HELMHOLTZ Metadata Collaboration

NFDI MatWerk

MecaNano COST Action CA21121

JÜLICH Forschungszentrum

PIDs in the Helmholtz Knowledge Graph:
inferring types, resolving entities

Dr. Volker Hofmann

https://orcid.org/0000-0002-5149-603X

Institute for Advanced Simulation –
Materials Data Science and Informatics (IAS-9)

IAS-9 Materials Data Science & Informatics

JÜLICH Forschungszentrum

www.helmholtz-metadaten.de

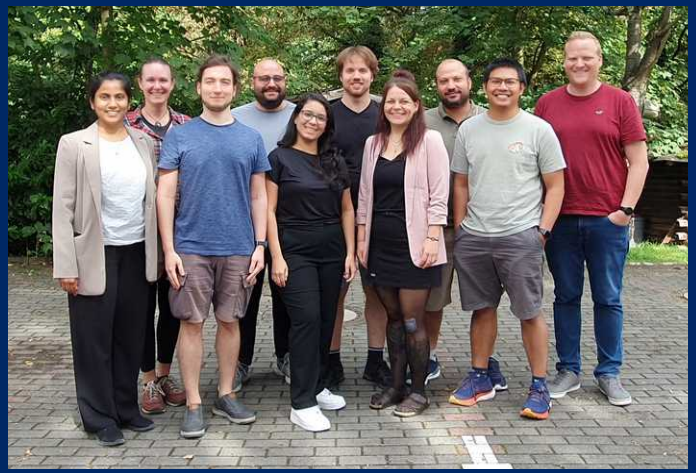
https://doi.org/10.5281/zenodo.10723293

HELMHOLTZ Metadata Collaboration

NFDI MatWerk

MecaNano COST Action CA21121

JÜLICH Forschungszentrum



Institute for Advanced Simulation –
Materials Data Science and Informatics (IAS-9)

IAS 9 Materials Data Science & Informatics

JÜLICH Forschungszentrum

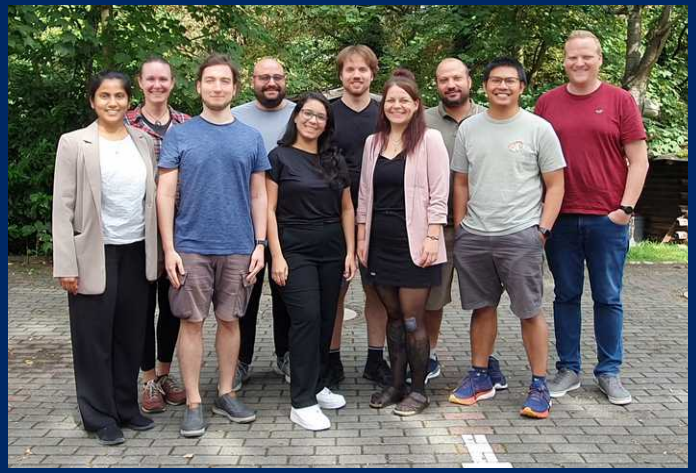
www.helmholtz-metadaten.de

HELMHOLTZ Metadata Collaboration

NFDI MatWerk

MecaNano COST Action CA21121

JÜLICH Forschungszentrum



Institute for Advanced Simulation –
Materials Data Science and Informatics (IAS-9)

IAS 9 Materials Data Science & Informatics

JÜLICH Forschungszentrum

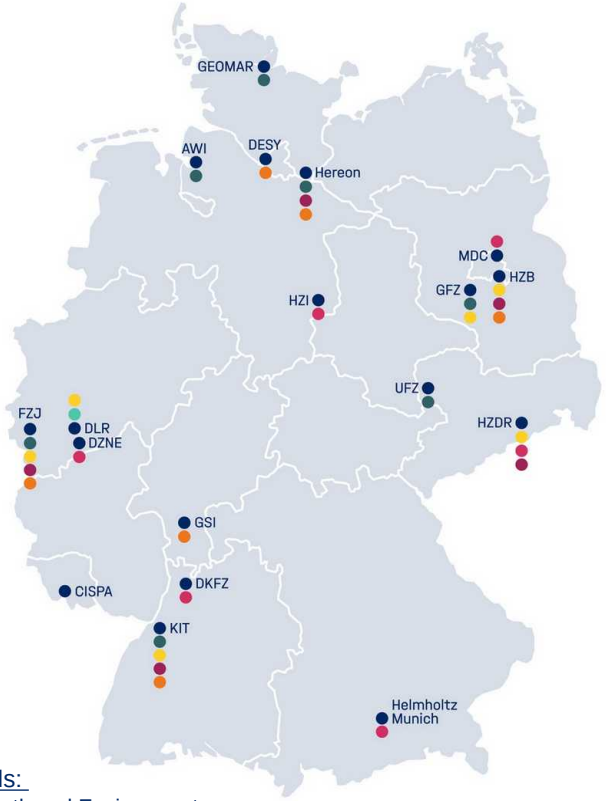
www.helmholtz-metadaten.de

HMC

Leverage the potential of metadata for the visibility and reusability of data across Helmholtz and beyond.

Establish a **Helmholtz FAIR data space**
in which data is found and re-used based on common practices and agreed upon norms

The Helmholtz digital ecosystem consists of data silos



Research Fields:
Energy, Earth and Environment
Health, Information
Aeronautics, Space and Tansport, Matter

18 independent research centres in 6 different research fields all of which host research data infrastructures

