HMC HELMHOLTZ Metadata Collaboration

IAS-9 Materials Data Science & Informatics

JÜLICH Forschungszentrum

How to UNHIDE and improve the metadata landscape of research software in Helmholtz.

Jens Bröder[1], F. D'Mello[1], G. Preuß[2], S. Fathalla[1], P. Buttigieg[3], O. Mannix[2], V. Hofmann[1], S. Sandfeld[1]

[1]Institute for Advanced Simulation, Materials Data Science and Informatics (IAS-9), Forschungszentrum Jülich GmbH

[2]Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB)

[3]Alfred Wegner Institute for Polar and Marine Research

deRSE 2024 | 07.03.2024

www.helmholtz-metadaten.de

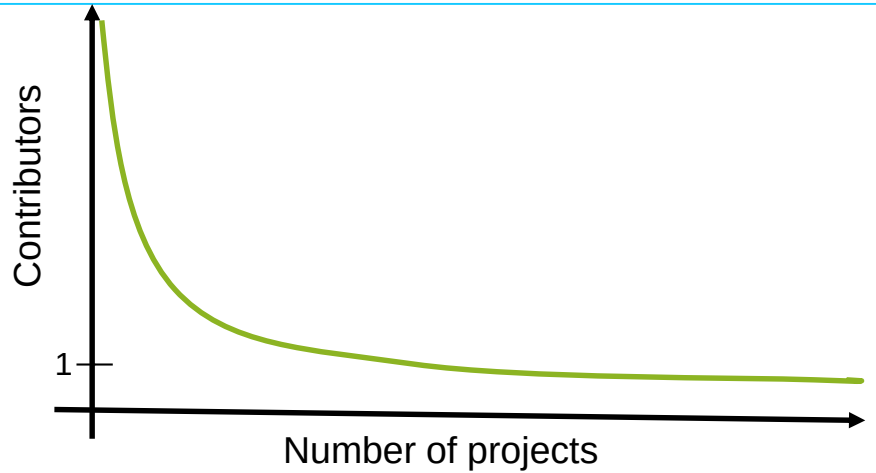
HELMHOLTZKG

Better metadata for research software now!

Problems with RSE:

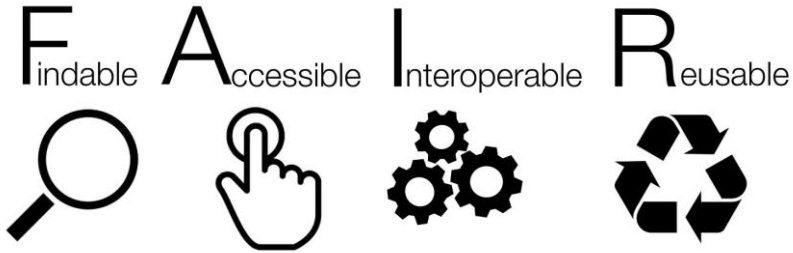
- Funding: Software is usually a by product
- Experience/Quality: Developer on the side
- Persistence: usually single developer
- Wrong incentive: I do my own software
- Hidden: uploaded somewhere

➡ Most projects die over a few years.



Better metadata helps:

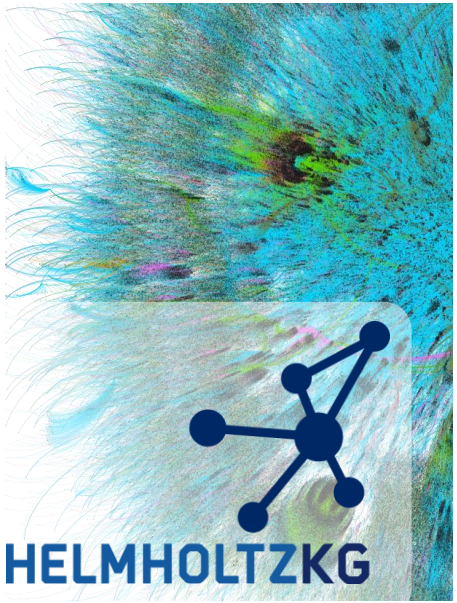
- Find existing software for your problem.
- Understanding software.
- Foster reuse of software and collaboration.



[1] M. D.Wilkinson. Scientific Data 3 (2016), pp. 1–9., also see www.go-fair.org

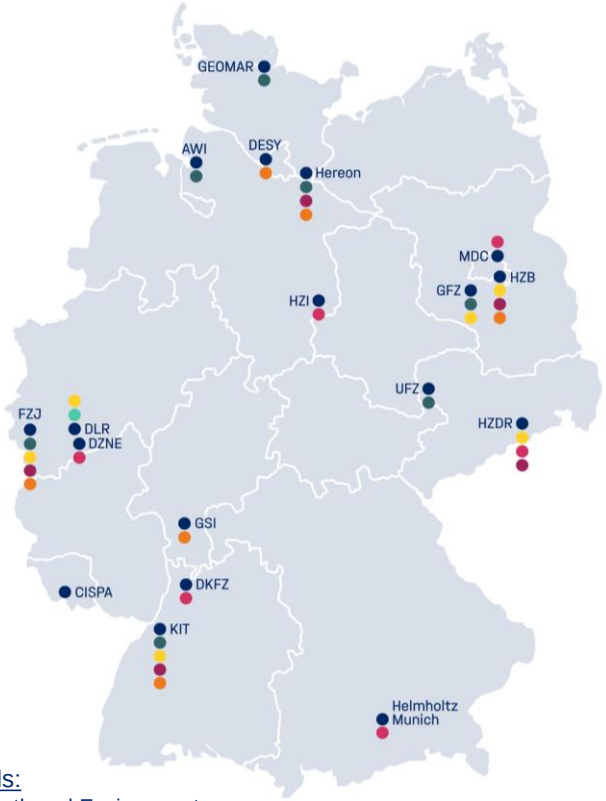
Outline

1. **UnHIDE initiative and the Helmholtz knowledge graph**:
   - Why, what, where?



2. **Software metadata in detail**:
   - General problems & status
   - Metadata in the graph
   - How to improve?

What is the (meta)data reality for the Helmholtz association?



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 digital infrastructure

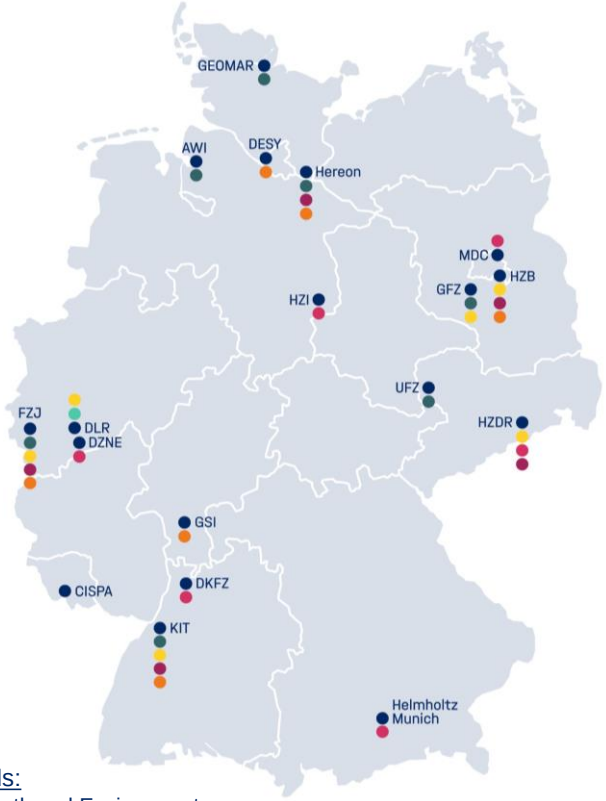
libraries
metadata on published research & data sets

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

code repositories, also external
code & software

research infrastructures
heterogenous methods and approaches

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 infrastrutures
3. Improve quality and interoperability of Helmholtz metadata
4. Make digital assests easily findable
5. Assess the status quo of Helmholtz Metadata

