HELMHOLTZ METADATA COLLABORATION

HZB Helmholtz Zentrum Berlin

AST | E&E | ENERGY | HEALTH | INFORMATION | MATTER | DATA COMMONS | HMC PROJECTS | HMC OFFICE

PIDs for samples in the PaN community

Helmholtz Metadata Collaboration Hub Matter / Helmholtz-Zentrum Berlin für Materialien und Energie

Oonagh Mannix, Heike Görzig, Rolf Krahl

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Helmholtz Metadata Collaboration

- Make Helmholtz Data **FAIR** - findable, accessible, interoperable and reusable
- Provide a **sustainable** service for efficient metadata handling
- Establish and support a **metadata community** in the respective research areas

FAIR: Findable, Accessible, Interoperable, Reusable

Findable

F1 metadata are assigned a globally unique and eternally persistent identifier.

F2 data are described with rich metadata.

F3 metadata clearly and explicitly include the identifier of the data it describes

F4 metadata are registered or indexed in a searchable resource.

Accessible

A1 metadata are retrievable by their identifier using a standardized communications protocol.

A2 metadata are accessible, even when the data are no longer available

Interoperable

I1 metadata use a formal, accessible, shared and broadly applicable language for knowledge representation.

I2 metadata use vocabularies that follow FAIR principles

I3 metadata include qualified references to other metadata

Reusability

R1 metadata are richly described with a plurality of accurate and relevant attributes.

R2 metadata are released with a clear and accessible data usage license.

R3 metadata are associated with detailed provenance

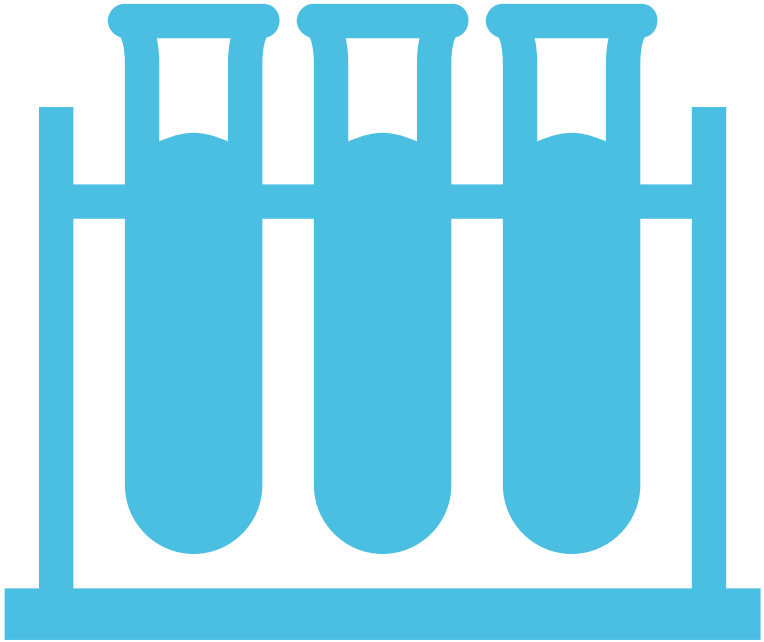
R4 metadata meet domain-relevant community standards.

Wilkenson et al. *Scientific Data* 2016 https://doi.org/10.1038/sdata.2016.18

FAIR and samples

In the literature….

- FAIR principles apply not only to data but to everything that led to the data[1,2]
- This explicitly includes physical samples and analogue artefacts (and their digital representation)
- Physical samples and artefacts can be thought of as data and care needs to be taken to ensure they are FAIR