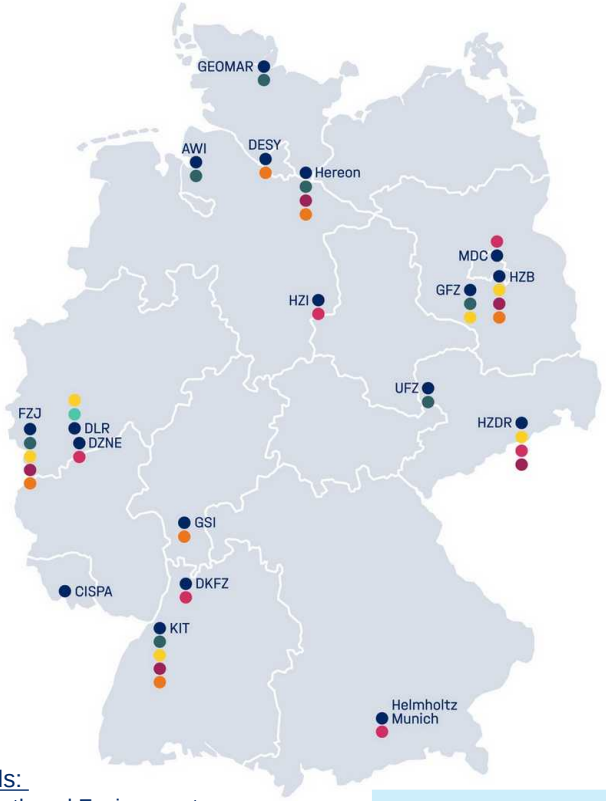
libraries
metadata on published research & data sets

data repositories
research data (cold, medium-hot or hot)

code repositories
code & software

research infrastructures
heterogenous methods and approaches

The Helmholtz digital ecosystem consists of data silos



Research Fields:
Energy, Earth and Environment
Health, Information
Aeronautics, Space and Tansport, Matter

18 independent research centres in 6 different research fields all of which host research data infrastructures

libraries
metadata on published research & data sets

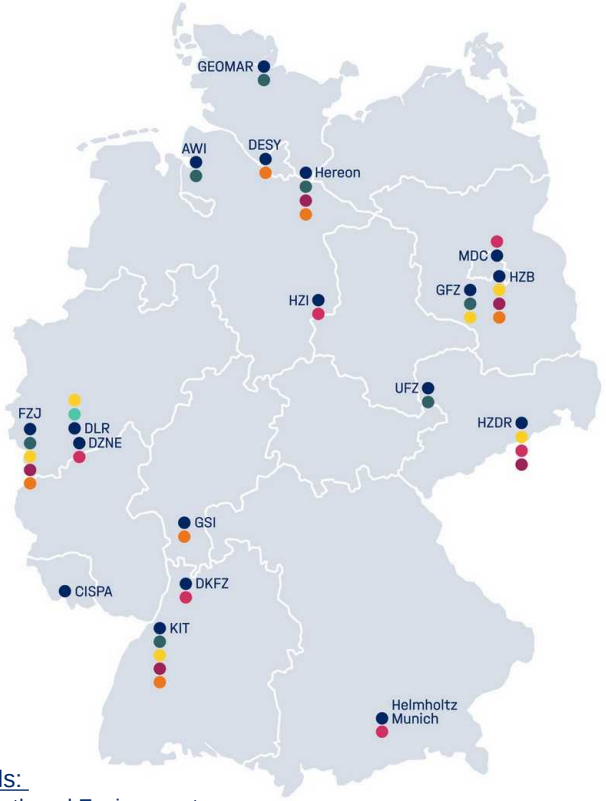
data repositories
research data (cold, medium-hot or hot)

code repositories
code & software

research infrastructures
heterogenous methods and approaches

How can we interconnect data from these sources?

UNHIDE - unified Helmholtz information and data exchange



Research Fields:
Energy, Earth and Environment
Health, Information
Aeronautics, Space and Tansport, Matter

UNHIDE Scope

1. Create a lightweight interoperability layer by consolidating (meta)data in an association wide knowledge graph.
2. Increase visibility of Helmholtz digital infrastructures
3. Improve quality and interoperability of Helmholtz metadata
4. Make digital assets easily findable
5. Assess the status quo of Helmholtz Metadata

HELMHOLTZKG

Helmholtz Knowledge Graph

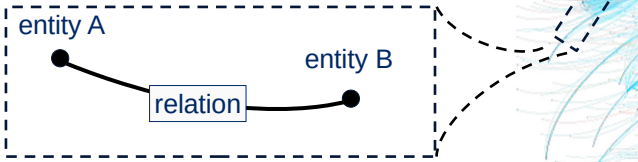
dev.container hosted on

HIFIS HDF CLOUD

research data infrastructures

provided records

connected records



HELMHOLTZKG

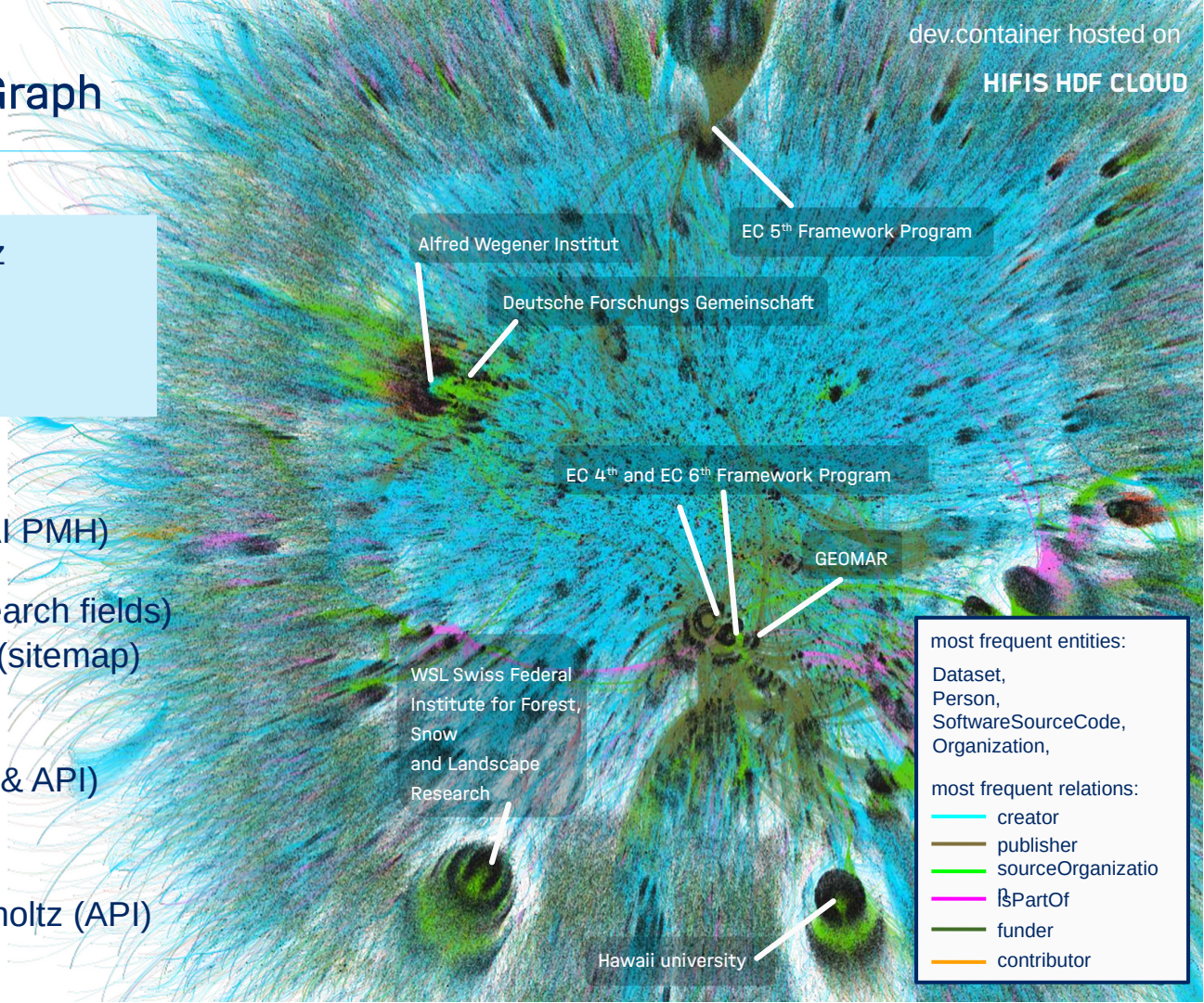
Helmholtz Knowledge Graph

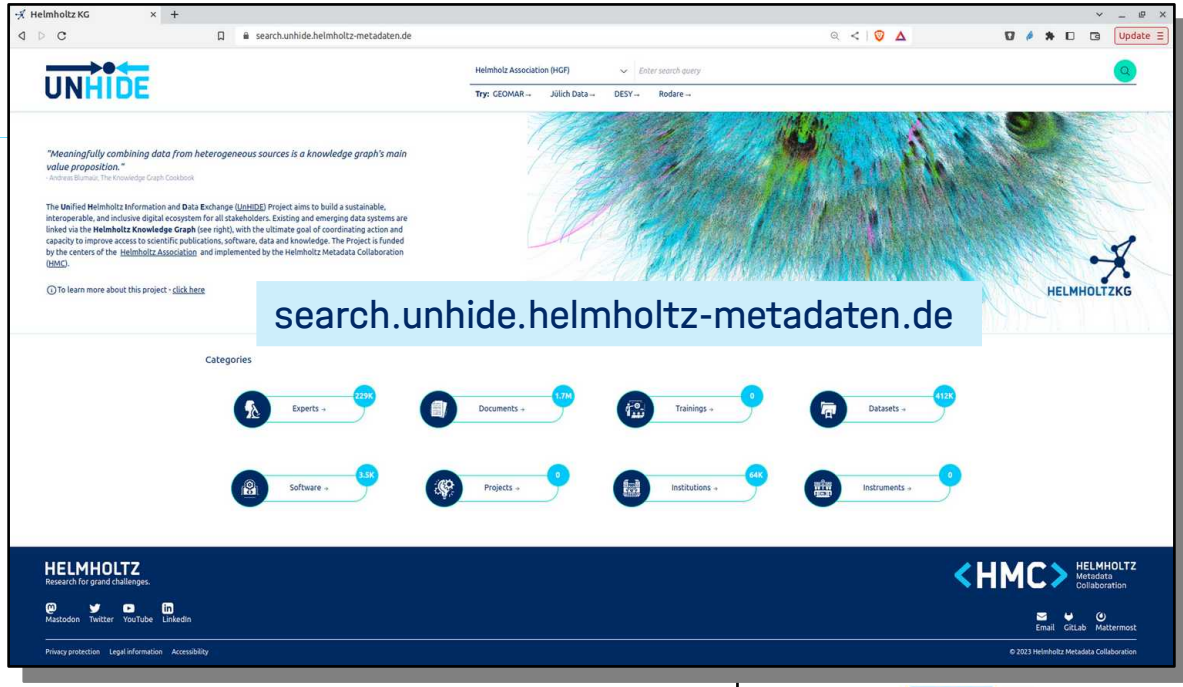
dev.container hosted on
HIFIS HDF CLOUD

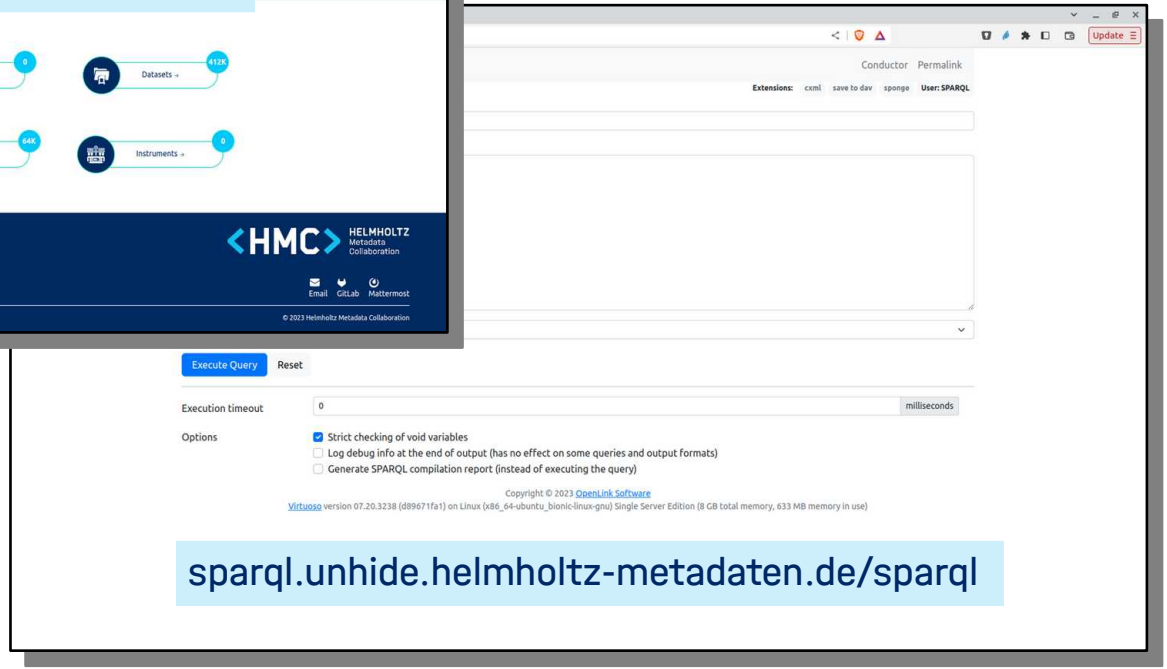
data from 32 different Helmholtz providers in one place!

> 2.15 mio records

- Libraries
  16/18 Helmholtz libraries (OAI PMH)
- data repositories (from 3 research fields)
  Rodare, Pangea, Jülich data (sitemap)
- code repositories
  12 (all) GitLab Instances (Git & API)
- global resources
  DataCite sub-graph for Helmholtz (API)

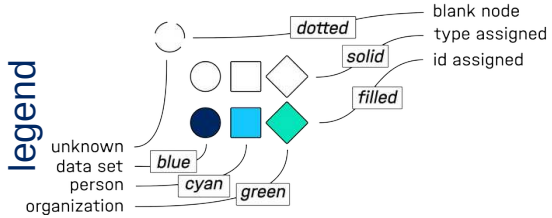



Helmholtz KG
search.unhide.helmholtz-metadaten.de
Update
UNHIDE
Helmholz Association (HGF)
Enter search query
Try: GEOMAR → Jülich Data → DESY → Rodare →
"Meaningfully combining data from heterogeneous sources is a knowledge graph's main value proposition."
- Andreas Blumauer, The Knowledge Graph Cookbook
The Unified Helmholtz Information and Data Exchange (UNHIDE) Project aims to build a sustainable, interoperable, and inclusive digital ecosystem for all stakeholders. Existing and emerging data systems are linked via the Helmholtz Knowledge Graph (see right), with the ultimate goal of coordinating action and capacity to improve access to scientific publications, software, data and knowledge. The Project is funded by the centers of the Helmholtz Association and implemented by the Helmholtz Metadata Collaboration (HMC).
To learn more about this project - click here
HELMHOLTZKG
search.unhide.helmholtz-metadaten.de
Categories
Experts → 229K
Documents → 1.7M
Trainings → 0
Datasets → 412K
Software → 3.5K
Projects → 0
Institutions → 64K
Instruments → 0
HELMHOLTZ
Research for grand challenges.
Mastodon Twitter YouTube LinkedIn
Privacy protection Legal information Accessibility
HMC HELMHOLTZ Metadata Collaboration
Email GitLab Mattermost
© 2023 Helmholtz Metadata Collaboration


Update
Conductor Permalink
Extensions: cxml save to dav sponge User: SPARQL
Execute Query
Reset
Execution timeout
0
milliseconds
Options
Strict checking of void variables
Log debug info at the end of output (has no effect on some queries and output formats)
Generate SPARQL compilation report (instead of executing the query)
Copyright © 2023 OpenLink Software
Virtuoso version 07.20.3238 (d89671fa1) on Linux (x86_64-ubuntu_bionic-linux-gnu) Single Server Edition (8 GB total memory, 633 MB memory in use)
sparql.unhide.helmholtz-metadaten.de/sparql

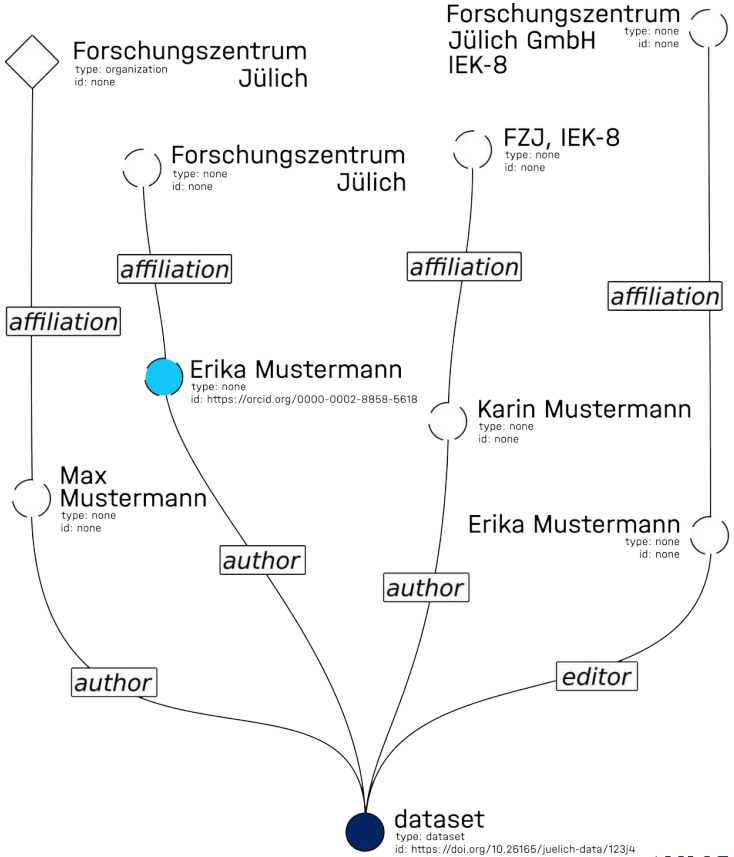
Improving metadata quality at the source

Provided data records are often incomplete, heterogeneous & messy




```
{  "@context": "http://schema.org",
   "@id": "https://doi.org/10.26165/juelich-data/123j4",
   "@type": "dataset",
   "name": "Name of the dataset.",
   "author": [ { "affiliation": { "@type": "Organization",
                                  "name": "Forschungszentrum J\u00fclich" },
                 "name": "Max Mustermann" },
               { "@id": "https://orcid.org/0000-0002-8858-5618"
                 "affiliation": { "name": "Forschungszentrum J\u00fclich" },
                 "name": "Erika Mustermann" },
               { "affiliation": { "name": "FZJ, IEK-8" },
                 "name": "Karin Mustermann" } ],
   "editor":   { "affiliation": { "@type": "Organization",
                                  "name": "Forschungszentrum J\u00fclich GmbH, IEK-8" },
                 "name": "Erika Mustermann" }}
```

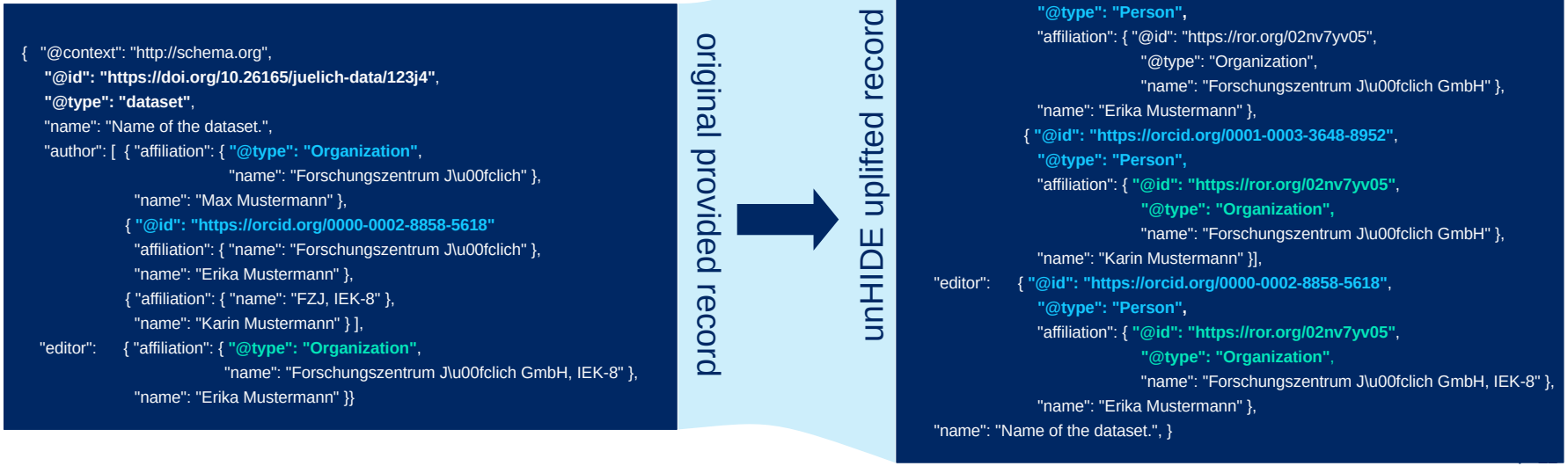

original provided record



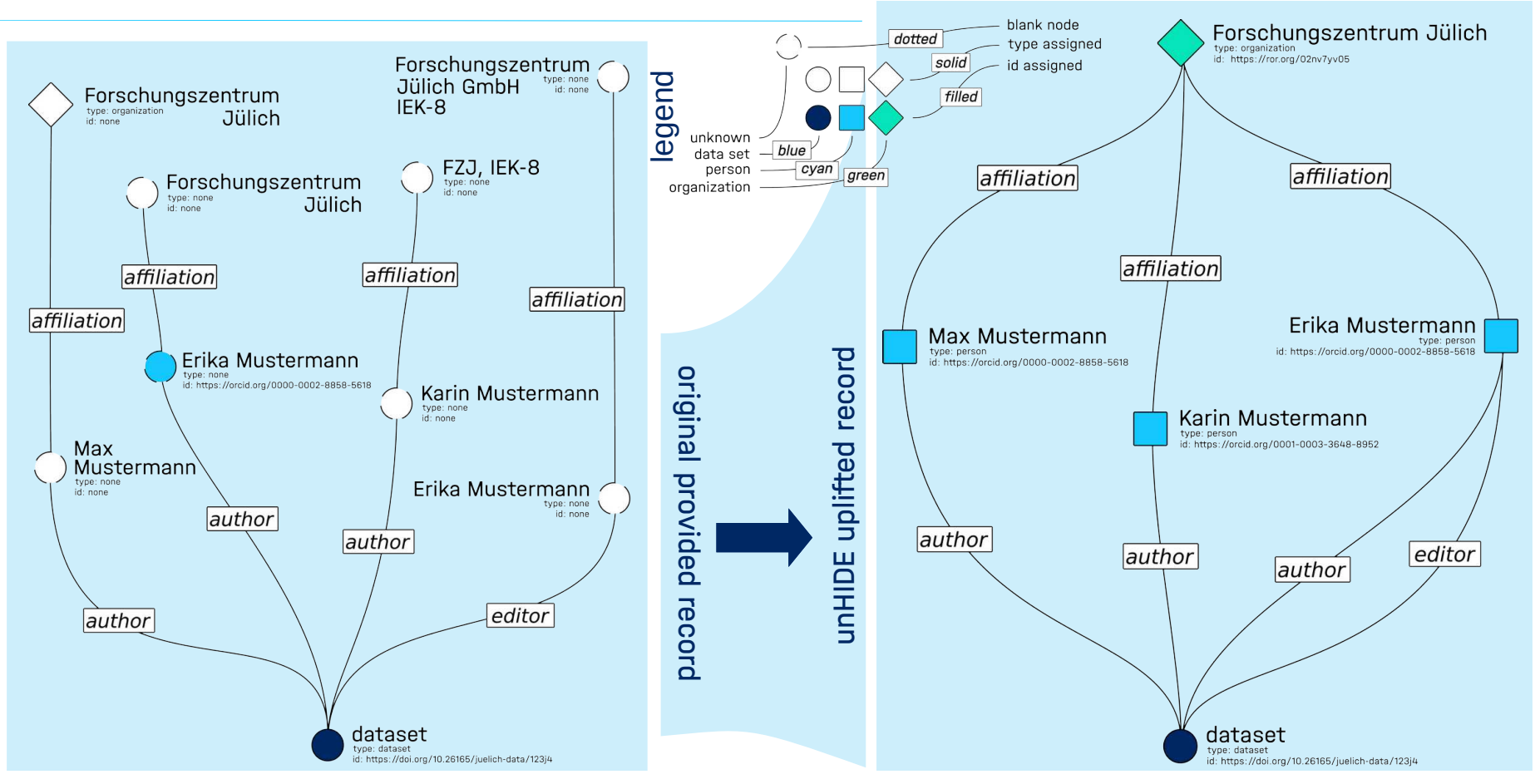
Improving metadata quality at the source

The assembled data allows uplifting records through type inference, general harmonization which allows resolving entities and assigning IDs.

Uplifting is recorded in reversible patches and will be fed back to data providers



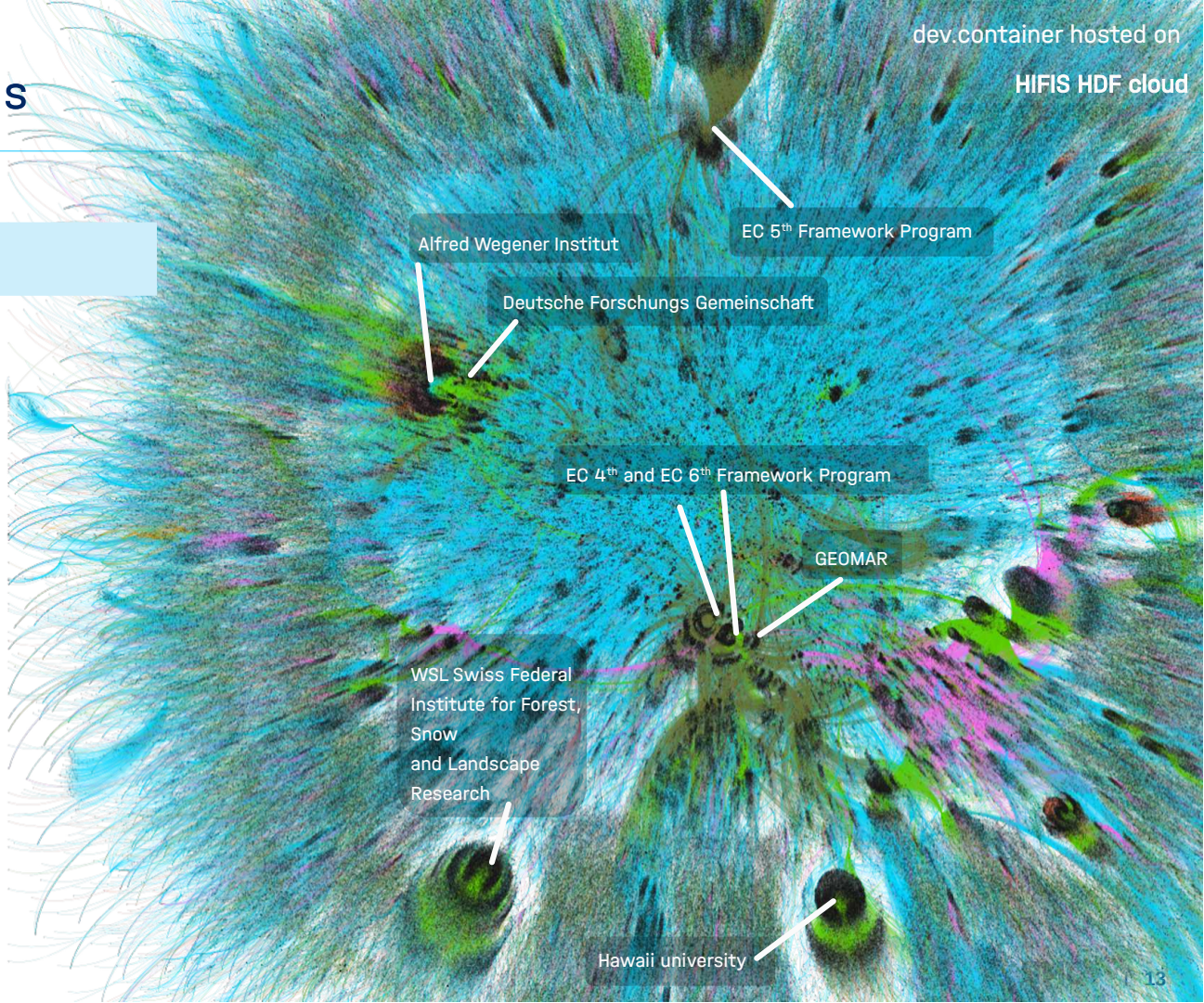
Uplifting increases structuredness of the graph



The KG data and its PIDs

Overall > **2.15 mio** records

- n triples: **71 Mio**
- n typed entities: **16.4 Mio**
  - with URI: 790 k
  - blank nodes: 15.5 Mio