HELMHOLTZKG

Helmholtz Knowledge Graph

hosted on
HIFIS HDF cloud

research data infrastructures

Public (web) metadata records

connect records

conceptual & code co-development
with https://oceaninfohub.org/

HELMHOLTZKG

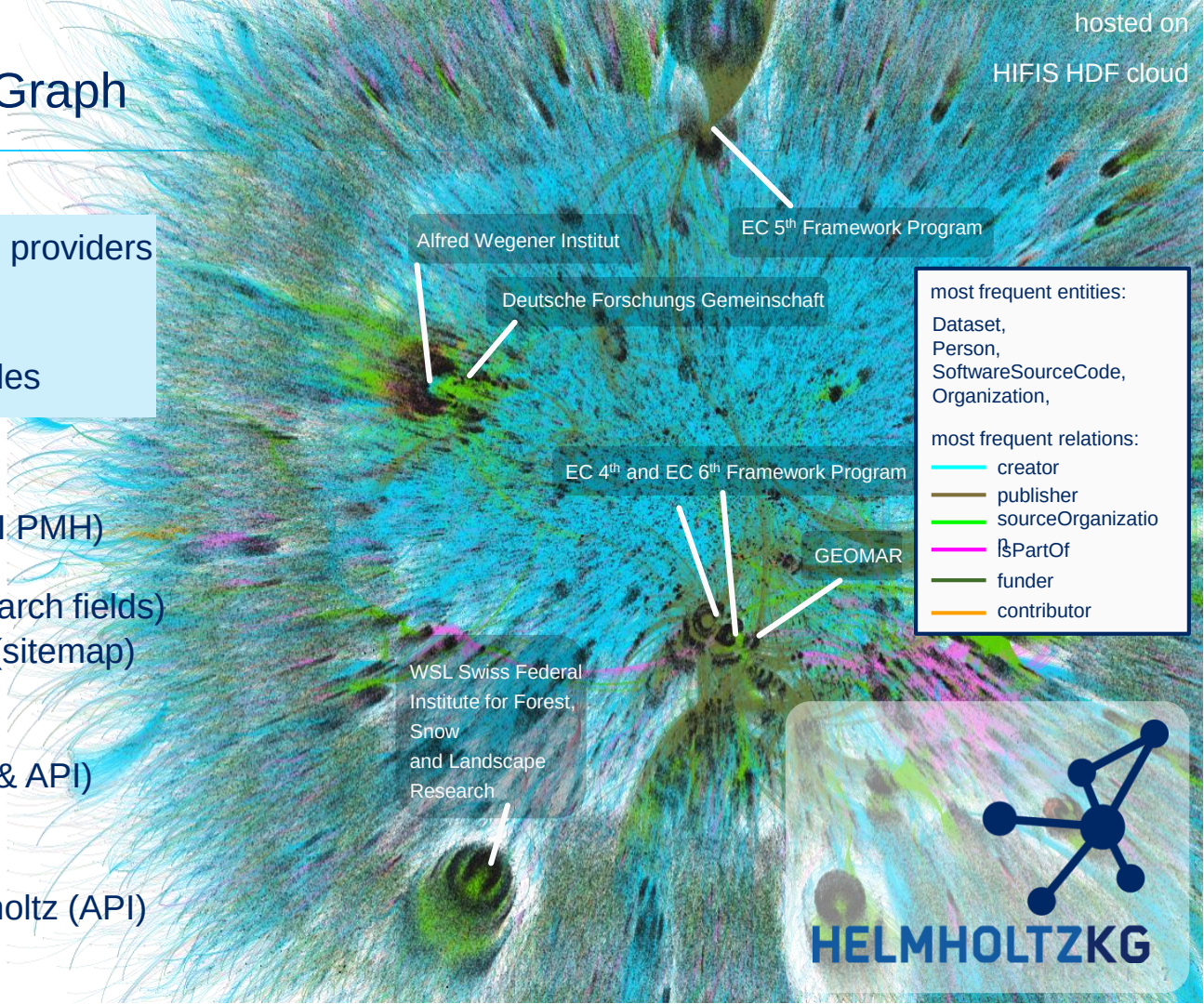
Helmholtz Knowledge Graph

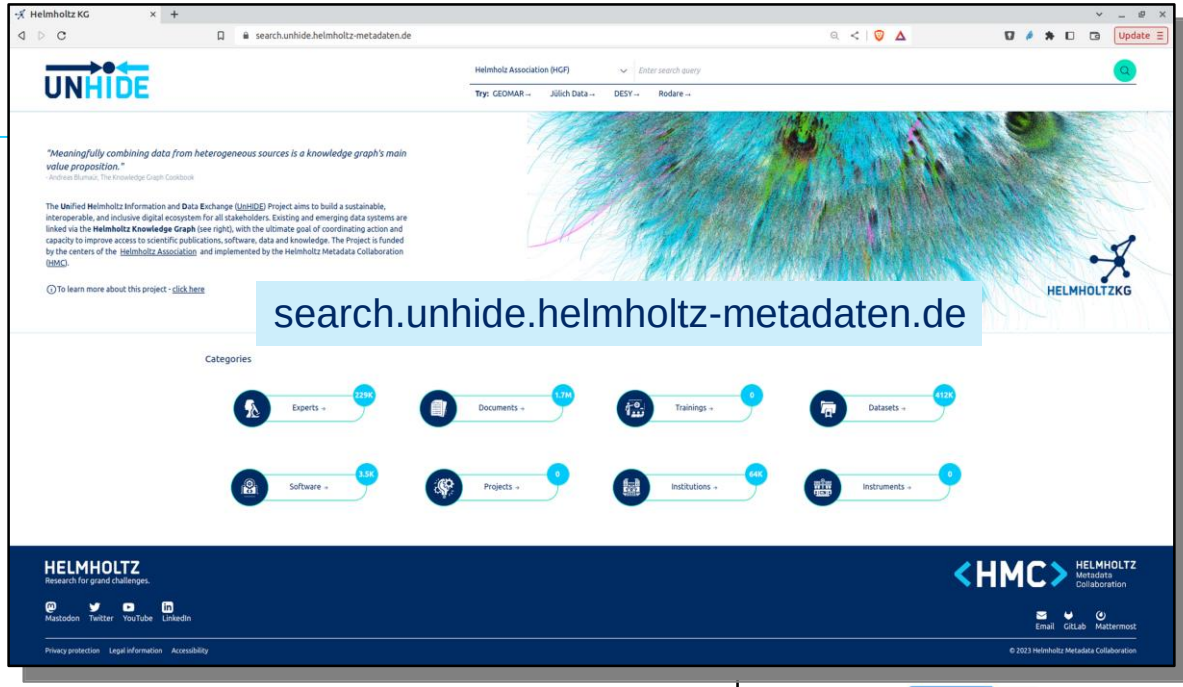
hosted on
HIFIS HDF cloud

data from 32 different Helmholtz providers in one place!

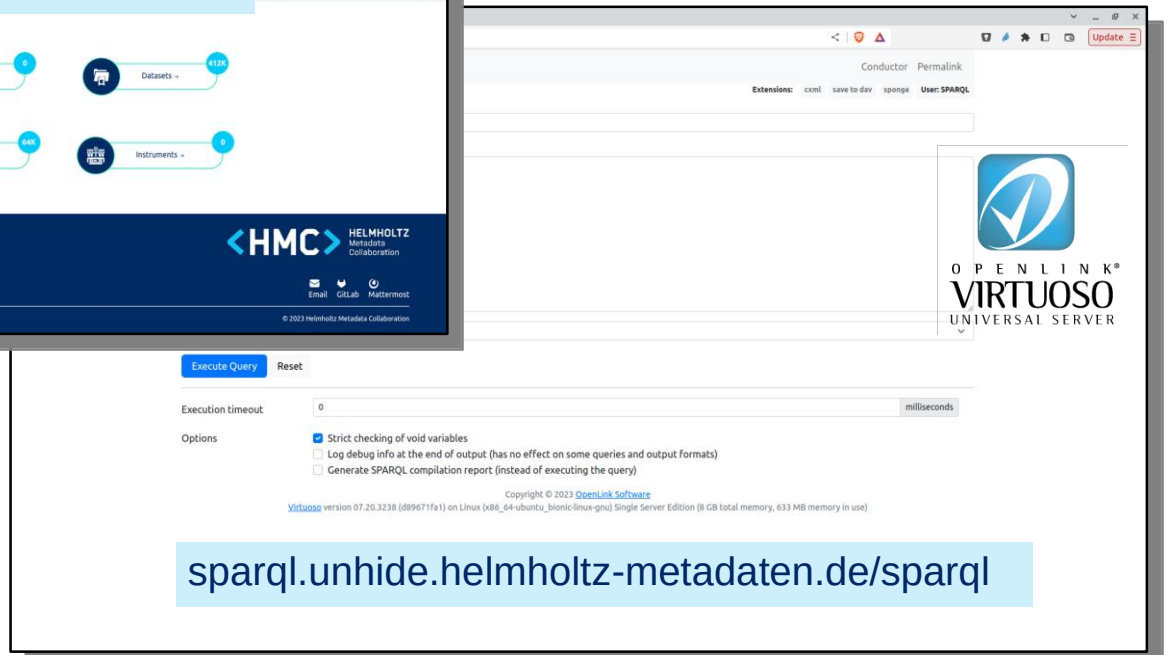
> 2.15 mio records | 7.1 mio triples

- 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)





search.unhide.helmholtz-metadaten.de



sparql.unhide.helmholtz-metadaten.de/sparql

docs.unhide.helmholtz-metadaten.de

2. Software metadata in detail

- General problems & status
- Metadata in the graph
- How to improve?

Metadata in usage for (research) software

What?	For Software
Standards	Git, language specific, citation.cff, codemeta, …
Size	small
Metadata enrichment/ Platforms	- Manually - Tracks the whole thing - Simple type of change (text)
Re-use case	- Similar purpose (usually) - As a whole - Install whole, use part (library)

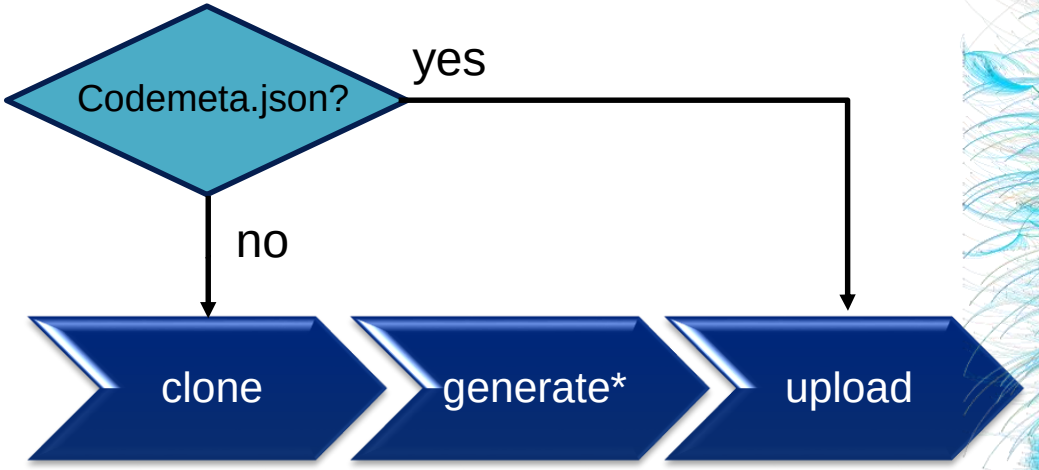
For datasets:

Different story!
Much, much harder

Data pipeline for software in detail:

Git Harvester:

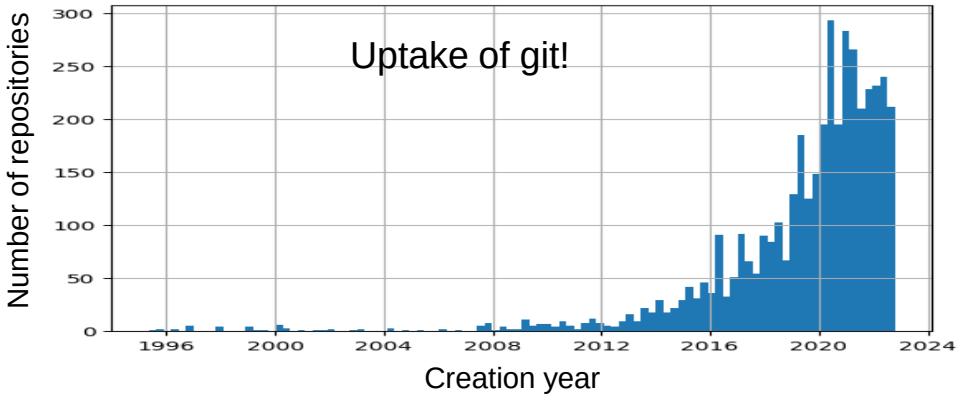
PREFECT

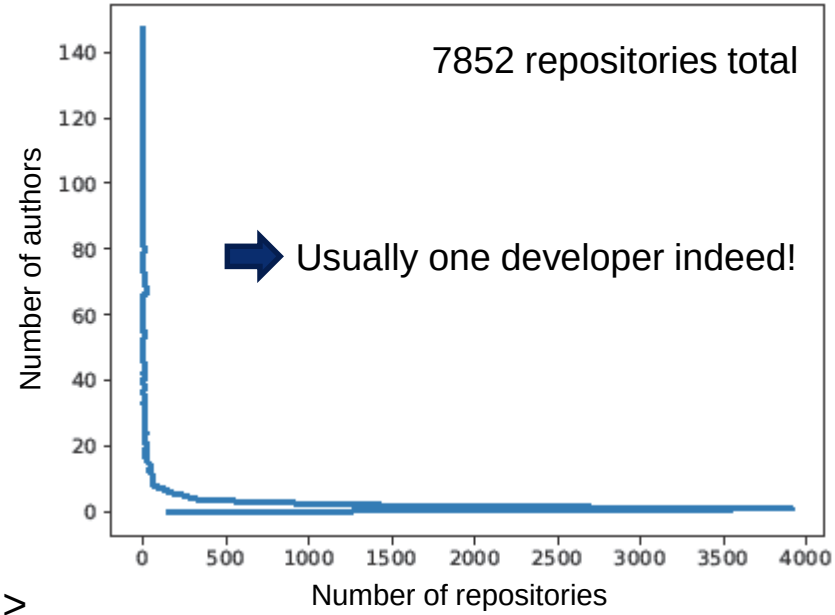


*Run Codemeta-harvester[1] (codemetapy[2], github-linguist[3], git commands)

[1] https://github.com/proycon/codemeta-harvester
[2] https://github.com/proycon/codemetapy
[3] https://github.com/github/linguist/

Overview of graph content with respect to RSE software

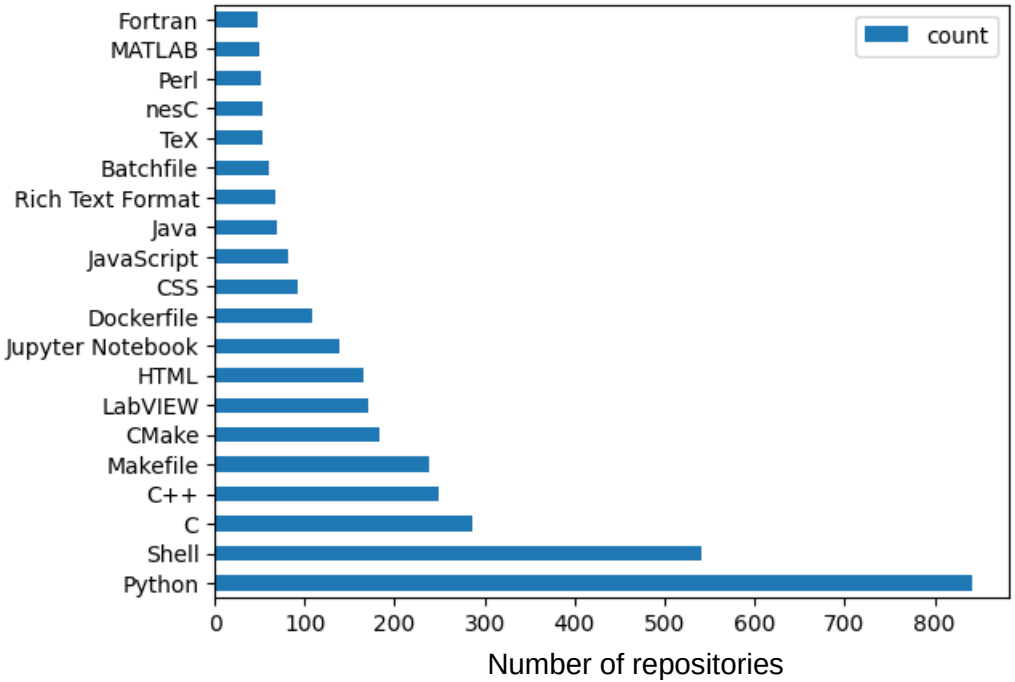




Statistics:

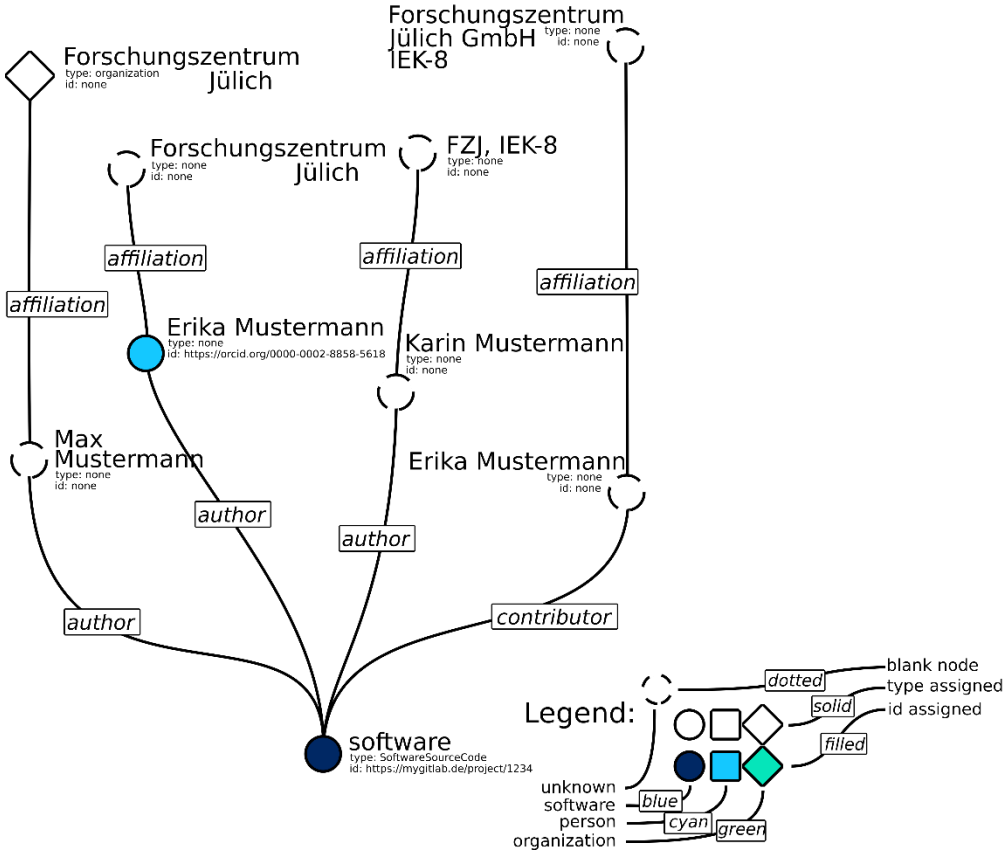
- 600+ active, 800+ WIP (no release)
- 2500+ licenses (MIT > GPL-3 > Apache > BSD > other GPL > others)
- 4500+ tags (possible releases)

Overview of Language use



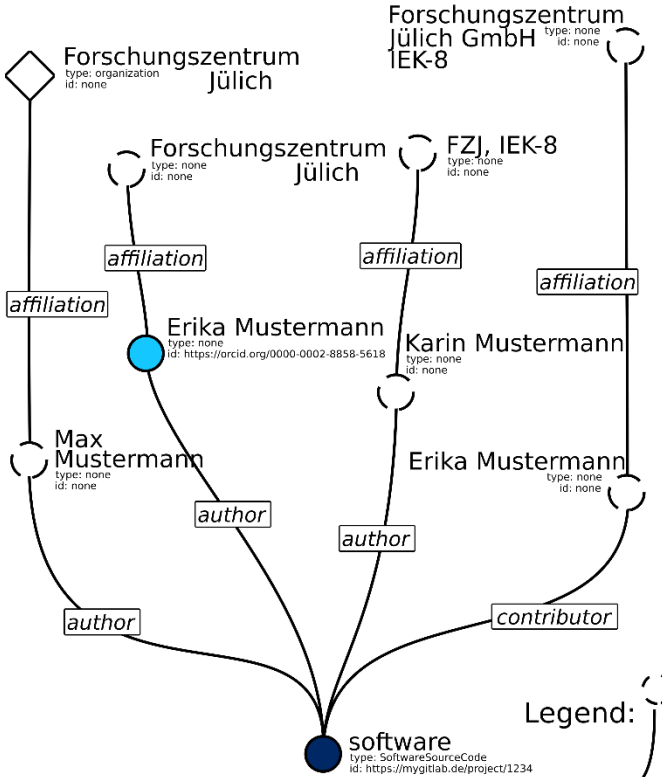
Future: Improving software metadata

Missing PIDS and semantics:

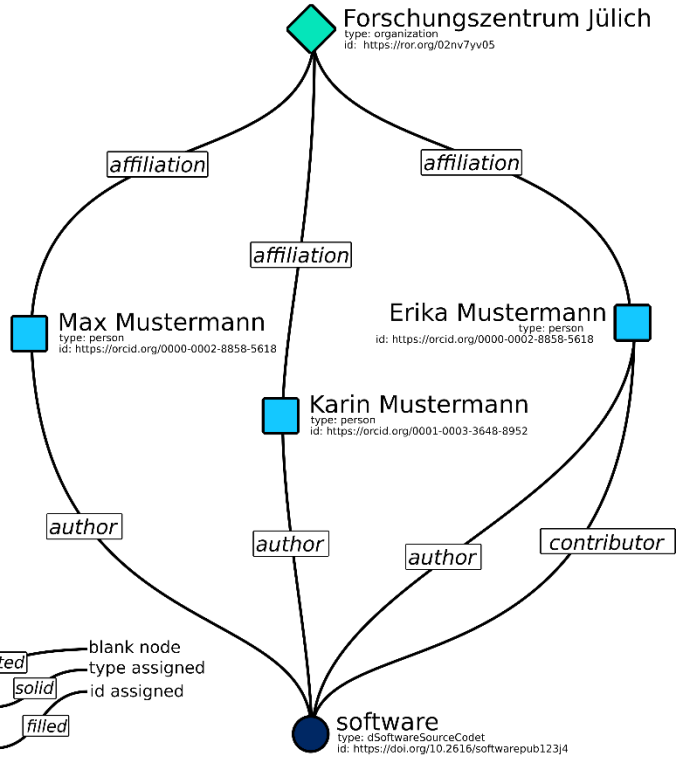


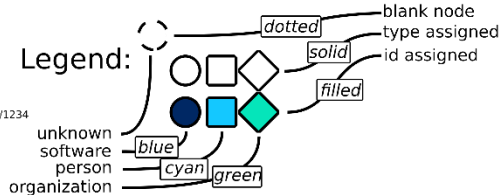
Future: Improving software metadata

Missing PIDS and semantics:



Complete PIDS and semantics:





the UNHIDE team

Total: ~2-3 FTEs for one year

Strategy

Volker Hofmann
Strategy Development
FZJ / Information

(Pier Luigi Buttigieg)
Former Member
Strategy Development
AWI

Development

Jens Bröder
Tech Lead
FZJ / Information

Gabriel Preuß
Full stack
HZB / Matter

Said Fathalla
Triple Store; SPARQL
FZJ / Information

Fiona D'Mello
Web FrontEnd
FZJ / FDC

Community

Oonagh Mannix
Outreach
HZB / Matter

(Nina Weisweiler)
Former Member
Infrastructures, GFZ

Acknowledgements:

Anton Pirogov: Technical design discussions,
Mustafa Soylu: Deployment help,
Vivien Serve, Markus Kubin and
Özlem Özkan: UX-workshops,
Pedro V. Barranco: Data/endpoints investigations,
Annika Strupp: early discussion about sense
Stefan Sandfeld (director): Providing resources.
Other HMC staff contributing to visions.

Outlook:

- A medata bot, with automatic suggestions, improve gitlab/github apis to better expose metadata, software publications
- Harvest also data from github. Problem: finding software by affiliation is hard.

Outlook:

- A medata bot, with automatic suggestions, improve gitlab/github apis to better expose metadata, software publications
- Harvest also data from github. Problem: finding software by affiliation is hard.

Sum up & take away:

- The goal of the Helmholtz knowledge graph is to make the metadata in Helmholtz visible and improve it ultimately at the source. https://helmholtz-metadaten.de/en/unhide_helmholtz-kg
- We provide federated search for software project through metadata of all Helmholtz gitlab instances

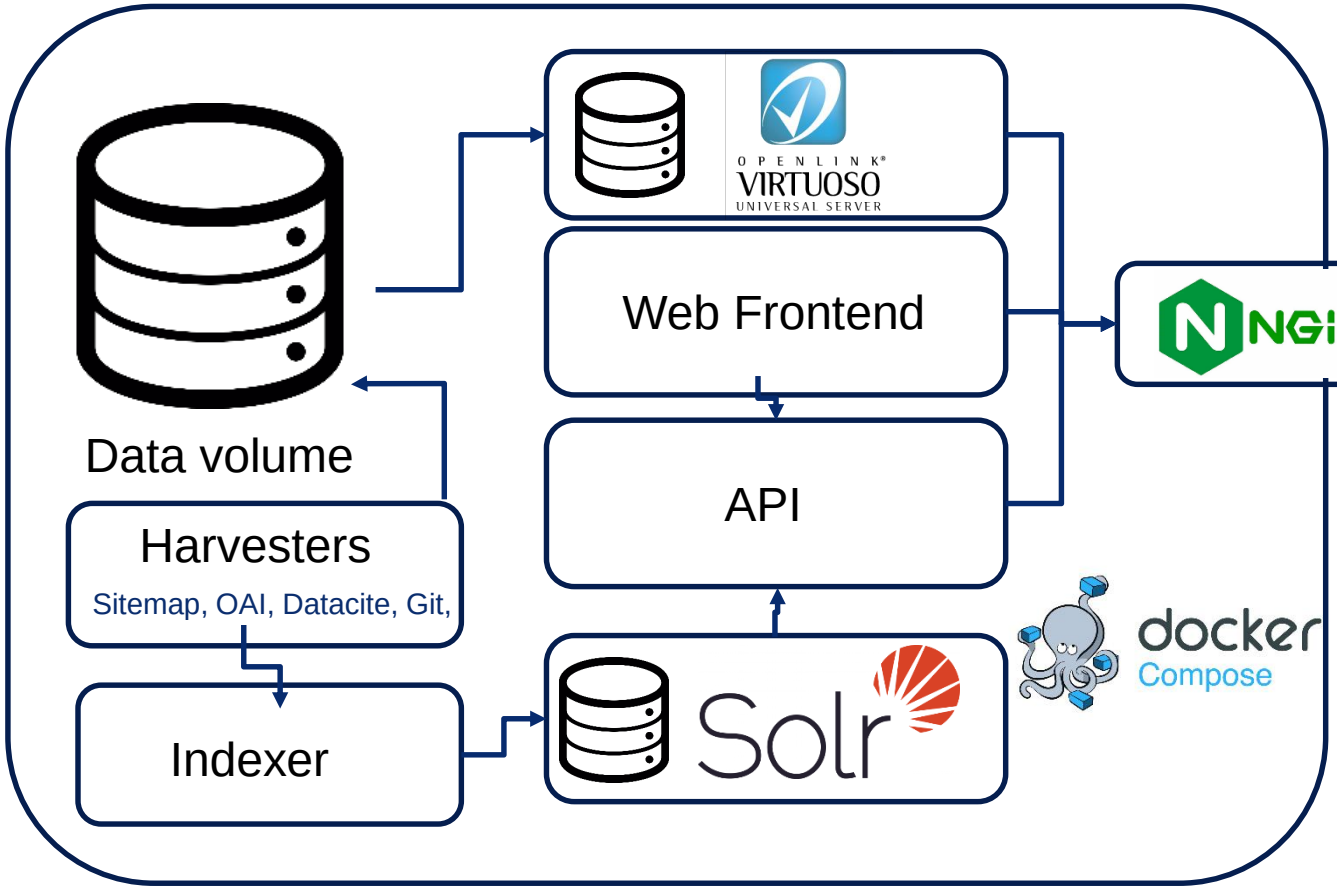
Thank you for your attention!

www.helmholtz-metadaten.de

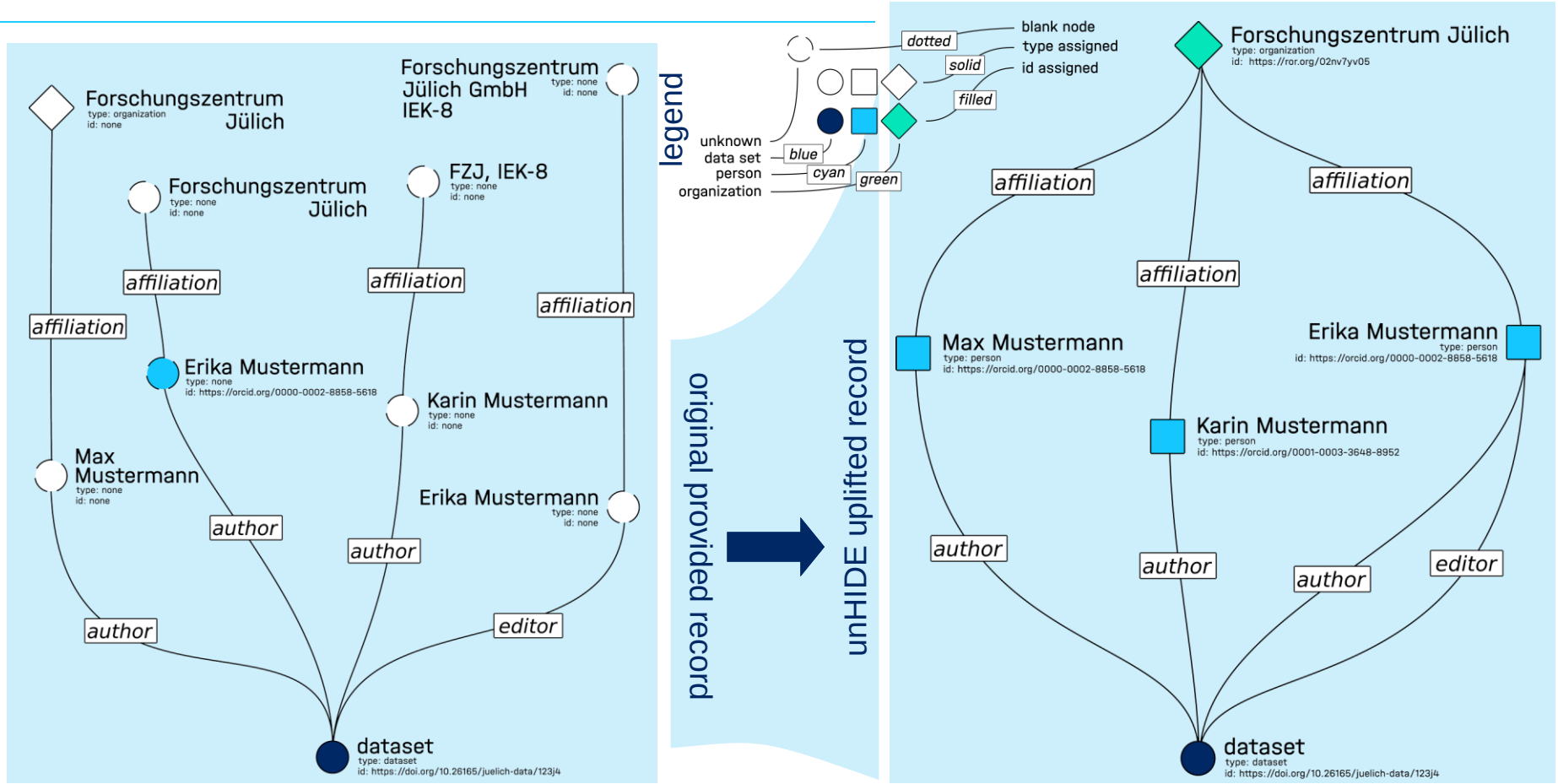
Technical view behind graph:

HELMHOLTZ
Data Federation | HDF

HELMHOLTZKG

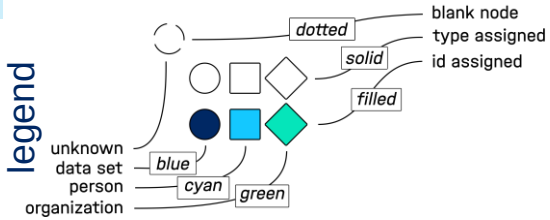


Uplifting increases structuredness of the graph



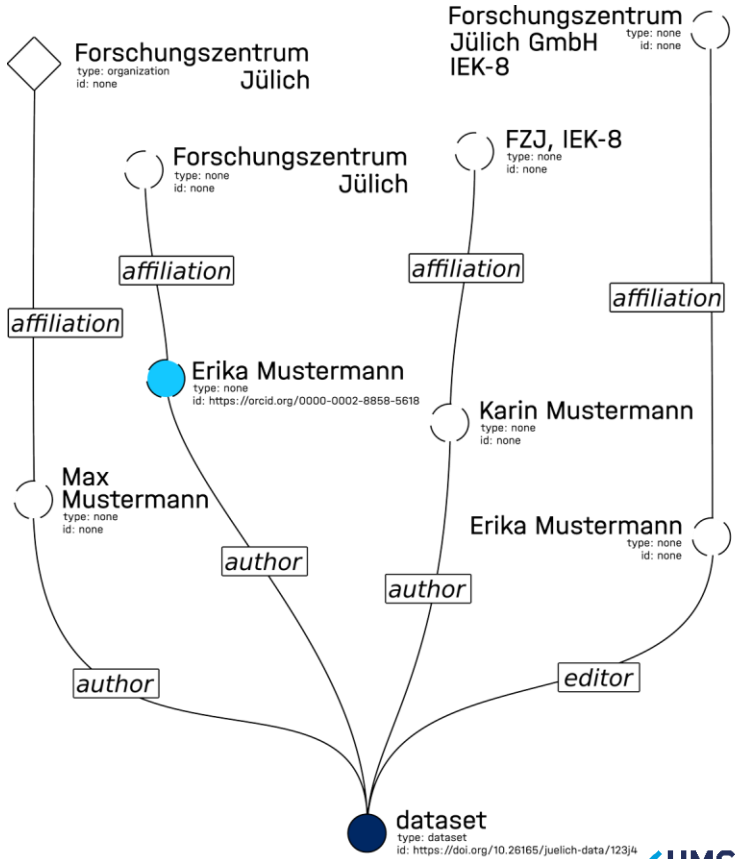
Improving metadata quality at the source

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


```
{   "@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

unHIDE uplifted record

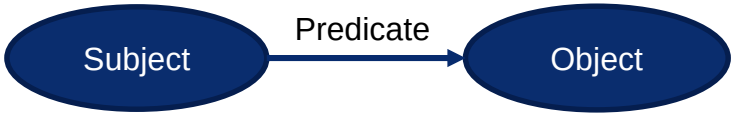

```
{   "@context": "http://schema.org",
    "@id": "https://doi.org/10.26165/juelich-data/123j4",
    "@type": "dataset",
    "author": [  { "@id": "https://orcid.org/0000-0003-3648-8952",
                   "@type": "Person",
                   "affiliation": { "@id": "https://ror.org/02nv7yv05",
                                    "@type": "Organization",
                                    "name": "Forschungszentrum J\u00fclich GmbH" },
                   "name": "Max Mustermann" },
                 { "@id": "https://orcid.org/0000-0002-8858-5618",
                   "@type": "Person",
                   "affiliation": { "@id": "https://ror.org/02nv7yv05",
                                    "@type": "Organization",
                                    "name": "Forschungszentrum J\u00fclich GmbH" },
                   "name": "Erika Mustermann" },
                 { "@id": "https://orcid.org/0001-0003-3648-8952",
                   "@type": "Person",
                   "affiliation": { "@id": "https://ror.org/02nv7yv05",
                                    "@type": "Organization",
                                    "name": "Forschungszentrum J\u00fclich GmbH" },
                   "name": "Karin Mustermann" }],
    "editor":      { "@id": "https://orcid.org/0000-0002-8858-5618",
                     "@type": "Person",
                     "affiliation": { "@id": "https://ror.org/02nv7yv05",
                                      "@type": "Organization",
                                      "name": "Forschungszentrum J\u00fclich GmbH, IEK-8"
},
                     "name": "Erika Mustermann" },
    "name": "Name of the dataset.", }
```

Basics Linked Data in JSON-LD serialization example

Library:

```
{"@context": {
    "dc11": "http://purl.org/dc/elements/1.1/",
    "ex": "http://example.org/vocab#",
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "ex:contains": {"@type": "@id"}
  },
  "@graph": [
    {"@id": "http://example.org/library",
      "@type": "ex:Library",
      "ex:contains": "http://example.org/library/the-republic},
    {"@id": "http://example.org/library/the-republic",
      "@type": "ex:Book",
      "dc11:creator": "Plato",
      "dc11:title": "The Republic",
      "ex:contains": "http://example.org/library/the-republic#introduction"},
    {"@id": "http://example.org/library/the-republic#introduction",
      "@type": "ex:Chapter",
      "dc11:description": "An introductory chapter on The Republic.",
      "dc11:title": "The Introduction"}
  ]
}
```

Semantic/RDF triple:



Person:

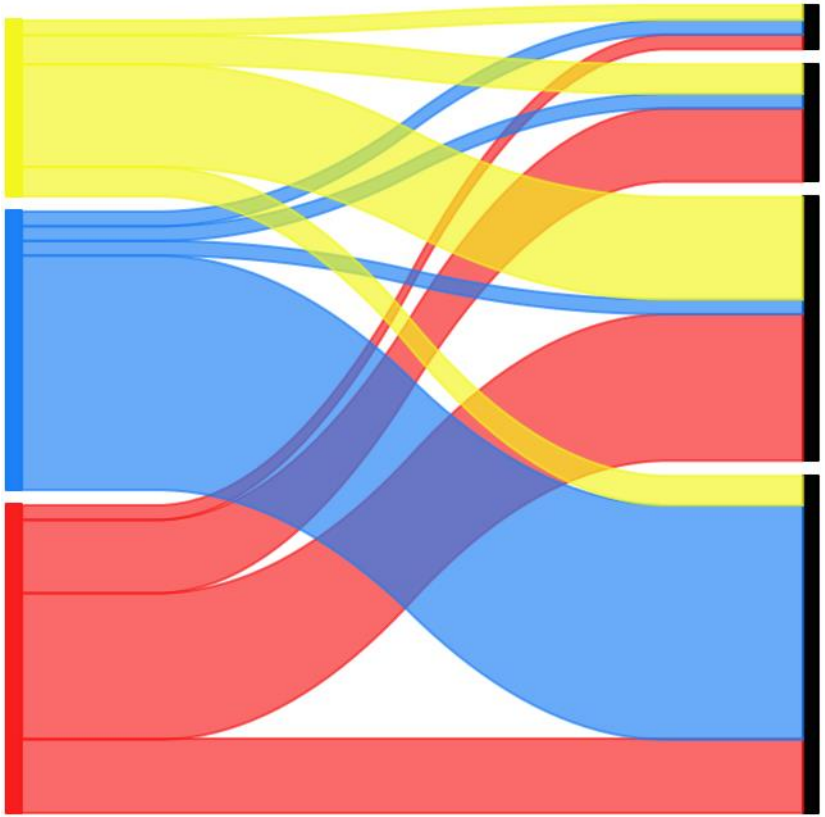
```
{
  "@context": "http://schema.org/",
  “@id”: “http://orcid.org/0000-0002-1584-4316“,
  "@type": "Person",
  "name": "Jane Doe",
  "jobTitle": "Professor",
  "telephone": "(425) 123-4567",
  "url": "http://www.janedoe.com"
}
```

The Helmholtz (Meta)data ecosystems

Data Stewards

Administration staff

Scientists



research infrastructures

code repositories

data repositories

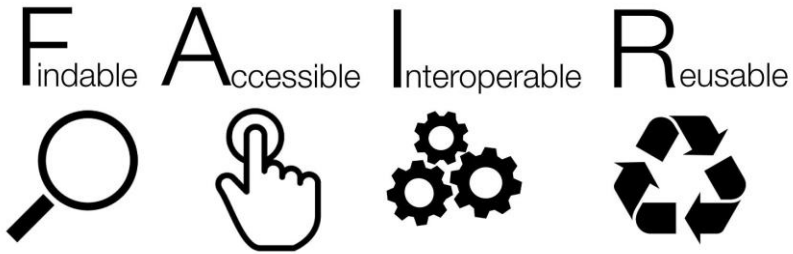
libraries

* Only abstract descriptive!

Motivation: sustainable science

In the ideal world:

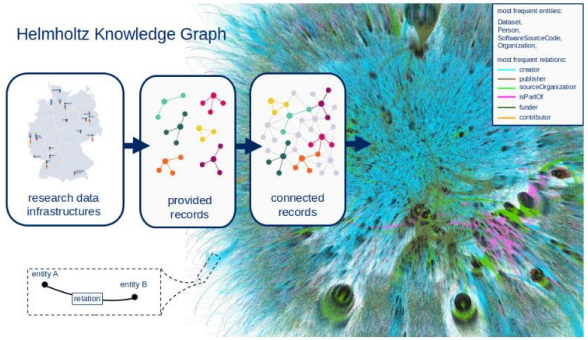
- Human **knowledge accumulates** as best and sustainable as possible. (Knowledge transfer)
- Research is (easily) **reproducible**. (for simulations that should be doable right?)
- Data, Software and Results are **interoperable** and **re-usable**. (drives Progress)



How can you bring your research a bit closer to these ideals?

[1] M. D.Wilkinson. Scientific Data 3 (2016), pp. 1–9., also see www.go-fair.org

unHIDE - unified Helmholtz Information and data exchange



Introduction & Scope

Research across the Helmholtz Association (HGF) depends and thrives on a complex network of inter- and multidisciplinary collaborations which spans across its 18 Centres and beyond.

However, the (meta)data generated through the HGF's research and operations is typically siloed within institutional infrastructure and often within individual teams. The result is that the wealth of the HGF's (meta)data is stored and maintained in a scattered manner, and cannot be used to its full value to scientists, managers, strategists, and policy makers.

To address this challenge, the Helmholtz Metadata Collaboration (HMC) is launching the **unified Helmholtz Information and Data Exchange (unHIDE)**. This initiative seeks to create a lightweight and sustainable interoperability layer to interlink data infrastructures and provide greater, cross-organisational access to the HGF's (meta)data and information assets. Using proven and globally adopted knowledge graph technology (Box 1), unHIDE will develop a comprehensive association-wide Knowledge Graph (KG) the "Helmholtz-KG": a solution to connect (meta)data, information, and knowledge.

> *Box 1*
>
> What is a Knowledge Graph?
>
> - A "graph", from graph theory, is a structure that models pairwise connections between objects using "nodes" connected by "edges".
> - A "knowledge graph" uses such a graph structure to capture knowledge about how a collection of

docs.unhide.helmholtz-metadaten.de

general: hmc@fz-juelich.de
strategy: v.hofmann@fz-juelich.de
tech: j.broeder@fz-juelich.de