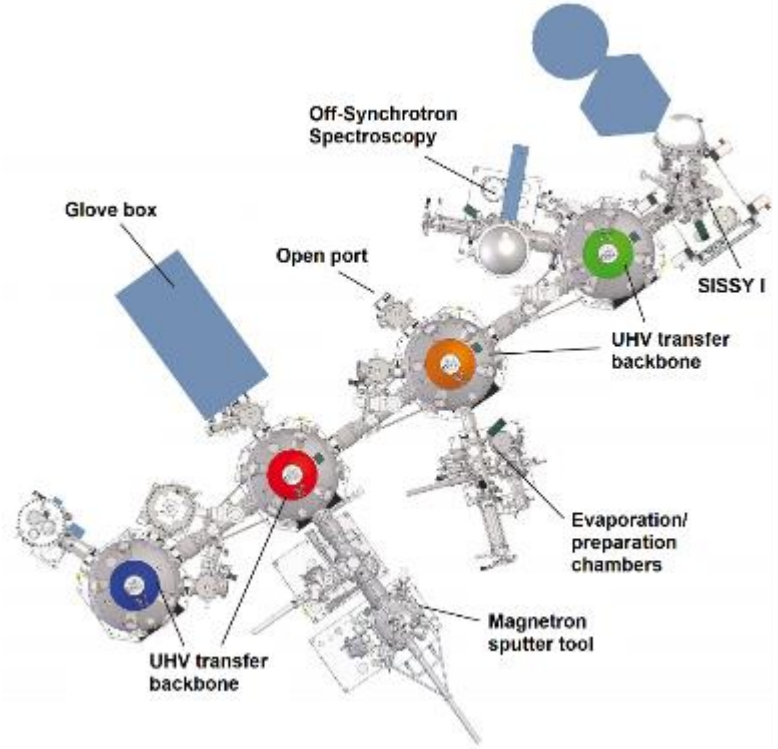


1. The Beijing Declaration on Research Data: https://doi.org/10.5281/zenodo.3552330
2. Plomp *Data Science Journal* 2020 http://doi.org/10.5334/dsj-2020-046

Sample is an entity under investigation

In-depth look at FAIR principles in practice

- Apply FAIR principles to sample workflows
- Zero order solution in place linking experimental workflow and the sample and workflow of the sample
  - Assigned machine identifier in sample database
  - GUI for human identifier
  - Treatment stored in database
  - Potential for further enrichment



Günther et al. *ICALEPCS* 2021

FAIR: Findable, Accessible, Interoperable, Reusable applied to samples

HMC HELMHOLTZ METADATA COLLABORATION

Findable

F1 metadata are assigned a globally unique and eternally persistent identifier. ☒

F2 data are described with rich metadata.

F3 metadata clearly and explicitly include the identifier of the data it describes ☒

F4 metadata are registered or indexed in a searchable resource. ☒

Accessible

A1 metadata are retrievable by their identifier using a standardized communications protocol. ☒

A2 metadata are accessible, even when the data are no longer available ☒

Interoperable

I1 metadata use a formal, accessible, shared and broadly applicable language for knowledge representation.

I2 metadata use vocabularies that follow FAIR principles

I3 metadata include qualified references to other metadata ☒

higher order infrastructure required ->

Sample PID

Wilkenson et al. *Scientific Data* 2016 https://doi.org/10.1038/sdata.2016.18

Sample PIDs and FAIR

Findability

F1 (meta)data are assigned a globally unique and eternally persistent identifier.

F2 data are described with rich metadata.

- **For entities which change their state after (meta)data has been produced (e.g. a physical sample undergoing destructive sampling)**, rich metadata will allow a user to understand (and if possible replicate) the state of a entity before, during, or after (meta)data collection.

F3 metadata clearly and explicitly include the identifier of the data it describes

F4 (meta)data are registered or indexed in a searchable resource.

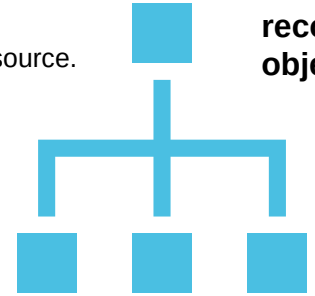
Interoperability

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.

- (Meta)data never occur in isolation, and Findable and Accessible links to other (meta)data - stored in separate records - are typically needed for the Helmholtz community (and wider world) to understand their context and provenance. In this way**, metadata records link together and enrich related digital objects and streams.**

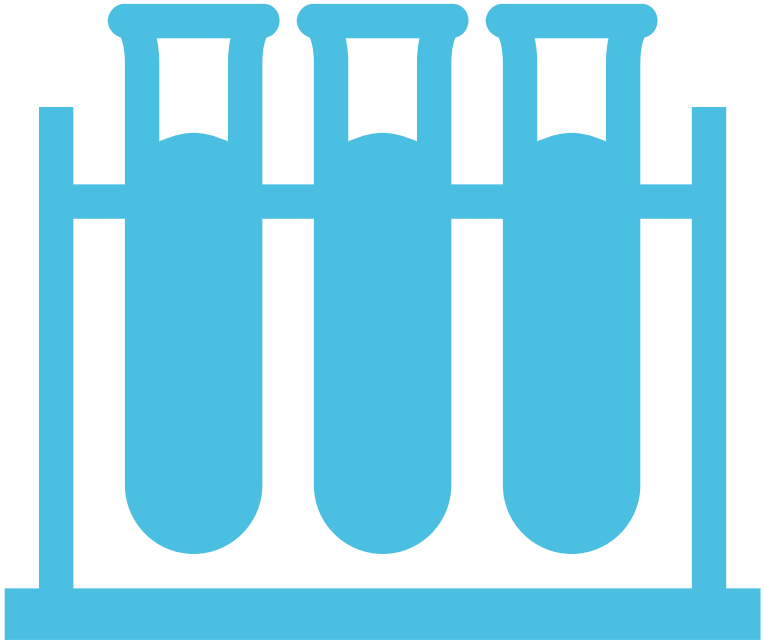


An interpretation of the FAIR principles to guide implementations in the HMC digital ecosystem (2021)

