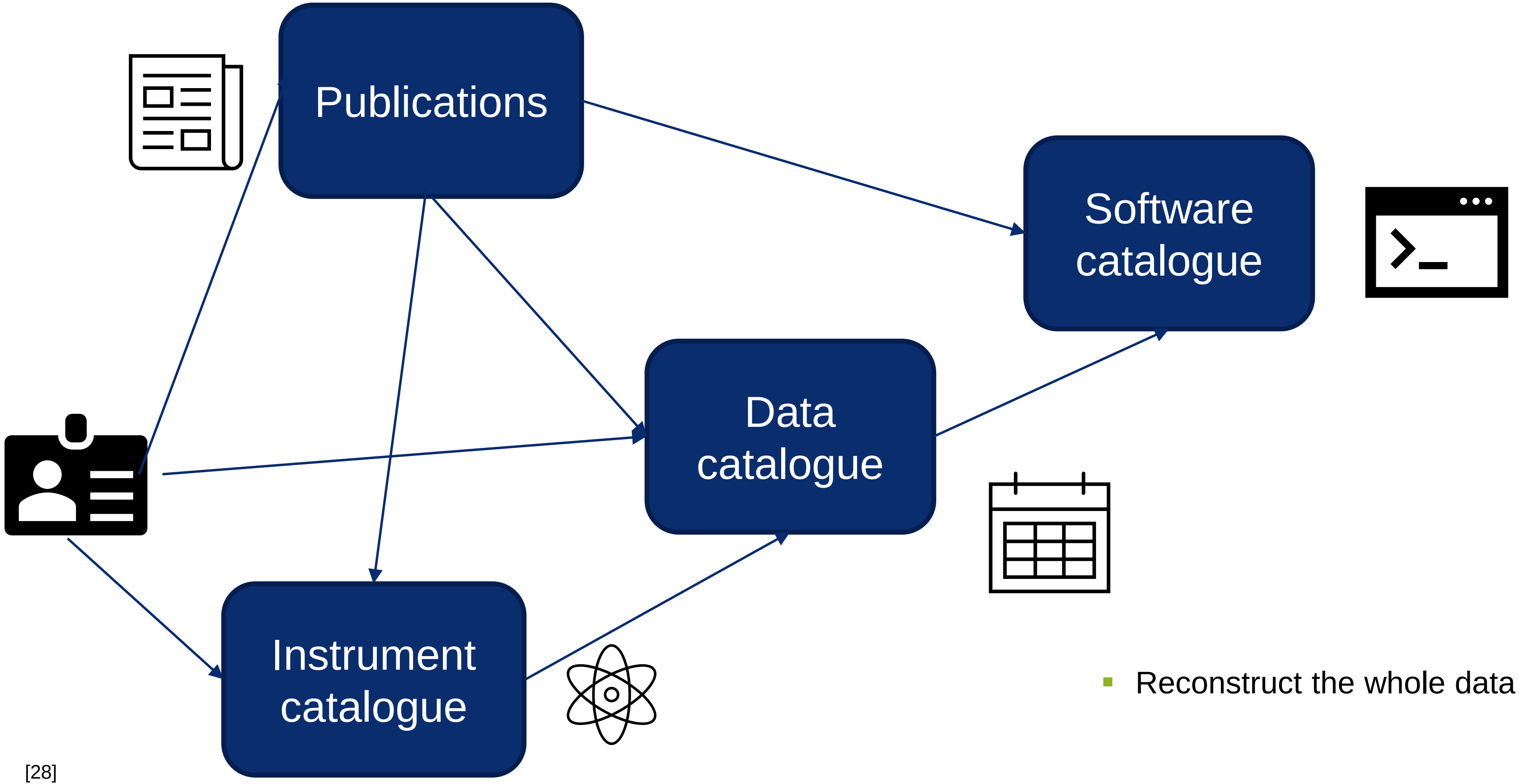
Ansible, Bash, and GIMP use the GNU GPLv3.