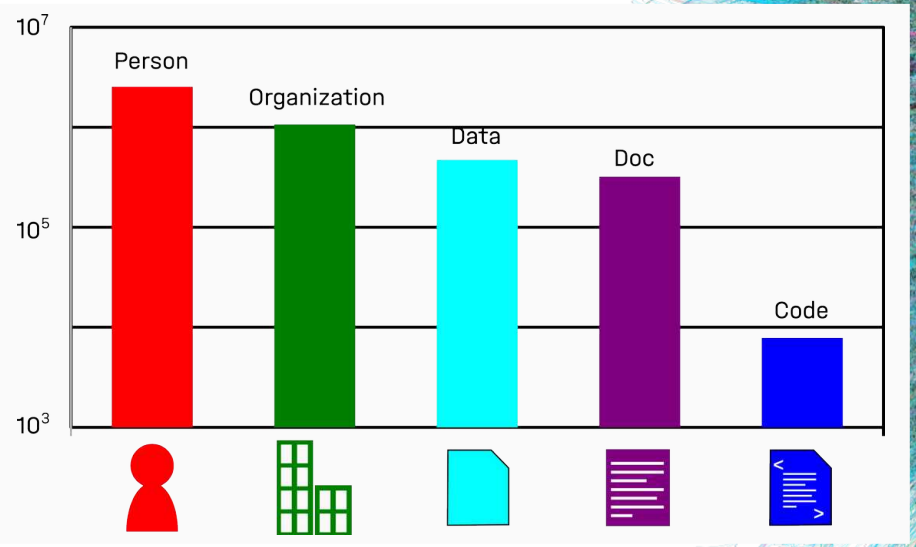
The KG data and its PIDs

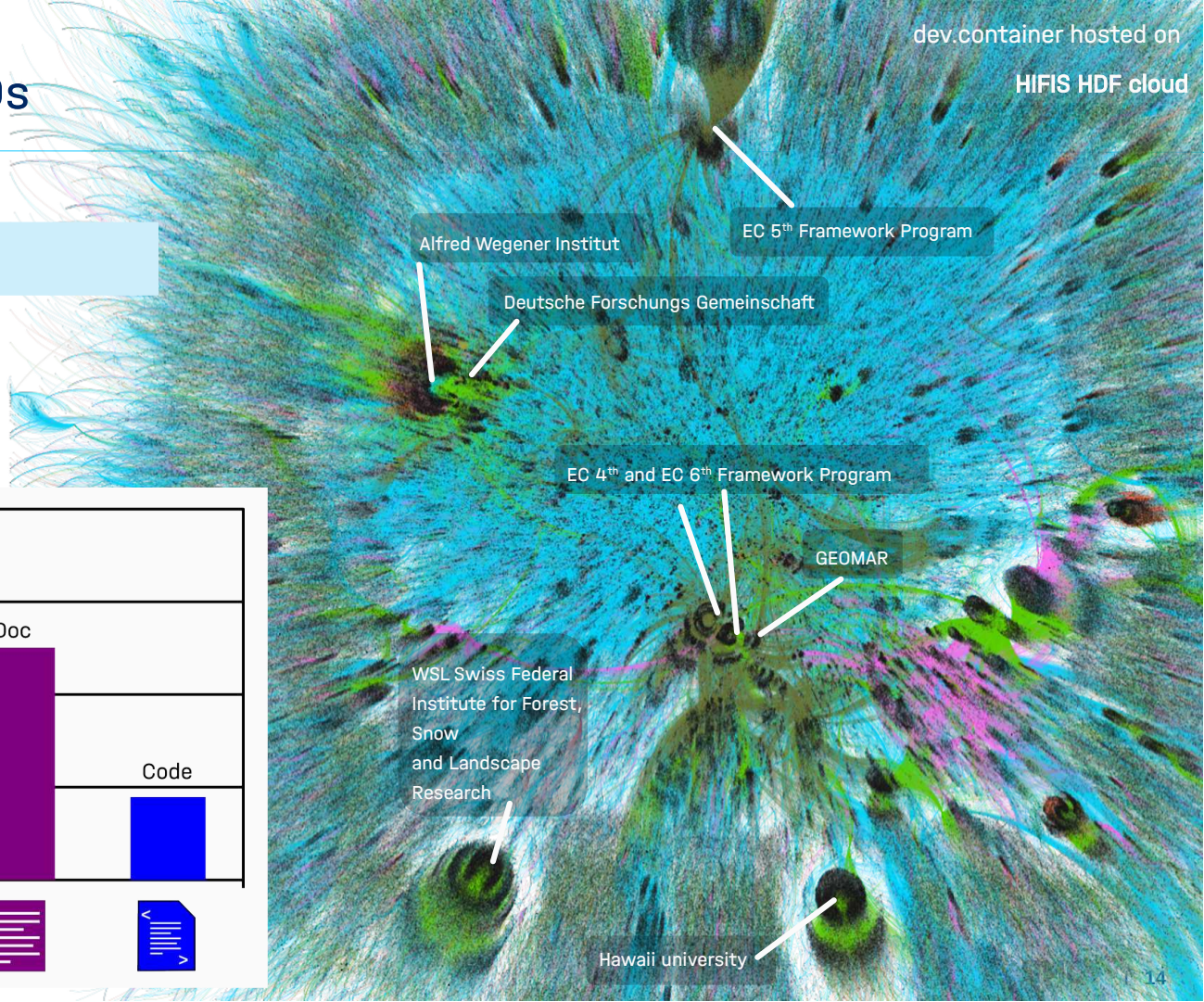
dev.container hosted on

HIFIS HDF cloud

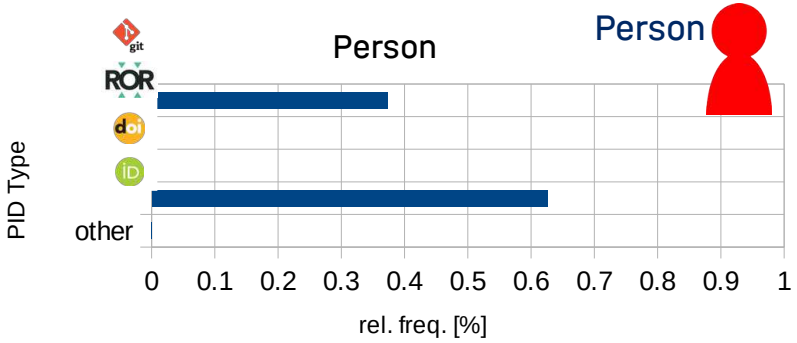
Overall > **2.15 mio** records

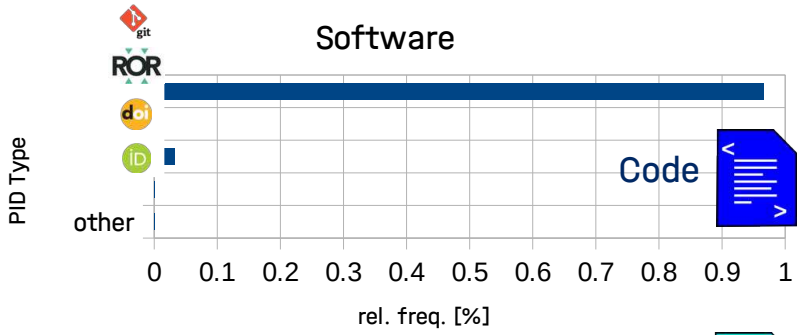
- n triples: **71 Mio**
- n typed entities: **16.4 Mio**