Sample PIDs and PaN facilities

Specific considerations for PaN

- Sample preparation, measurement (at PaN facility) and analysis are often performed by different people or teams
- Measurement can be destructive (radiation damage), or the sample can be ephemeral (pump-probe, high pressure)
- Often without sample information giving meaning to the data is impossible
- PaN experiments are often challenging to reproduce, require peer review to perform, high quality/high interest data
- In PaN special case that often blurry line between sample and instrument

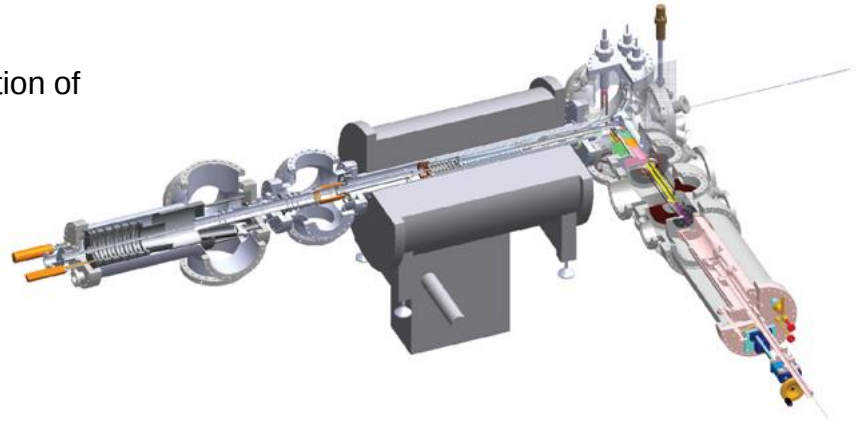


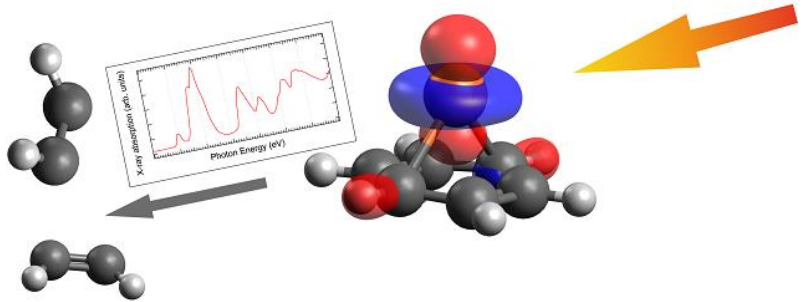
Ion trap: a case of ephemeral sample

Specific consideration of PaN

Sample is created during the experiment and exists for the duration of the experiment.

- Should it have a sample PID?
- Should some other identifier be used?
- What should such a PID look like?