[27]

From expands report (2022)



Link the different resources: enables discovery and interoperability



- Reconstruct the whole data life cycle

[28]

00 Recap: **FAIR Principles and PaN data life cycle**

Why FAIR

- FAIR is not Open Access
- Different solutions and different paces of implementation are suggested for the different communities
- **Findable**

  Persistent ID

  Metadata online

  Online Repository
- **Accessible**

  Community or generic repository
- **Interoperable**

  Community or generic standards

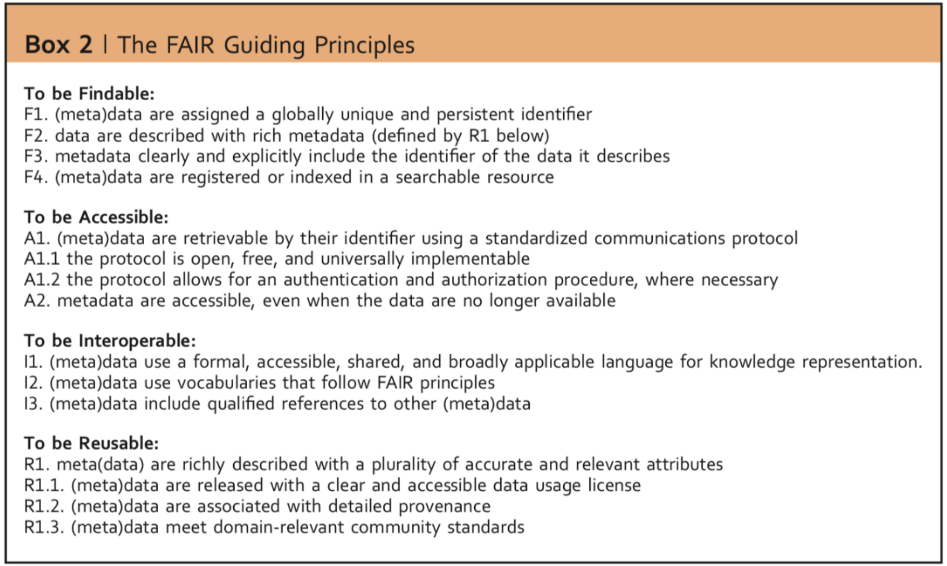
  Open file formats
- **Reusable**

  Rich, complete and clear documentation

  Licenced

Guidelines for promoting data visibility, reuse and ensure data quality

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

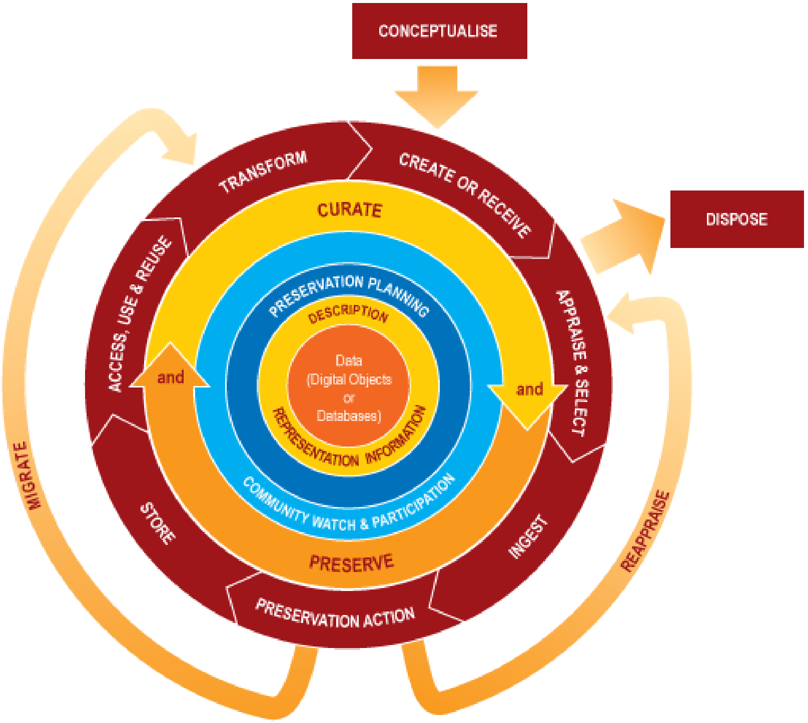
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Research data life cycle

Proposal submission and scheduling	Proposal ID and measurement descriptions
Data creation and processing	Design DMP, define derivative data, authoring
Data staging	Identify repository Assign metadata and data format
Data publishing	Establish data access conditions Associate PID and copyright and versioning
Data reuse	referencing to related publications and versions



Lib.ua.edu

Recap on How prepare data to FAIRness

- Identifier to the data, repository, metadata, licences
- Standard or not proprietary formats
- Keep data documented and versioned
- Project or community repository
- Check the publication licences
- Cross references to related data

HELMHOLTZ
METADATA
COLLABORATION

luigia.cristiano@helmholtz-berlin.de

And the Hub Matter Team

Thanks.

HMC Hub Matter

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES