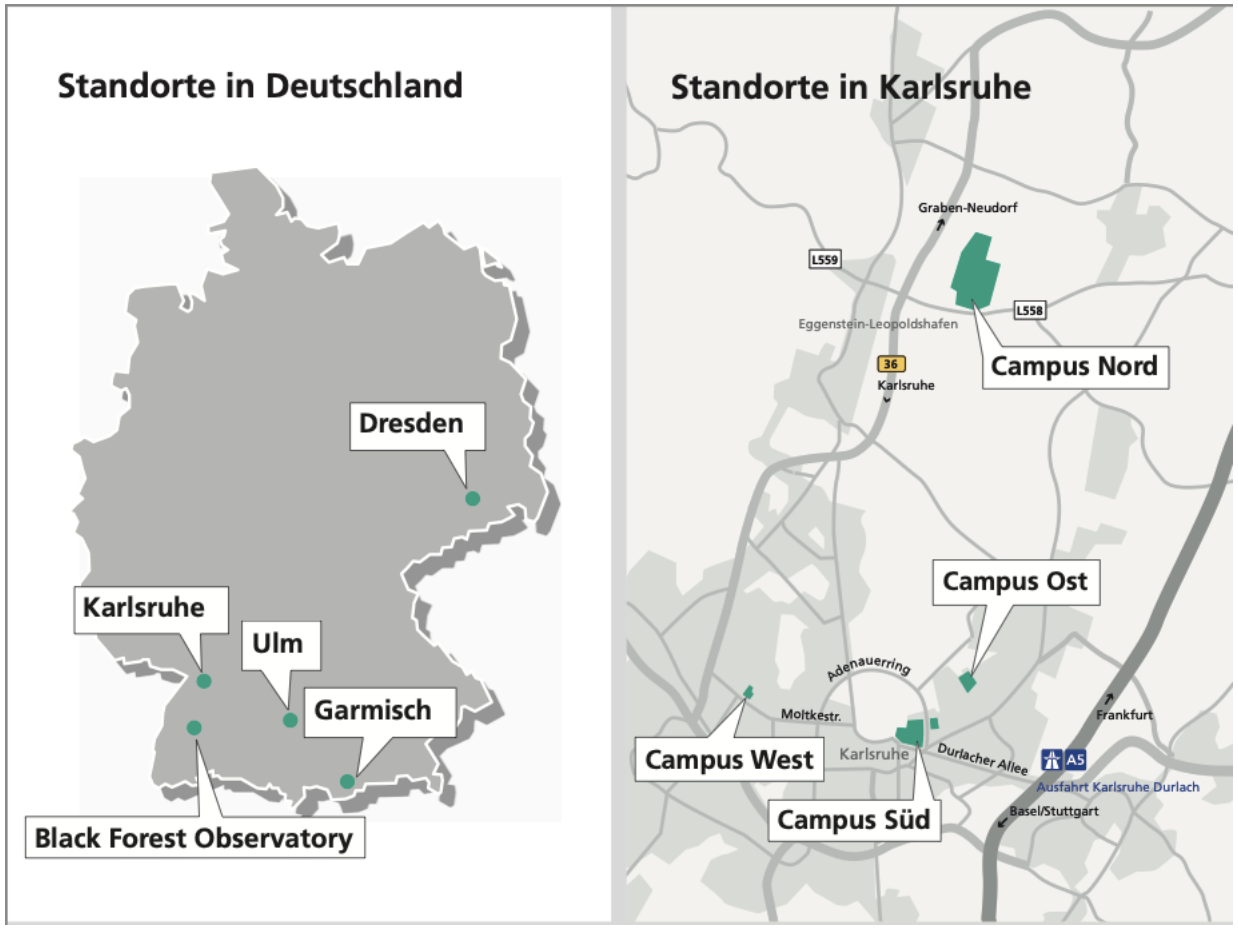
Research Data Management at Karlsruhe Institute of Technology (KIT)

Dr. Rossella Aversa

OAVdA, 05.06.2026

Karlsruhe Institute of Technology (KIT)

Scientific Computing Center



Karlsruhe Institute of Technology (KIT)

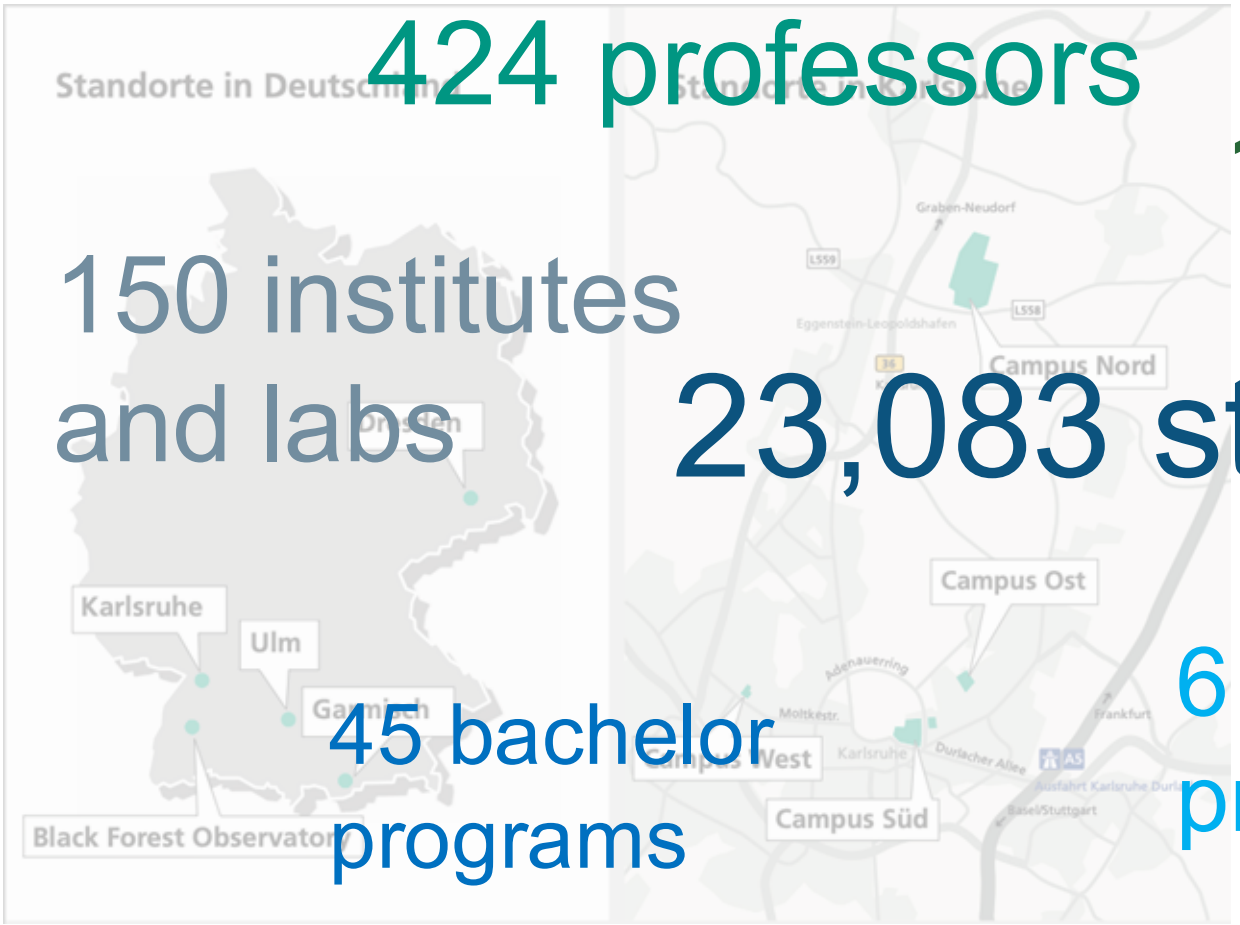
SCC Scientific Computing Center

424 professors

10,131 employees

150 institutes and labs

23,083 students



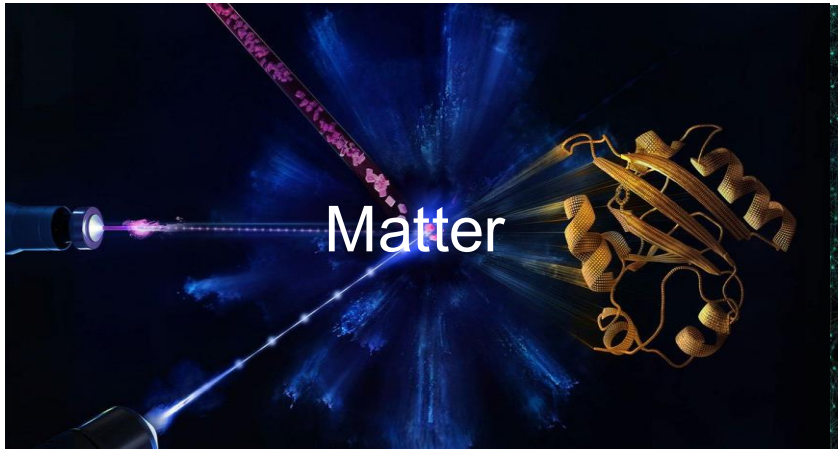
45 bachelor programs

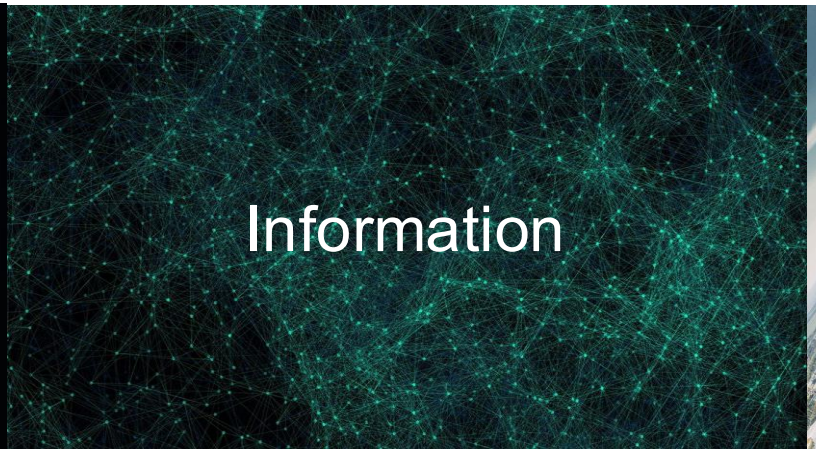
61 master programs

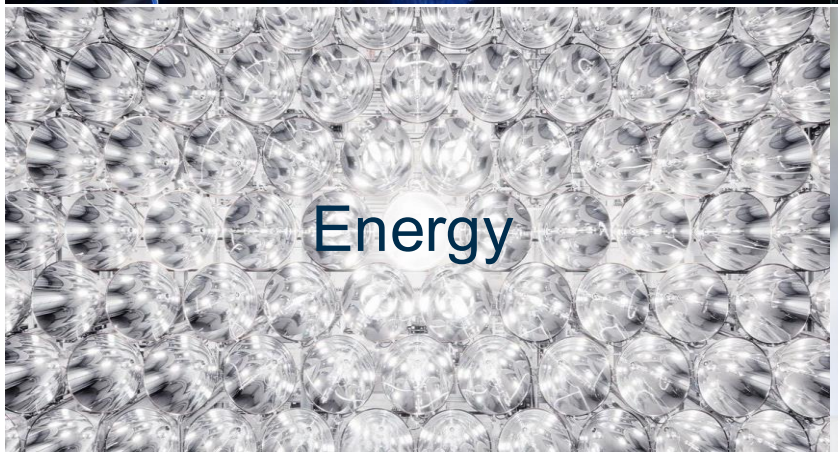
Source: KIT website

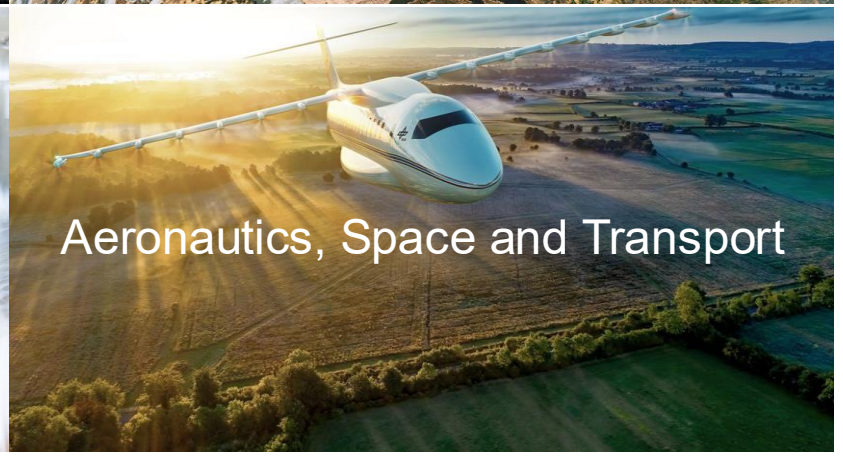
KIT

Helmholtz Association





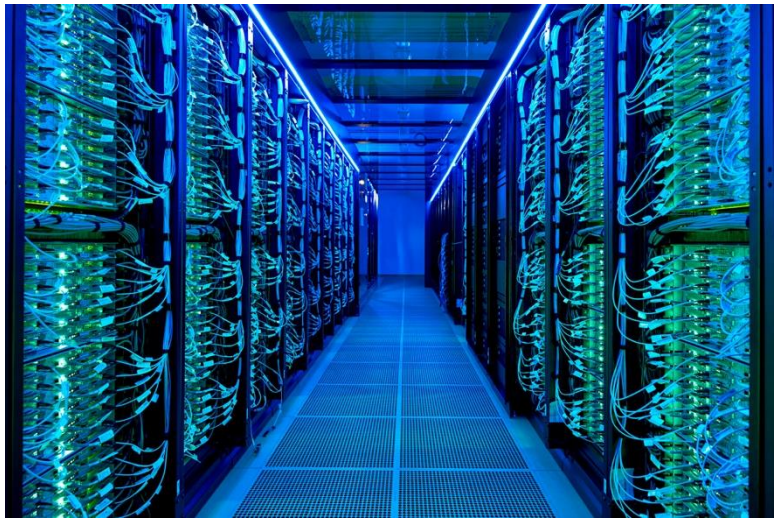




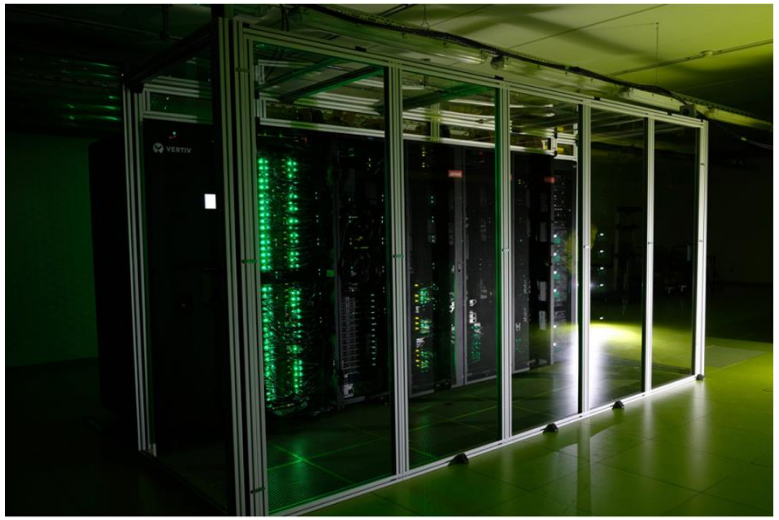
Source: Helmholtz Association website

KIT: Scientific Computing Center (SCC)

Scientific Computing Center



GridKa

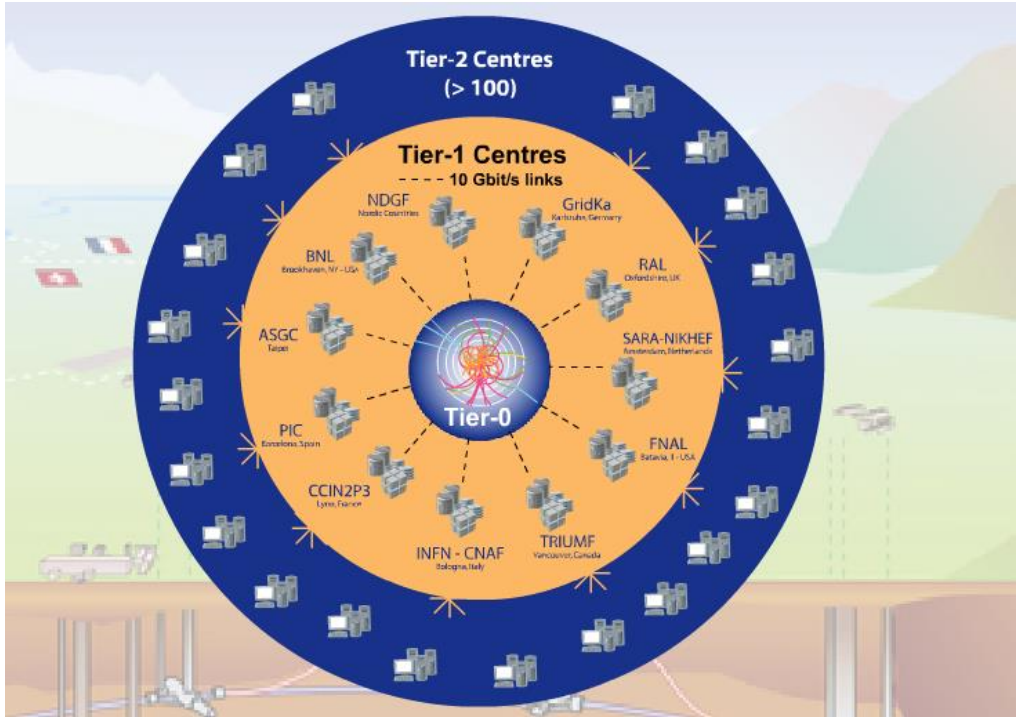


bwUniCluster

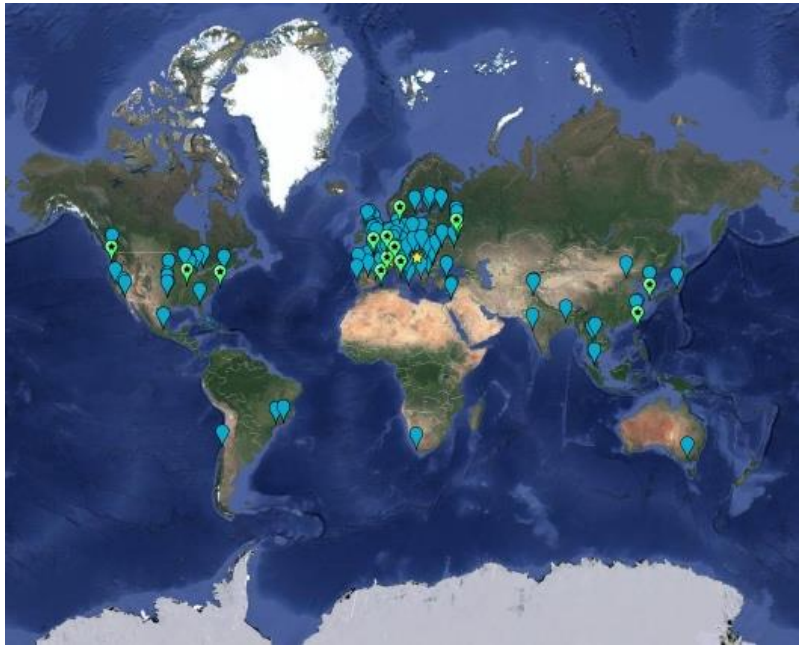
Source: SCC website

KIT involvement: Worldwide LHC Computing Grid (WLCG)

- Infrastructure for LHC data storage and analysis
- More than 170 institutions and 34 countries



Courtesy of Bob Jones (CERN)



- **Tier-0** CERN
- **Tier-1** ~ 13 sites
- **Tier-2** ~ 150 sites

GridKa

KIT involvement: Worldwide LHC Computing Grid (WLCG)

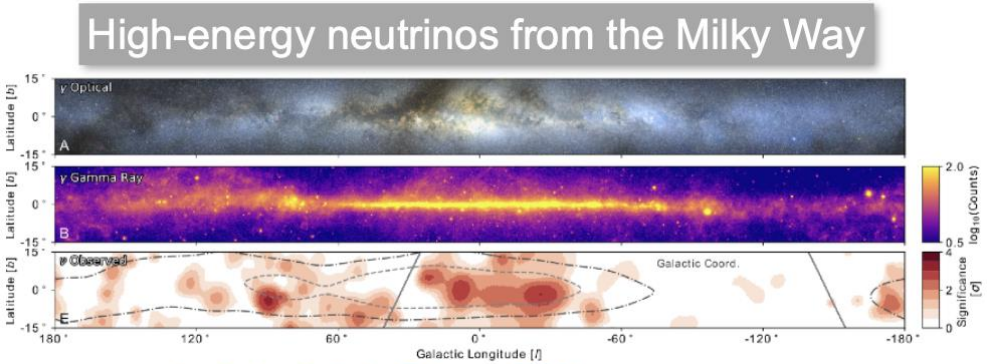
- Solar panels installed on GridKa data center and office buildings
- Unique in WLCG
- 500 kWp → GridKa could be powered 100% by solar energy on a sunny day

Courtesy of Achim Streit (KIT-SCC)

KIT involvement: IceCube

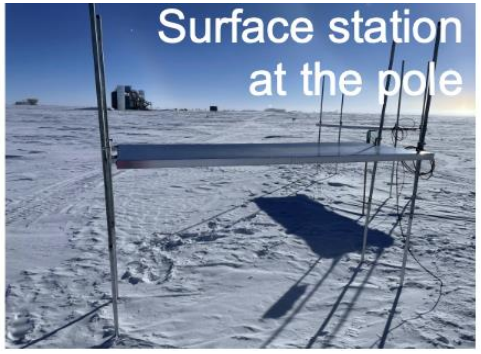
MATTER AND THE UNIVERSE
Matter and Radiation from the Universe

- Multi-purpose neutrino observatory at South Pole
- Indirect study of dark matter
- Cosmic ray physics
- Validation of hadronic interaction models



IceCube Collab., Science 2023

- Instrumentation (hardware, construction, installation)
- Software
- Data Analysis
- HPC Tier-1



Courtesy of Markus Roth (KIT) and Kathrin Valerius (KIT)

KIT involvement: high-energy universe

MATTER AND THE UNIVERSE
Matter and Radiation from the Universe

Einstein Telescope

Gravitational waves

KATRIN @ KIT

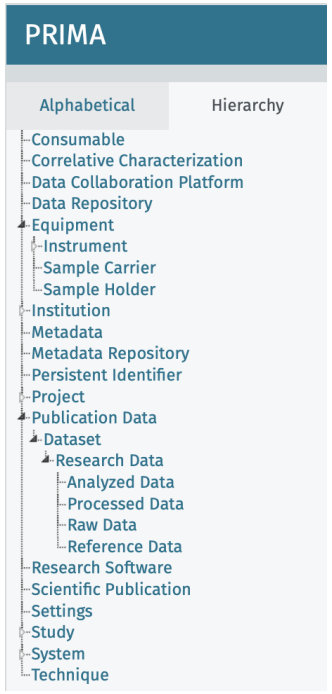
Neutrino mass

Courtesy of Andreas Haungs (KIT) and Kathrin Valerius (KIT)

Data Exploitation Methods (DEM)

Scientific Computing Center

Semantics



PRIMA Thesaurus

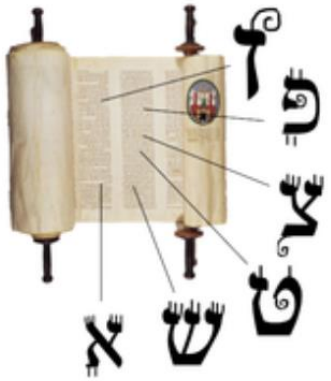
Schemas

ML data basic schema

Graphs

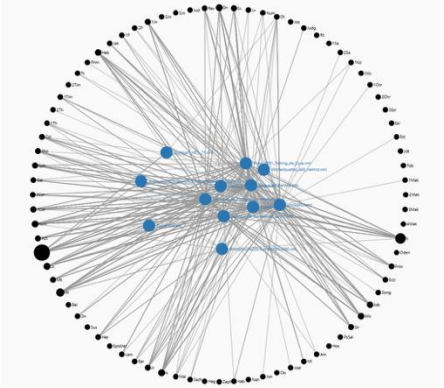
NovaCrate

Image segmentation



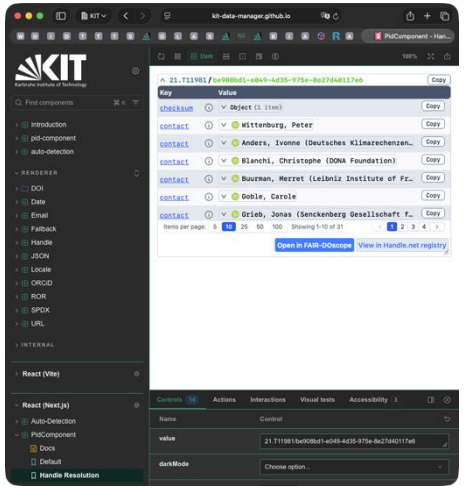
ToRoll

Data Analysis



Preliminary

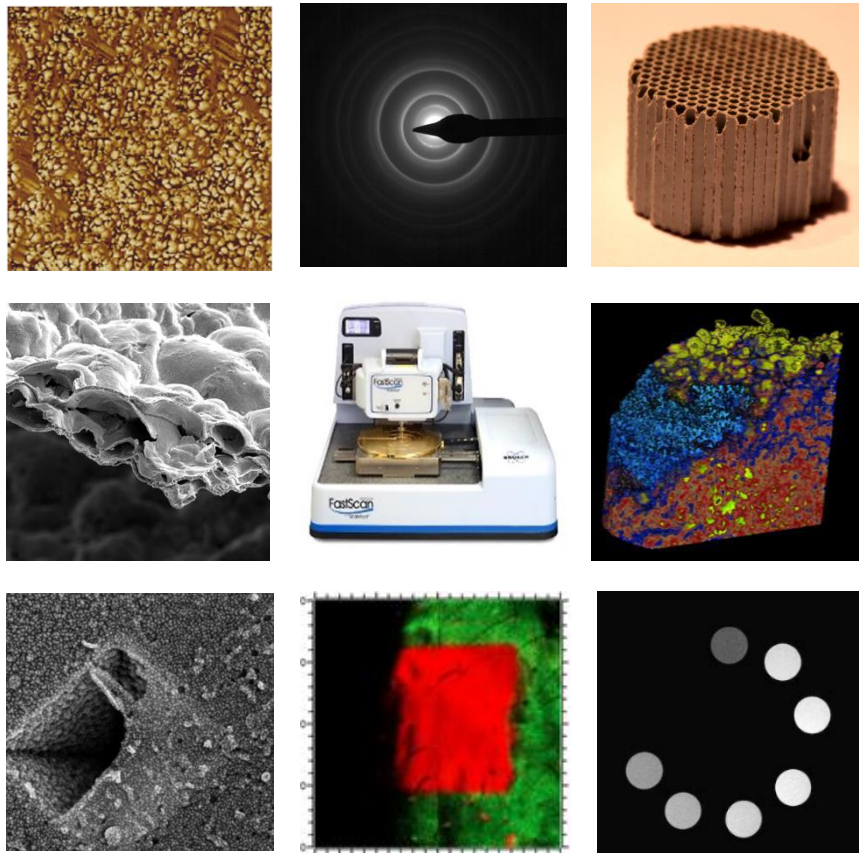
RDM software



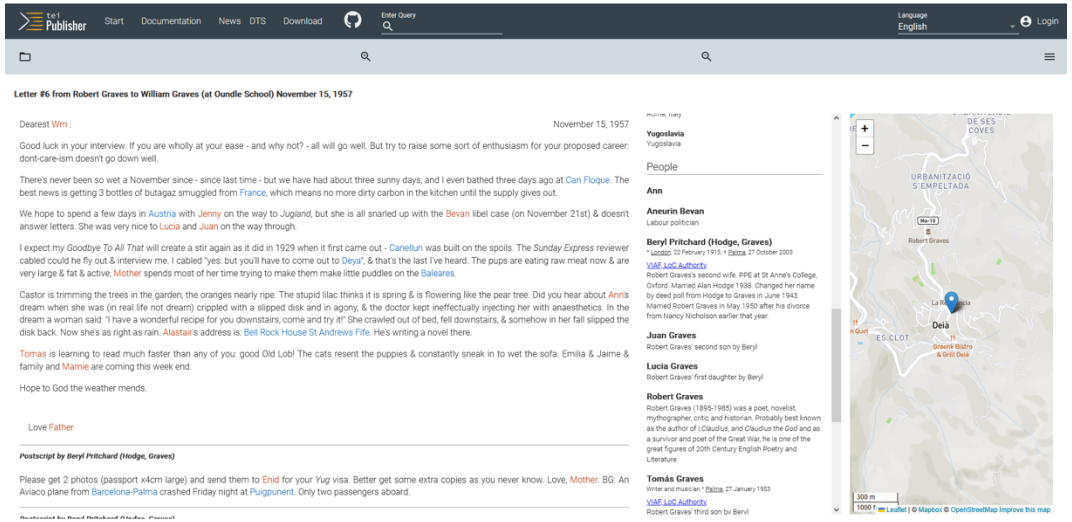
Typed PID Maker

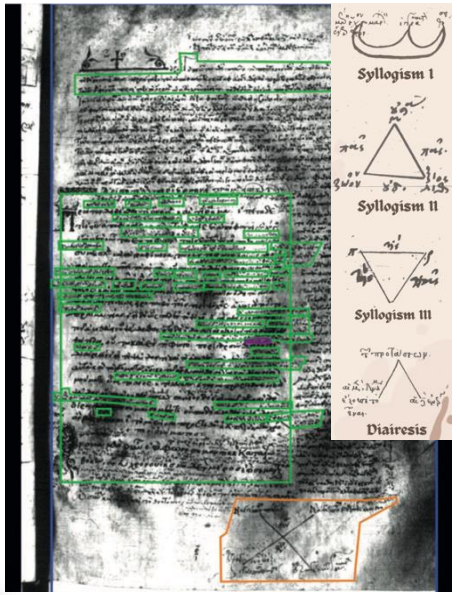
Current DEM Applications

Materials Science and Nanoscience

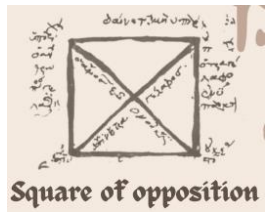


Digital Humanities





Aristotele, *De interpretatione*

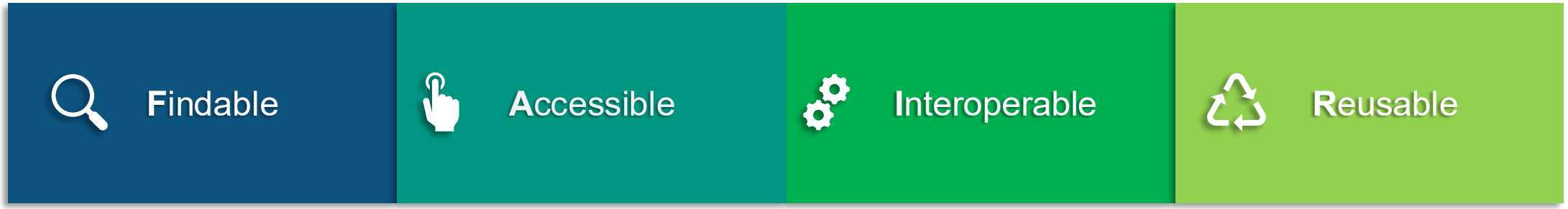


Courtesy of Richard Thelen (KIT), Neil MacKinnon(KIT), Regina Ciancio (CNR-IOM), Germaine Götzelmann (KIT), Danah Tonne (KIT), Lilly Osburg (KIT)

FAIR Digital Objects

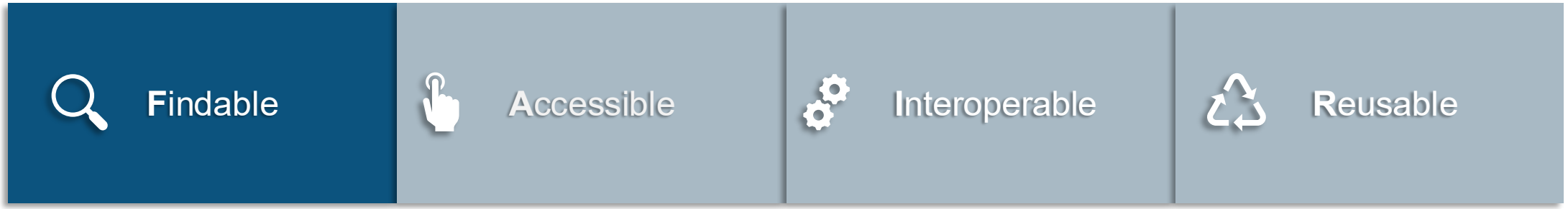
Digital Object: any identifiable piece of digital information that can be stored, managed, accessed, and exchanged by computer systems. It may represent data, software, metadata, a document, an image, a workflow, or another digital resource.

The FAIR Guiding Principles



https://www.go-fair.org/fair-principles/

The FAIR Guiding Principles

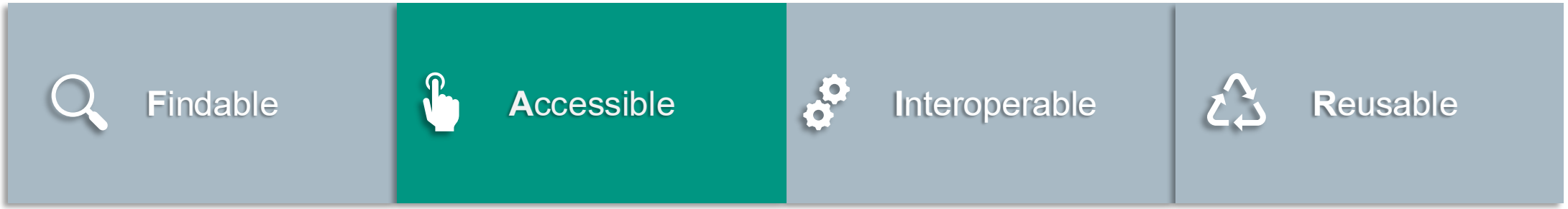


(Meta)data should be easy to find for both humans and computers

Globally unique persistent identifiers (PID)

https://www.go-fair.org/fair-principles/

The FAIR Guiding Principles

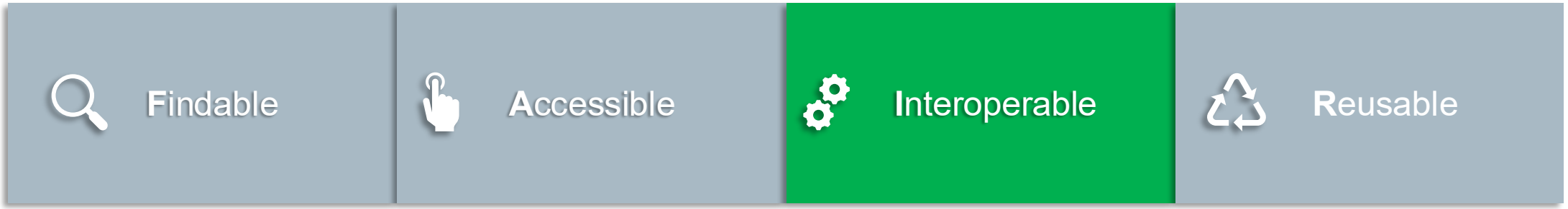


It should be known how (meta)data can be accessed

(Meta)data repositories, authorization & authentication

https://www.go-fair.org/fair-principles/

The FAIR Guiding Principles

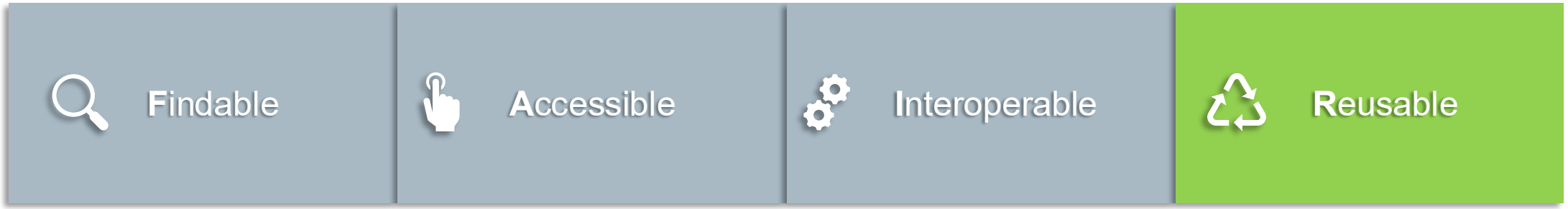


Data should be exchanged and interpreted by humans and computers

Structured metadata (schemas, vocabularies)

https://www.go-fair.org/fair-principles/

The FAIR Guiding Principles



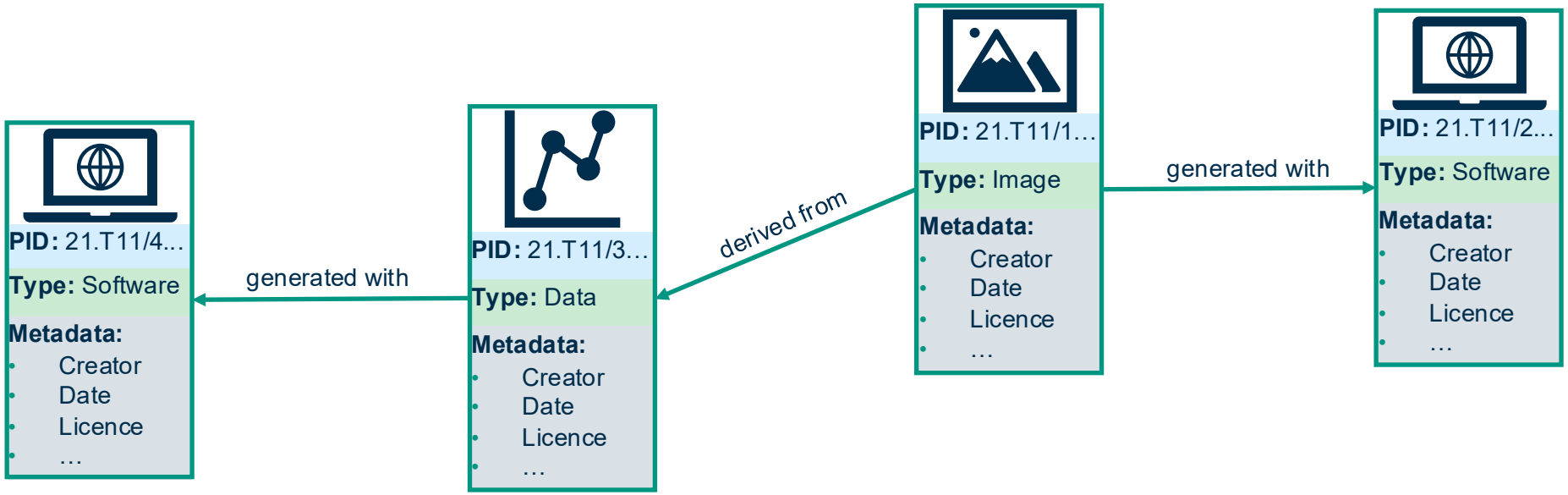
It should be clear how data can be reused and/or replicated

Licences, rich (provenance) metadata

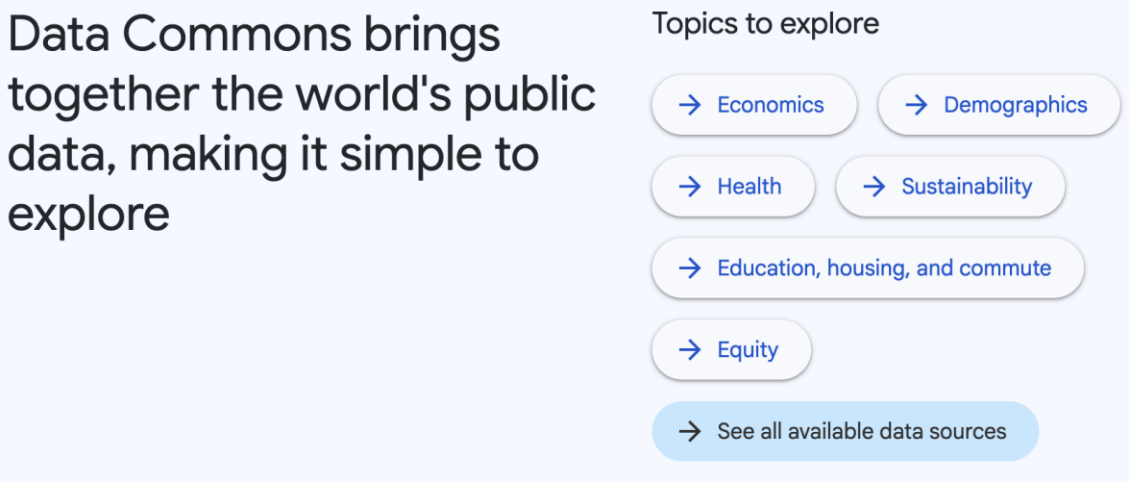
https://www.go-fair.org/fair-principles/

FAIR Digital Objects

FAIR Digital Object: digital object that is structured and described in a way that makes it FAIR. It contains the digital content and the essential machine-readable information.



Why are FDOs useful?



Data Commons synthesizes a single graph from these different data sources. It links **references** to the same entities (such as cities, counties, organizations, etc.) across different datasets to nodes on the graph, so that users can access data about a particular entity **aggregated from different sources without data cleaning** or joining.

Data Space: framework containing a set of data sources (and their relations) regardless of their format, location, or model.

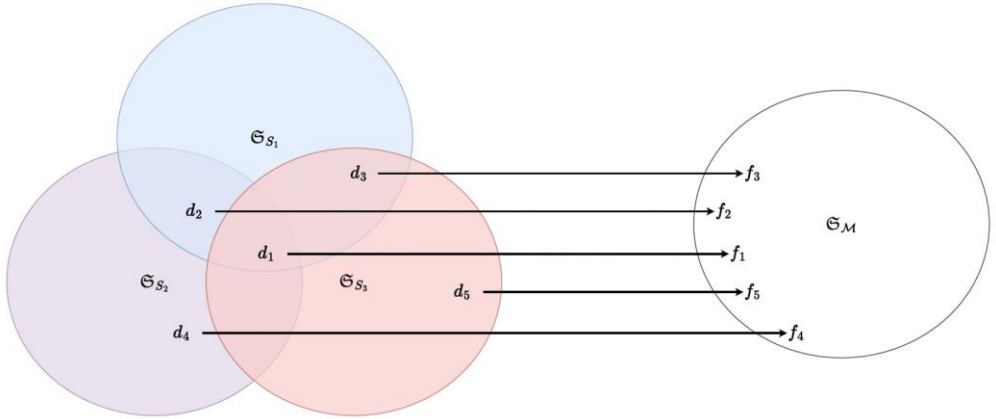
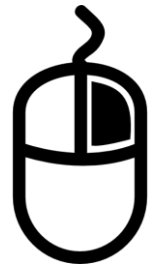


Figure 3.2.: Illustration of fully aligned FAIRified data spaces by projection to a meta-level.

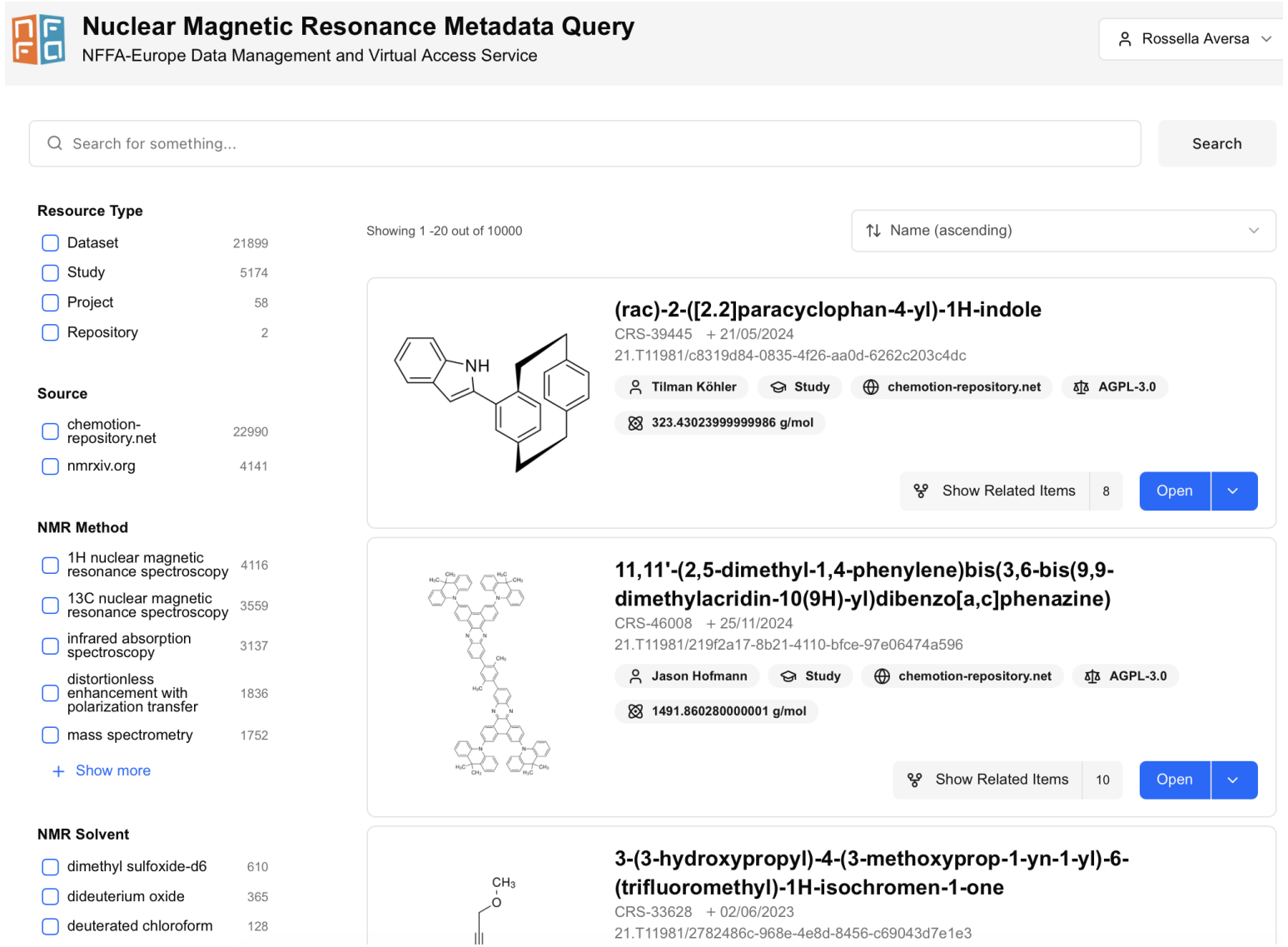
Courtesy of Nicolas Blumenröhr (KIT)

What are the advantages of FDOs?

- **Standardized descriptions:** harmonised representation, alignment of different data spaces, better data reusability/reproducibility, reduced duplication of effort
- **Standardized type definitions:** easier exchange of information across systems and organizations
- **PIDs:** persistently findable, reference (pointer) independent of storage location, no effect if location changes
- **Lightweight layer:** retrieve data from multiple repositories without touching or migrating the data
- **Customizable:** can include/link additional info which is not included in the specific repository
- **Machine-actionable:** machine-readable format, automated workflows/data pipelines, AI-ready without requiring manual intervention (at least quicker and more straightforward, no need of multiple APIs)
- **Operation FDOs:** set to act on specific target FDOs (according to their type)