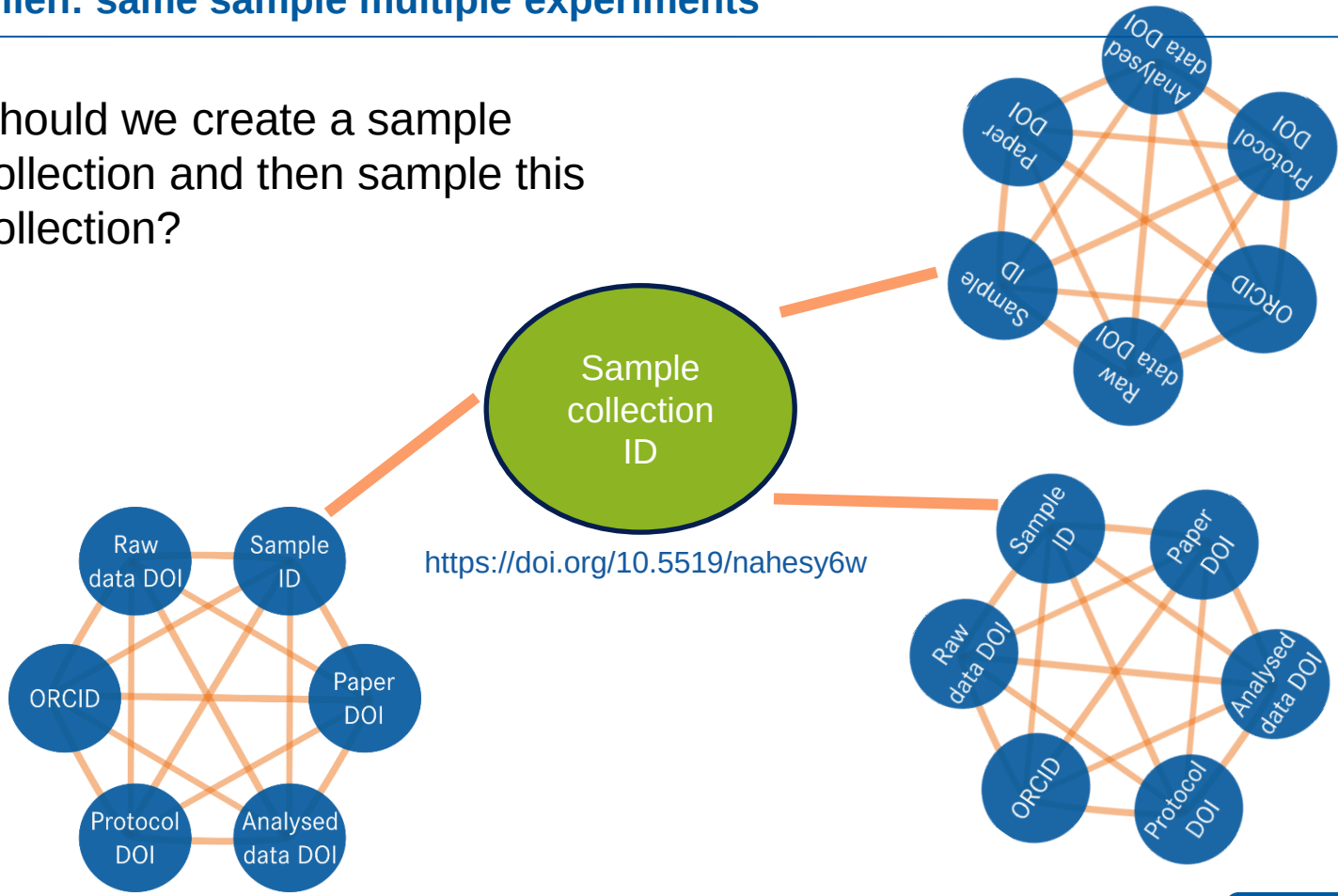




Hirsch et al. *J. Phys. B: At., Mol. Opt. Phys*. 2009 http://dx.doi.org/10.1088/0953-4075/42/15/154029

Pollen: same sample multiple experiments

Should we create a sample collection and then sample this collection?



https://doi.org/10.5519/nahesy6w

Common sample identifier types

- IGSN
  - https://www.igsn.org/
- Research resource identification (RRID)
  - https://www.rrids.org/
- Digital object identifier (DOI)
  - https://www.doi.org
- RDA 23 things physical samples
  - https://docs.google.com/document/d/1vzWIX77WC1rkGQOligPwCrqLCklsQx2FFwo7O7S8H-8/edit
- Extensive list here:
  - Damerow et al. *Data Science Journal*, 2021 http://doi.org/10.5334/dsj-2021-011

IGSN

- Metadata schema initially developed for geographical samples; looking to extend to other disciplines
- FAIR workflows to establish IGSN for samples in the Helmholtz association through the FAIR Wish project

Projects working on sample identifiers

- Digital LEAPS
  - Proposal: Surveying technology for advancing remote services (STARS)
  - Proposed starting date Dec 1st 2021
  - Contact Klaus Kiefer: klaus.kiefer@helmholtz-berlin.de
- DAPHNE NFDI consortium
  - https://www.daphne4nfdi.de/
  - Metadata schemata, data formats and sample descriptions used at the participating institutions are comprehensively documented A basic sample persistent identifier capability for cases where it is not possible/sensible to use sample PIDs developed elsewhere has been deployed
- Distributed system of scientific collections (DISSCO) project
  - https://www.dissco.eu/
- iSamples
  - https://isamplesorg.github.io/home/
  - Towards an interdisciplinary cyberinfrastructure for materials samples
- Physical samples and collections in the research data ecosystem RDA IG
  - https://www.rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig

Summary

- Samples should be considered as a first-class citizen in discussions of FAIR data implementation.
- Infrastructure is required to achieve this.
- Persistent identifiers for samples are central here.
- The PaN community has specific needs here.
- Wider community is mobilizing to act – good time to get involved!