HMC Hub Matter

Assessing the FAIRness of a prototypical PaN instrument

at BESSY II
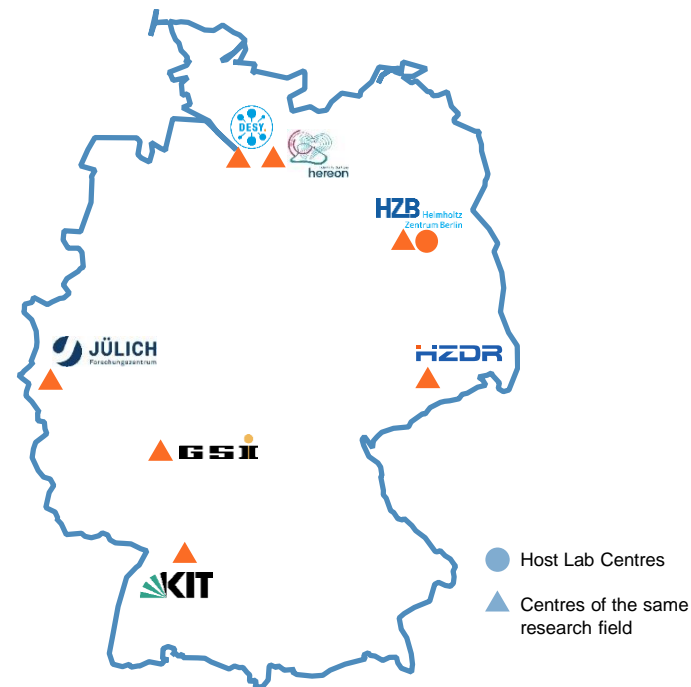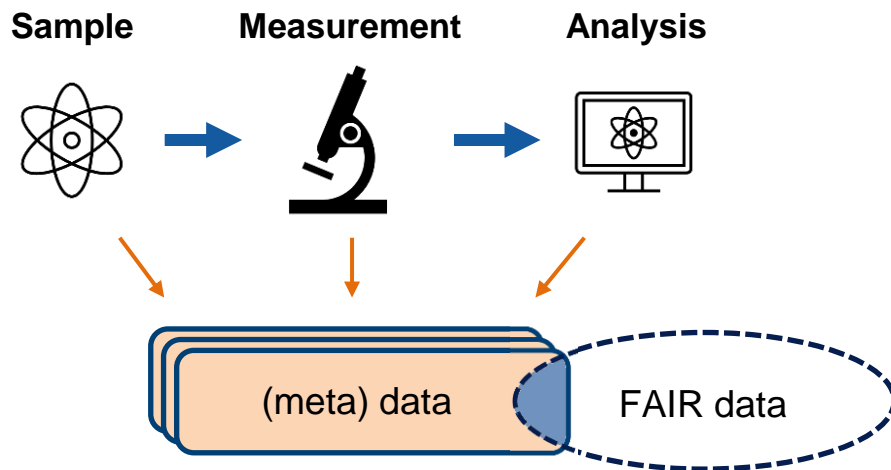
Markus Kubin and Gerrit Günther

Helmholtz Metadata Collaboration (HMC)

- Make Helmholtz Data **FAIR** - findable, accessible, interoperable and reusable

- Provide services for **sustainable** and efficient metadata handling

- Develop, share and **consolidate community-expertise** in metadata across Helmholtz

Research field matter:

- Particles, plasma physics, materials, biophysics …

- Technology of accelerators, detectors ...
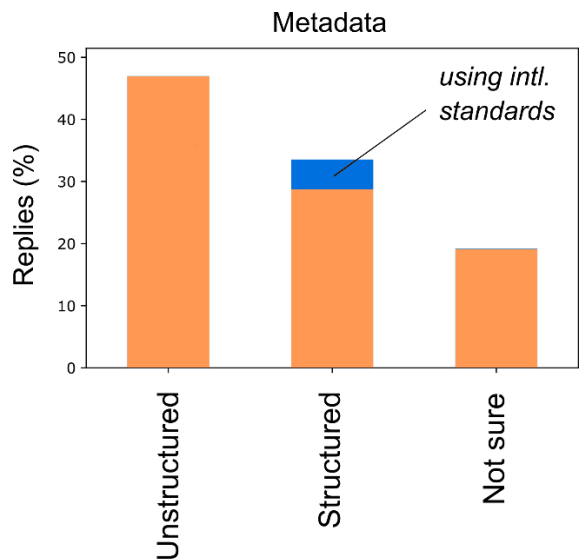
- User programs of photon, neutron, ion beam facilities …