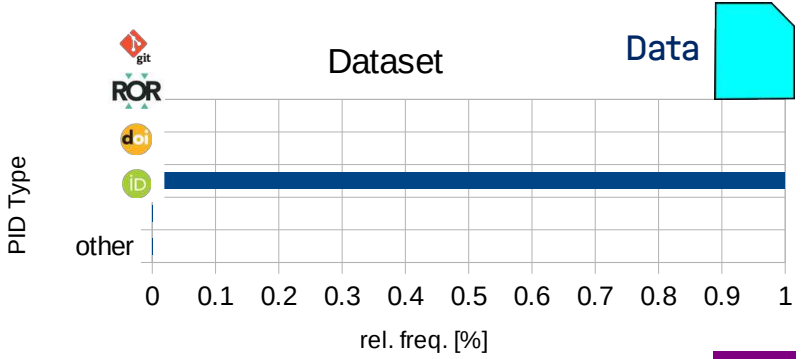


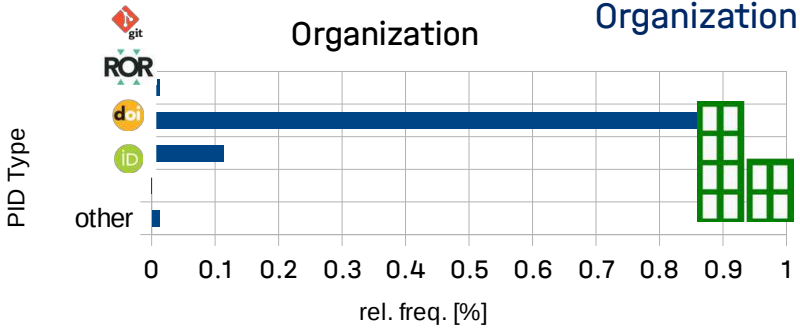


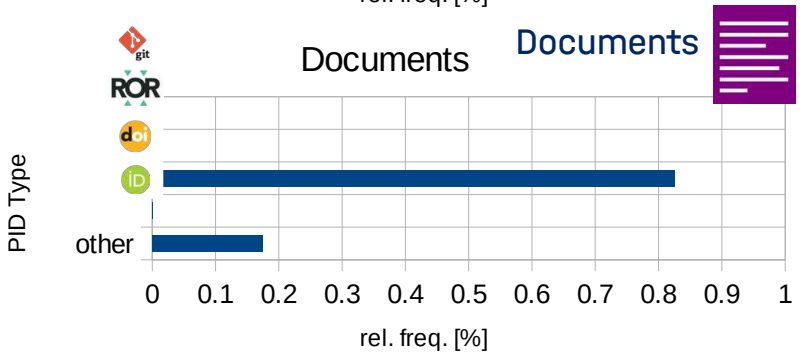
What PID for my resource?











What resource is behind my PID?

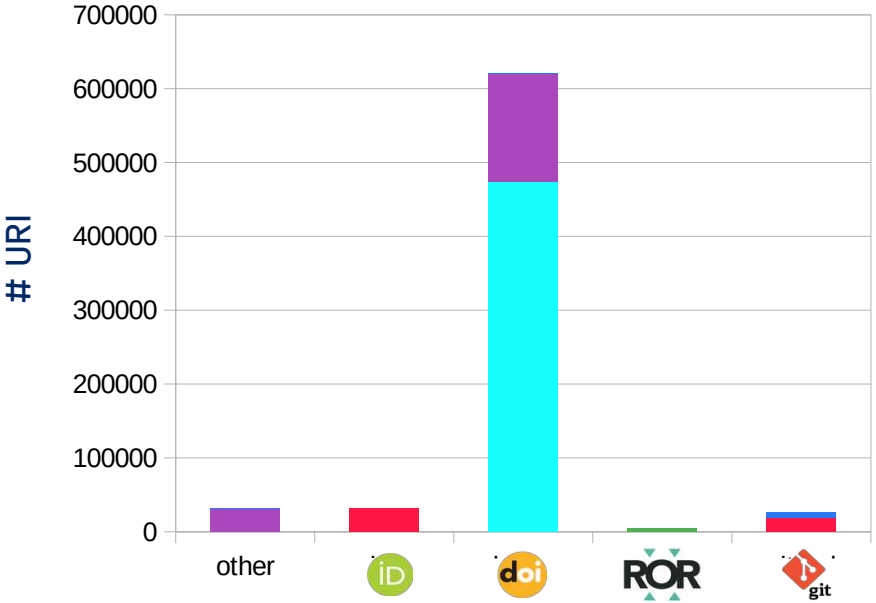
Person

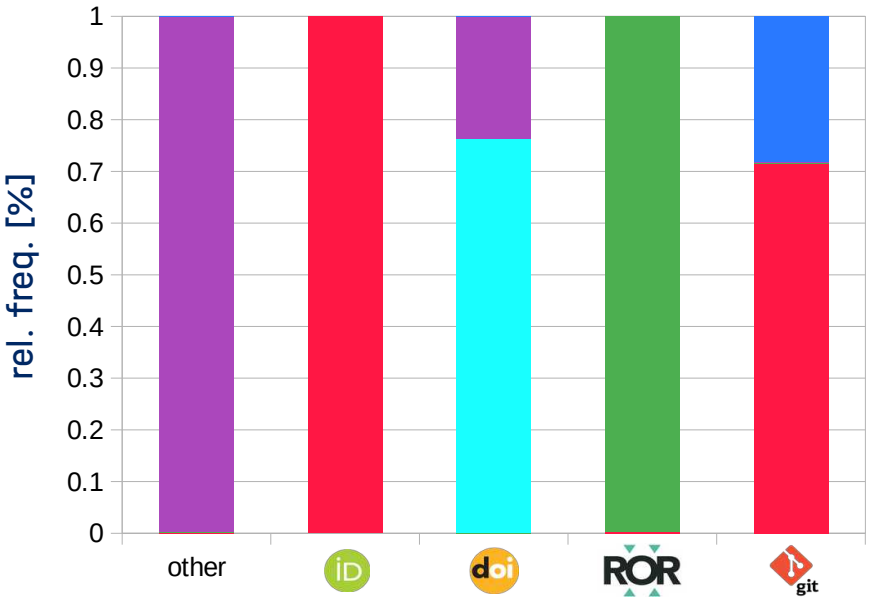
Organization

Code

Data

Documents





Take Home

- **DOIs ...and the rest**
  most common PIDs are DOIs, ORCID and ROR are a „not close“ 2nd.
- **What PID for my resource?**
  Code, data and documents – its clear what PID to use. No sufficient differentiation (Documents/Data)
- **Is a useful PID a useful identifier for my information?**
  Affiliation are mostly Strings – ROR Ids do not identify teh mostly desired “level” of organization
- **Type inference based on PIDs:**
  ORCIDs and ROR are useful as they are unambiguous

- **In an ideal world:** one type of PID for each semantic class

HELMHOLTZ Metadata Collaboration

NFDI MatWerk

MecaNano COST Action CA21121

JÜLICH Forschungszentrum

IAS-9 Director

Stefan Sandfeld

unHIDE & Helmholtz KG

Jens Bröder

Gabriel Preuß (HZB)
Said Fathalla
Fiona D'Mello
Pier Luigi Buttigieg (AWI)
Oonagh Mannix (HZB)

Department „Concepts & Tools for RDM"

Acknowledgements

www.helmholtz-metadaten.de

https://doi.org/10.5281/zenodo.10723293