Collaborative Graph Data (Structure) Editing

Leon Steinmeier (l.steinmeier@hzdr.de)
Helmholtz Metadata Collaboration

HZDR RSE Meetup November 2023

www.helmholtz-metadaten.de

The Helmholtz Metadata Collaboration

54 People working at **10** centres on the **Road to FAIR**

8 HMC Units at **6** host centres
Aeronautics, Space & Transport @ DLR
Earth & Environment @ Geomar
Energy @ KIT
Health @ DKFZ
Information @ FZJ
Matter @ HZB
FAIR Data Commons @ KIT & FZJ
HMC Office @ Geomar

● Host Centres
◆ Satellites
■ Other Centres

KEY STRATEGIC AREAS

1. Assessing and monitoring the state of FAIR data across Helmholtz
2. Facilitating connectivity of Helmholtz research data
3. Transforming (meta)data recommendations into implementations

GEOMAR

AWI
DESY
Hereon

MDC
GFZ Helmholtz-Zentrum POTSDAM
HZB Helmholtz Zentrum Berlin
HZI

JÜLICH FORSCHUNGSZENTRUM
DLR
DZNE

UFZ
HZDR

GSI

CISPA

dkfz.
KIT Karlsruher Institut für Technologie

Helmholtz Munich

‹HMC›

some spreadsheet

id	intensity	duration [seconds]
...	...	...
...	...	...

some README file

This data set is about some generic measurements.

device: meterbot 2000
operator: Mel (orcid.org/xxx...)
project: ERC 1337

Contemporary data

The problems:

- metadata is often optional and potentially unstructured
- data-metadata connection is usually relatively weak
- no terminology standardization
- no globally unique IDs for entities (e.g. researchers or devices)
- usually no data structure description (a.k.a. schema)
- no domain standards for (meta)data "richness"
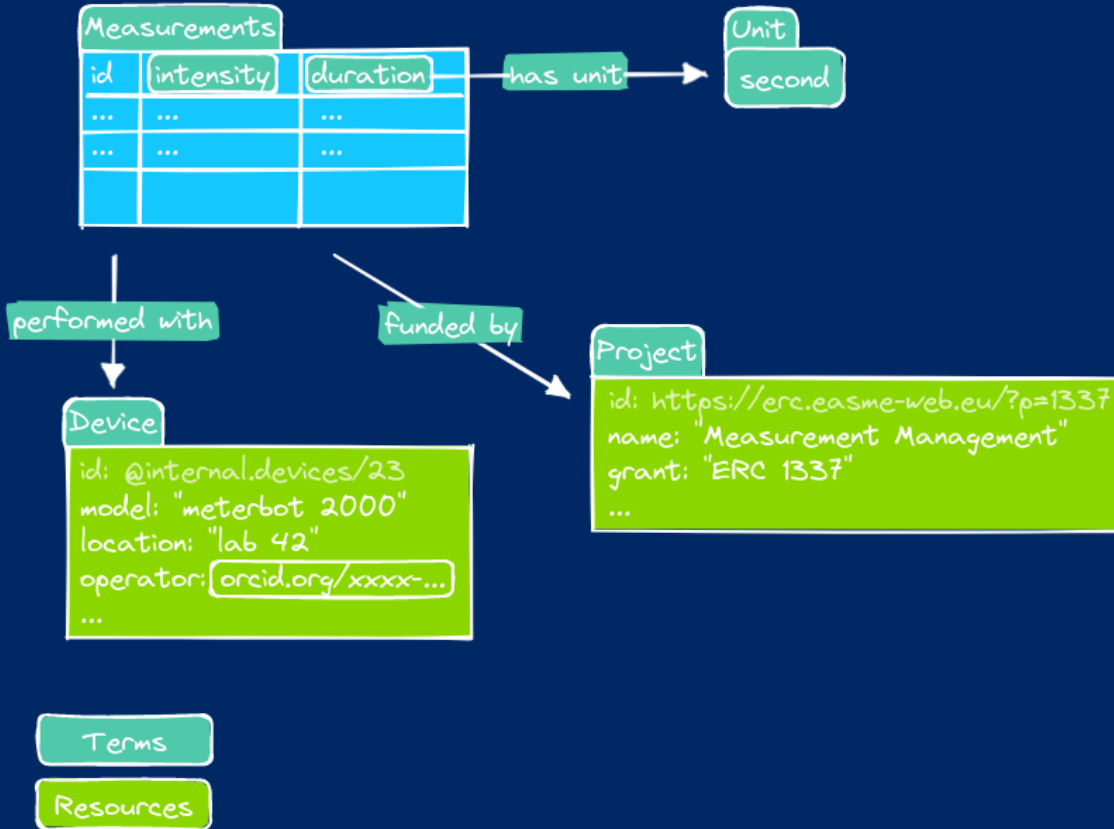
‹HMC›

Perfectly(?) FAIR data:

A semantic graph data set

- links data and metadata
- has an explicit structure
- can easily incorporate resource Ids and metadata
- has documented terminology via ontologies

...but what about

- terminology standardization,
- data structure harmonization,
- and (meta)data "richness"?

Collaborative graph data editing

via

- a graph data editor web app
- editing small "graph data sets" instead of large graphs
- collaboration on GitLab
- generating data entry forms from graph data sets
- publishing graph data sets via GitLab
- automatic, i.e. auto-complete-like suggestions during editing based on all public graph data

graph data set A

main

dev

...

graph data set B

main

feature x

...