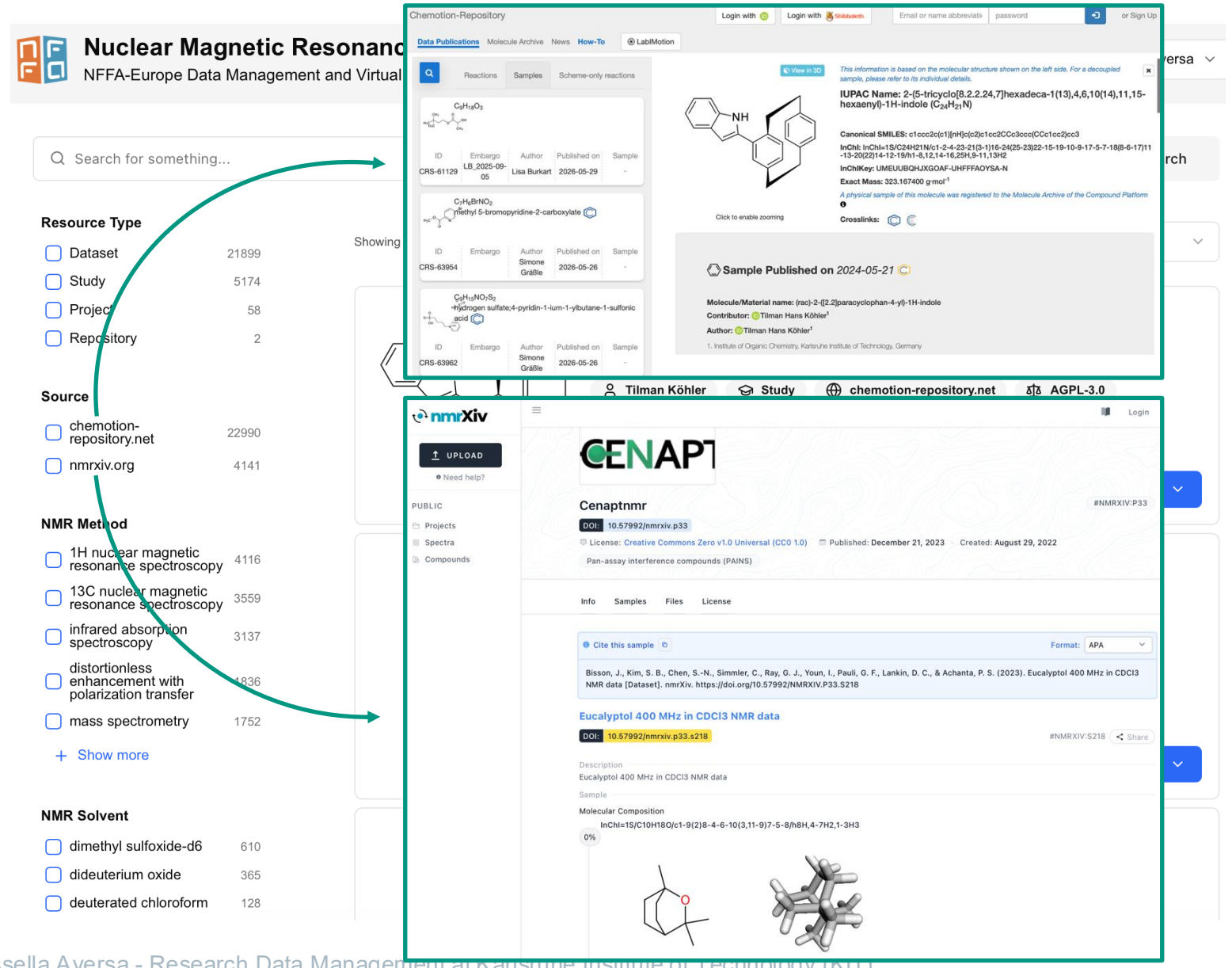


Example: NMR Search



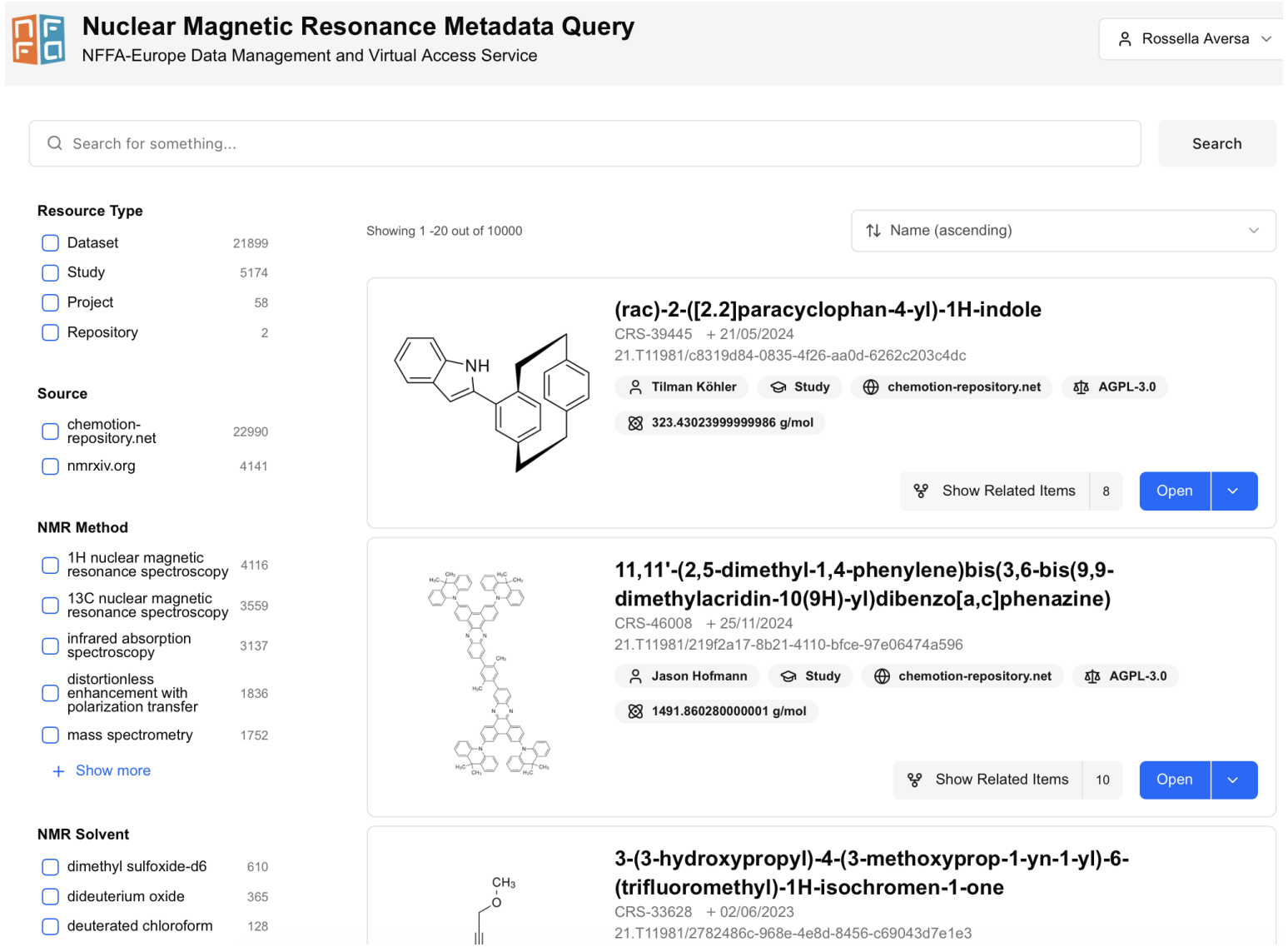
NMR Search

Example: NMR Search



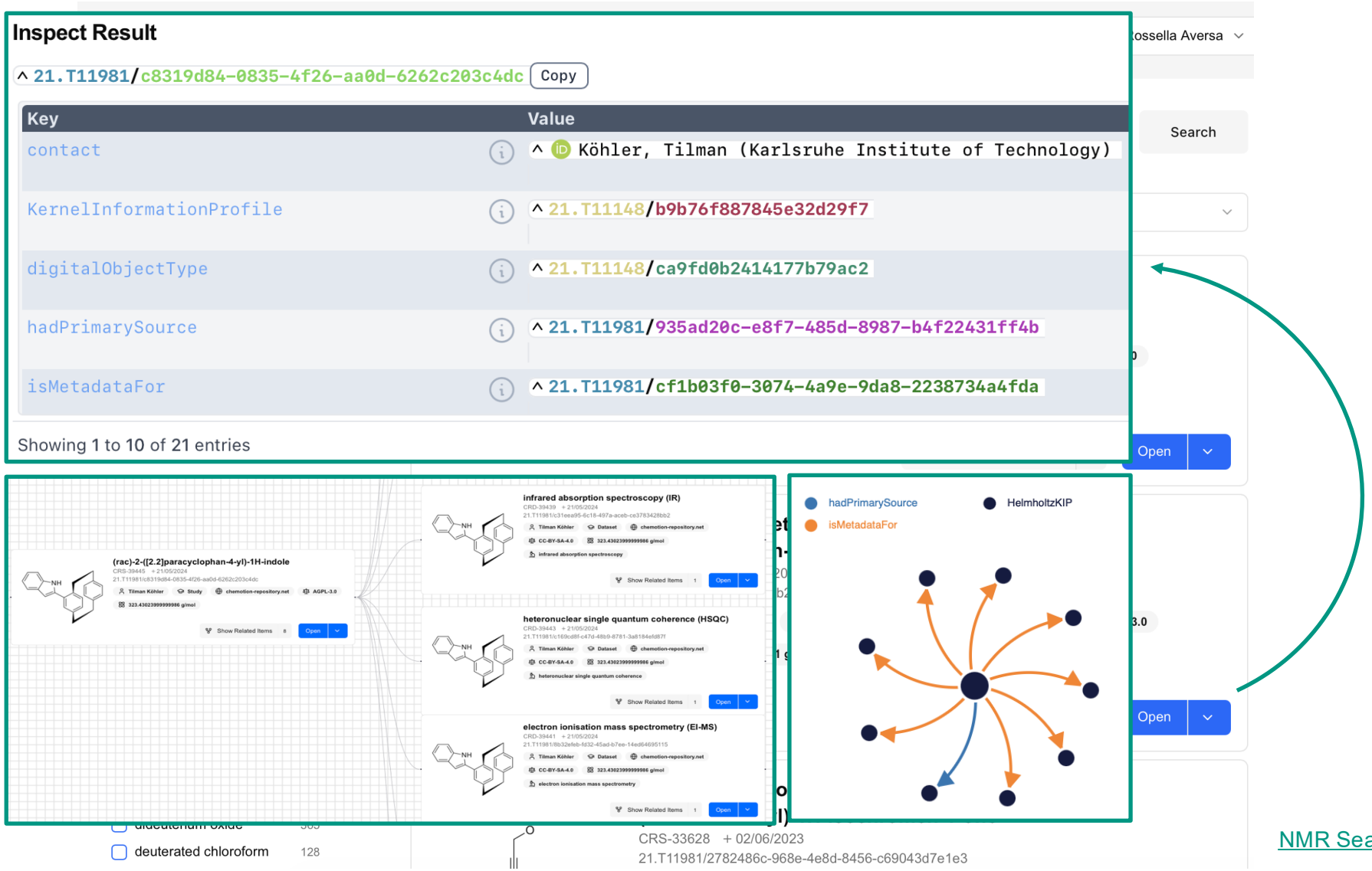
NMR Search

Example: NMR Search



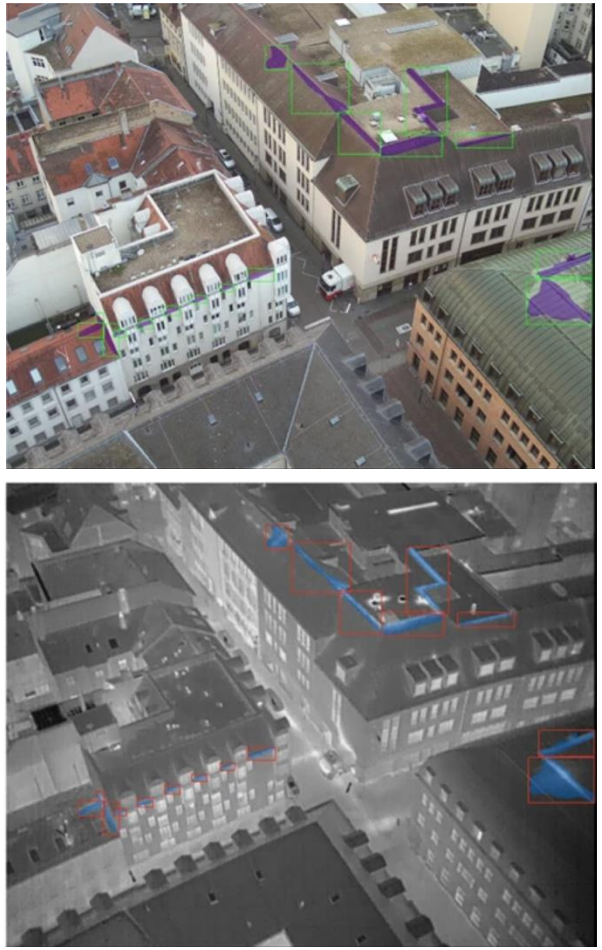
NMR Search

Example: NMR Search



NMR Search

Example: Thermal Bridges on Building Rooftops



PID Information Record

Type	Value
kernelInformationProfile	21.T11148/b9b76f887845e32d29f7
dateModified	2022-08-26T00:00:00+00:00
checksum	{ "sha512sum": "c51b5782d7f7b07607f7b8
dateCreated	2022-05-30T00:00:00+00:00
digitalObjectLocation	https://zenodo.org/record/6517768/files/Flu
version	1.0.0
digitalObjectType	21.11152/3052e4f1-4dd6-4c40-9a16-0a6b5
licenseUrl	https://creativecommons.org/licenses/by/4.0
hasMetadata	21.11152/ba370aa3-6422-428c-9ff7-c2ef4
isMetadataFor	21.11152/365fd8cf-8e86-41b8-9d0e-b816f
topic	http://vocabularies.unesco.org/thesaurus/cc
contact	https://orcid.org/0000-0002-9082-9095

FDO Badge

FAIR DO 21.11152/ebaeb5bc-0514-47c9-bcd2-98f0253843d8

Copy as

FAIR DO Graph

PID Component

Copy Code

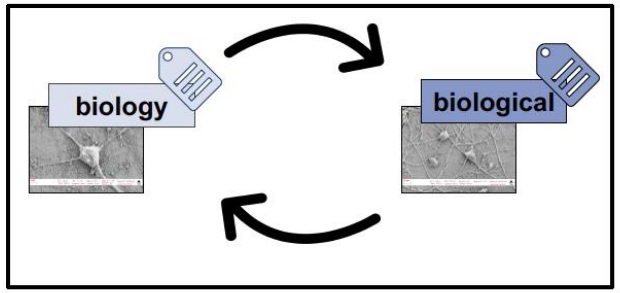
TBBR Dataset

Example: Relabeling of SEM training datasets

Data sets with different label terms are not directly reusable for Machine Learning (ML) applications.