Sample **Measurement** **Analysis**

(meta) data   FAIR data

Get in touch…

https://helmholtz-metadaten.de/en/matter/contact-us

● Host Lab Centres

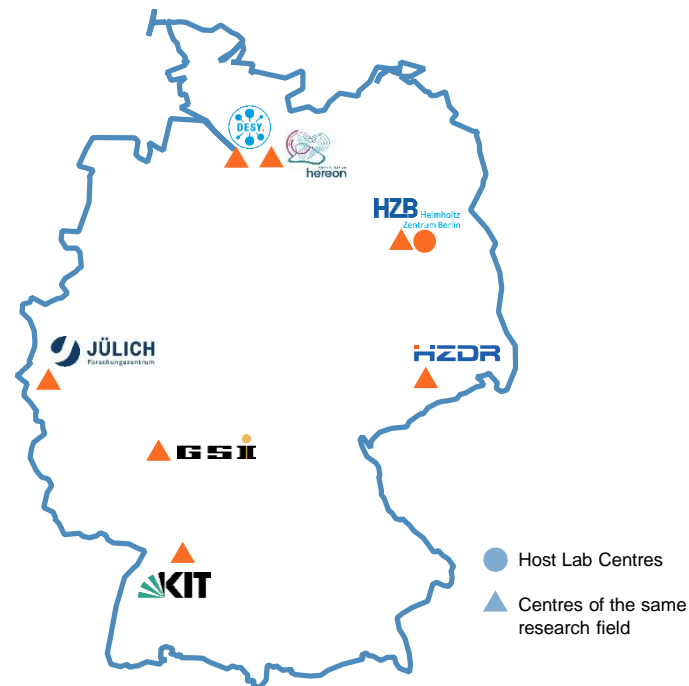▲ Centres of the same research field

Status of the community

Research field matter:

- Community survey 2021
    - Researchers in all Helmholtz centres / research fields
    - Survey still running

- First data (research field matter, n=104)

Metadata

using intl. standards

Replies (%)

Unstructured   Structured   Not sure

Will be published soon …

Host Lab Centres

Centres of the same research field

Get in touch…

https://helmholtz-metadaten.de/en/matter/contact-us

From FAIR guidelines

to FAIR indicators

Findable

Accessible

Interoperable

Reusable

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

FAIR Guiding Principles (2016)

M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.

Findable:

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

F2. data are described with **rich metadata** (defined by R1 below)

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

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

Accessible:

A1. (meta)data are **retrievable** by their **identifier** using a **standardized communications protocol**

A1.1 the protocol is **open, free, and universally implementable**

A1.2 the protocol allows for an **authentication and authorization** procedure, where necessary

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

Interoperable:

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

Reusable:

R1. meta(data) are **richly described** with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data **usage license**

R1.2. (meta)data are associated with **detailed provenance**

R1.3. (meta)data meet domain-relevant **community standards**

M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.