...

main

...

GitLab

neo4j

<HMC>
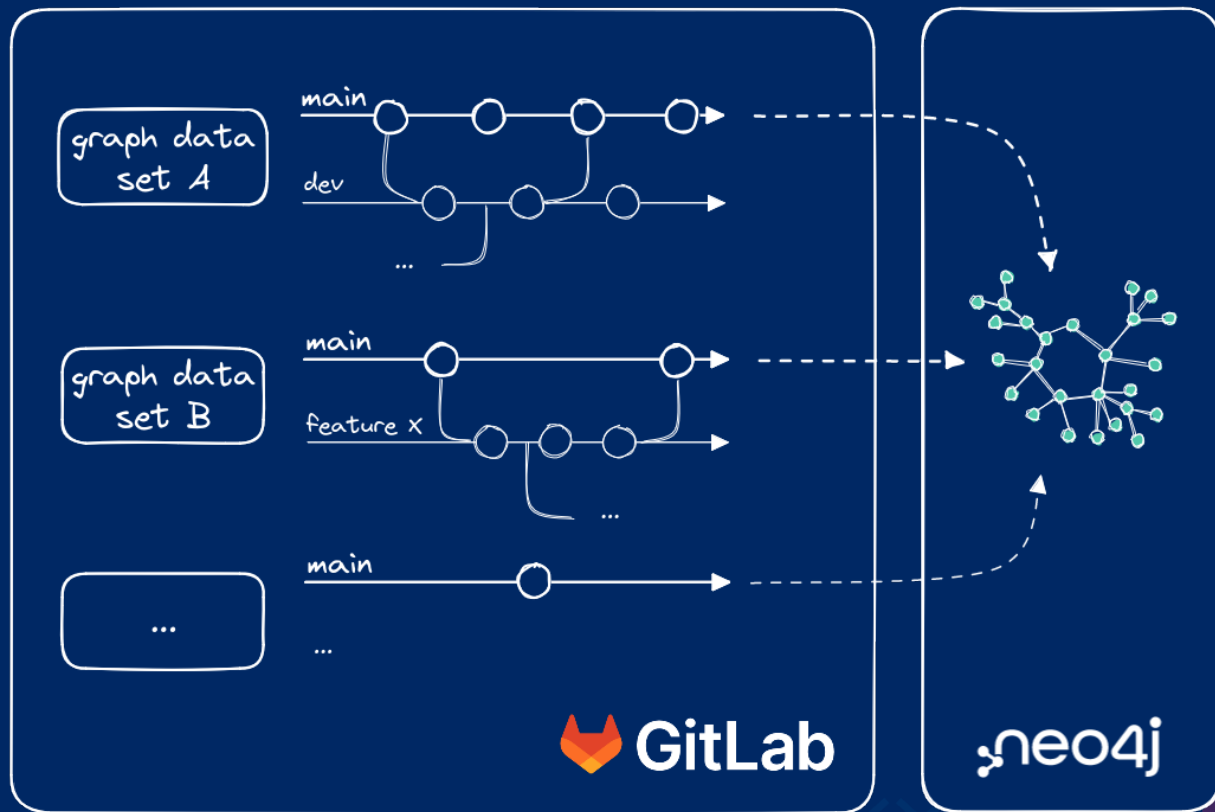
Software Design Challenges

1. Collaboration
   - Which parts of git to show to users and how?
   - Live collaboration (think of e.g. miro or mural)
2. "Auto-complete" suggestions
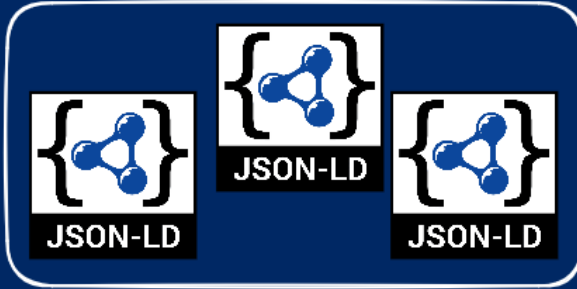   - So many possibilities...
3. One global GitLab
   - Which data/computing to do globally?
   - On which infrastructure?

Thank you for your attention!

Supplemental slides

The tech stack

- node.js backend server
- JSON-LD graph data format
- svelteKit javaScript framework for graph editor web app
- GitLab for version control and project management
- neo4j graph database as a basis for the auto-complete feature
- docker compose for portable deployment

...on its way to FAIR...

E.g. with:
- documented terminology
- resource Ids

some spreadsheet

id	intensity	duration	[seconds]
...	...	...	
...	...	...	

some README file

This data set is about some generic measurements. <

device: meterbot 2000
operator: Mel (orcid.org/xxx...)
project: ERC 1337

Ontology X

id: https://some.ontology/measurement
label: Measurement
synonym: Data Collection
description: Translate an Observation of something into a value.
...

Resource Registry Y

id: https://orcid.org/xxxx-...
name: Melanie
position: Scientist at X Labs
publications:
- ...

Semantics:

Terms — could have unique IDs, definitions, synonyms, etc.

Resources — could have unique IDs and all kinds of other properties

HMC

Semantics/Ontologies

- can also be edited with a graph data editor