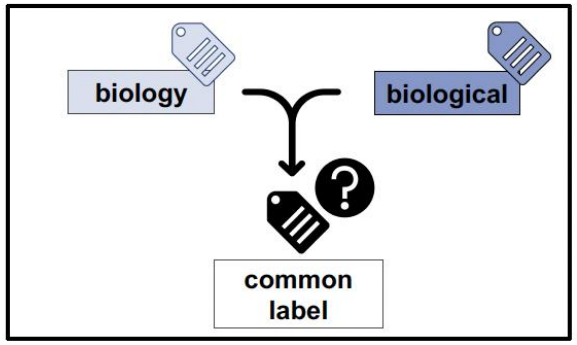
Automated relabelling is enabled by using a vocabulary-based schema and FAIR Digital Objects (FAIR DOs).

Usually different label terms are used to describe the same category, i.e. have the same meaning.

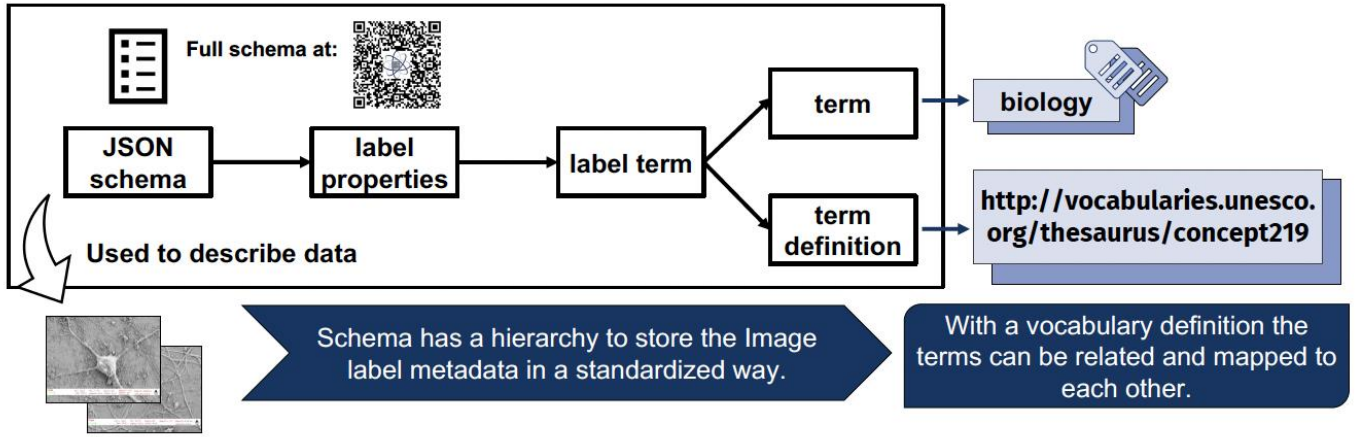


Example with *Scanning Electron Microscopy (SEM) image data. Provided by R. Aversa et. al. http://doi.org/10.23728/b2share.19cc2afd23e34b92b36a1dfd0113a89f

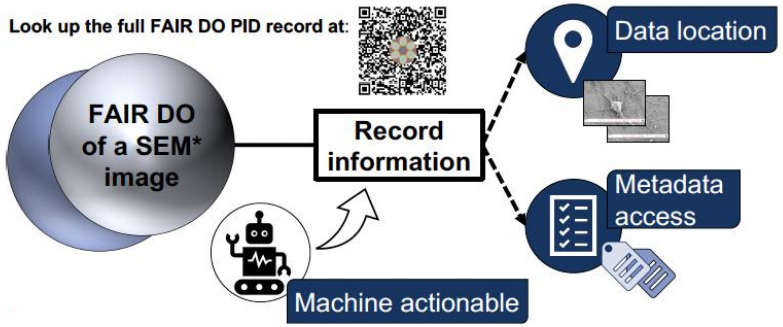
Labels can be mapped to each other, i.e. relabelled.



Metadata Schema for Label Information



FAIR DO for Machine Actionability



Conclusions

- Using a metadata schema and vocabularies for data description provides standardization.
- The record information of a FAIR DO contains all information required to access the data and its metadata.
- Machines can evaluate the information record and carry out the relabelling using that information.

Blumenröhr & Aversa 2022

FAIR Digital Objects

~ 37,000 FDOs

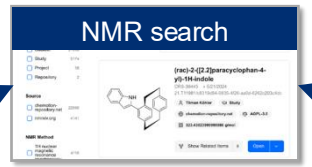
FAIR DIGITAL OBJECTS FORUM

RDA RESEARCH DATA ALLIANCE

FAIR Digital Objects

NEP **2023**

NMR search



NFDI4Chem

NMR data of eth. extract and fractions of swertia chirayita

21.T11981/5760c10e-6e64-41ea-824e-8dd4d3d2145d

Materials Science

Chemistry

Zenodo **2022**

Typed PID Maker

maintain
create
validate
PID

Zenodo **2022**

FAIR-DO Lab

Zenodo **2024**

PID-Component

GitHub **2024**

FAIR-DO Cookbook

HMC

GitHub **2023**

FAIR-DOscope

NFDI MatWerk

Sonderforschungsbereich 980 EPISTEME IN BEWEGUNG

HELMHOLTZ AI

Artificial Intelligence Cooperation Unit

Acta Mat. **2025**

BAM reference data

21.11152/fa8ce096-b98b-4ec9-8ca9-fa140454bb86

Materials Science

OM/ISCW. **2024**

DH benchmark

21.11152/a3f19b32-4550-40bb-9f69-b8ffd4f6d0ea

Digital Humanities

Nature SciData **2023**

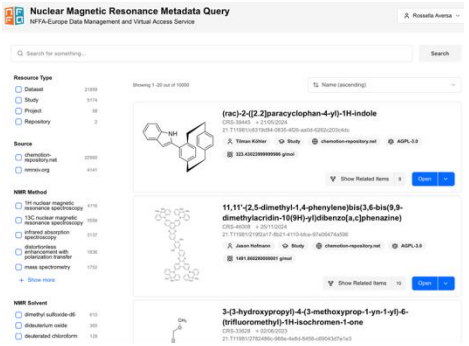
Thermal bridges on building rooftops

21.11152/6ea60288-d895-414e-80c0-26c9fdd662b2

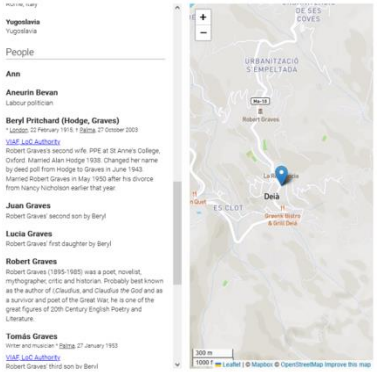
Energy

Looking for other applications

- **Data explorer:** overview of data (different repos, representations, infos)
  - important information can be harmonized
  - extra metadata (e.g., telescope/db-specific) in a different record for different actions

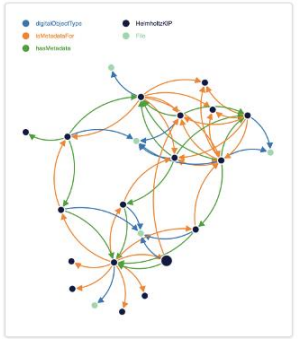


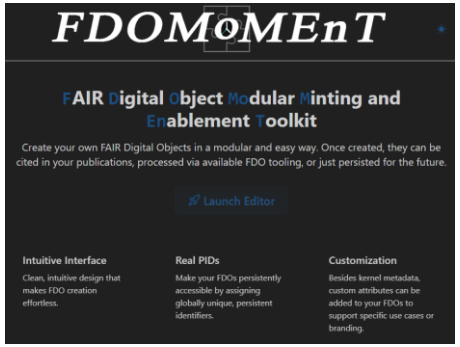
- **Interactive interfaces:** based on geolocation, portion of the sky, …
- Possible **operations** for machine-actionable datasets:
  - Spot if something is under embargo, for machine-driven decisions
  - Check for specific metadata in the record (because they are needed for training)
  - Show preview image
  - …



We can provide:

- assistance for modelling FDOs
- tools for creating FDOs
- specific visualisations (we have motivated experts)





CMB
~~Sky~~ is the limit