From 15 FAIR Guidelines (2016) **to 41 FAIR Indicators** (2020)

Example:

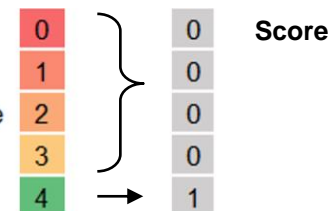**Guideline:**   F1. (Meta)data are assigned a globally unique and persistent ID

Indicators:

FAIR	ID	Indicator
F1	RDA-F1-01M	Metadata is identified by a persistent identifier
F1	RDA-F1-01D	Data is identified by a persistent identifier
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier
F1	RDA-F1-02D	Data is identified by a globally unique identifier

Rating: Essential *** – Important ** – Useful *
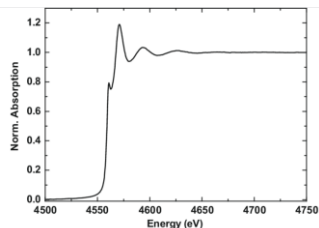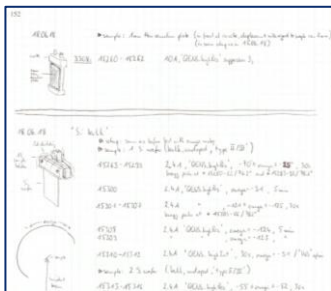
Metric:
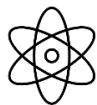
		Score
0 – not applicable	0	0
1 – not being considered this yet	1	0
2 – under consideration or in planning phase	2	0
3 – in implementation phase	3	0
4 – fully implemented	4 →	1

FAIR Data Maturity Model
Specification and Guidelines
2020

FAIR PRINCIPLES

Proposed RDA Recommendation
Produced by: **FAIR Data Maturity Model WG, 2019-2020**
https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

https://zenodo.org/record/3909563

A prototypical

PaN instrument at HZB / BESSY II

A prototypcial PaN instrument at HZB / BESSY II

Data collection

Automatic (meta)data collection:

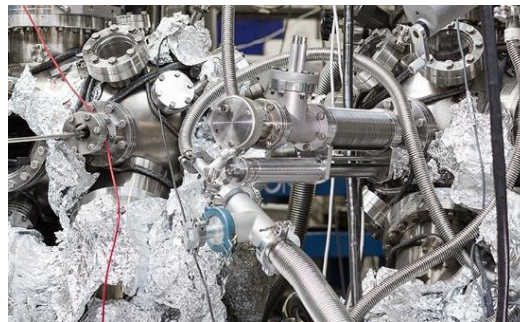- Metadata of sensors and motors etc.
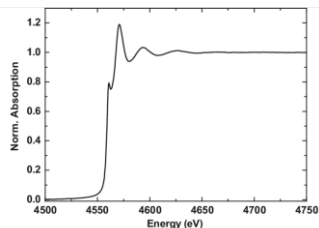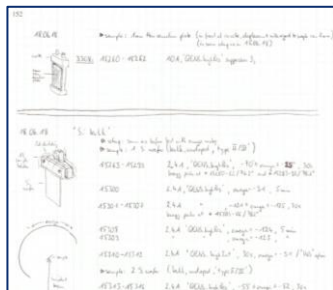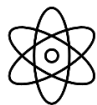- Detector data
- …

→ ASCII text files
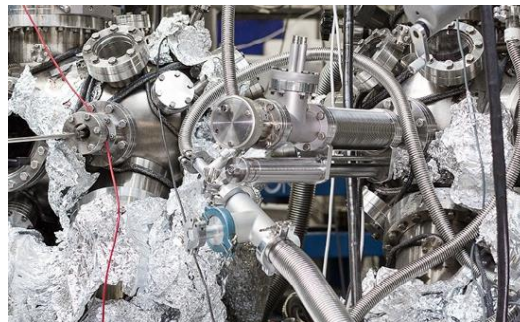
```
File  Edit  Format  View  Help
#    verticalh5 = 782,2000
#    poselliph1 = 113,530
#    poselliph3 = 243,1400
#    negelliph1 = 113,530
#    negelliph3 = 253,1300
#    negelliph5 = 703,1710
#    poselliph5 = 701,1710
# DIO-6612
#    id = 1
#    model = NI6612
#    vipath = Lise\DeviceNI6612\DeviceNI6612.vi
#    pxislot = 2
#    sim = false
#    activation = true
#    frontpanel = true
# Quantum Composer 9530
#    id = 2
#    model = Quantum9530
#    vipath = Lise\DeviceQuantum9530\DeviceQuantum9530.vi
#    sim = false
#    setconfig = false
#    frontpanel = true
#    type = serial
#    port = COM9
```
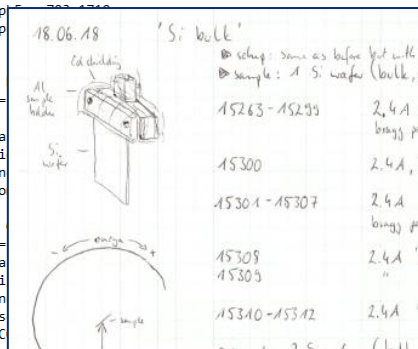
A prototypcial PaN instrument at HZB / BESSY II

Data collection

light

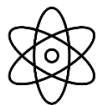## Automatic (meta)data collection:

- Metadata of sensors and motors etc.
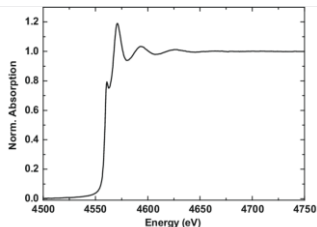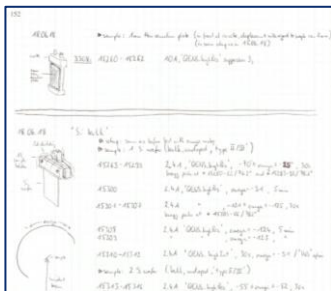- Detector data
- …

→ ASCII text files

Manual (meta)data collection:

- Sample information
- Manual instrument settings
- Calibration measurements
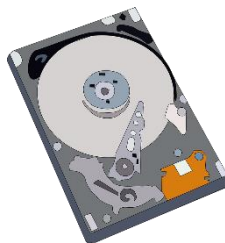- Background measurements
- …

→ Paper logbook

Data collection

light

Data processing

(group's share drive)

Conversion
to **NeXus** format

3.3.2.5. NXfluo

Status:

    application d...

Description:

    This is an ap...

Symbols:

    The symbol(...

    nE: Number...

Groups cited:

    NXdata, NXdetector, NXentry, NXinstrument, NXmonitor, NXmonochromator, N

Structure:

    entry: (required) NXentry

        title: (required) NX_CHAR

        start_time: (required) NX_DATE_TIME

        definition: (required) NX_CHAR

            Official NeXus NXDL schema to which this file conforms.

            Obligatory value: NXfluo

    INSTRUMENT: (required) NXinstrument

        SOURCE: (required) NXsource
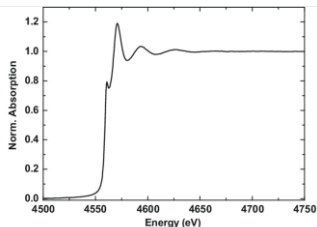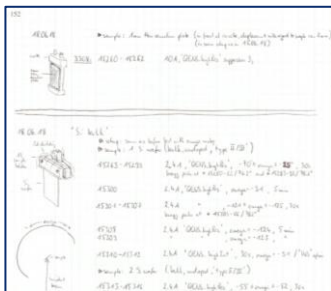
HDF5 / NeXusFile Format:

+ Hierarchical (meta)data

+ NeXus definition language

+ Persistent web description (definition/URL)

→ Machine readable / interoperable

A prototypcial PaN instrument at HZB / BESSY II

Data collection
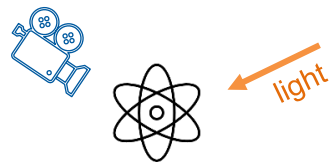
Data processing

(group's share drive)
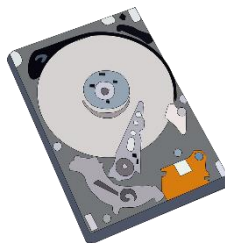
Conversion to **NeXus** format

Data repository: ICAT

Data publication

(manual PIDs)

http://doi.org/10.5442/ND000007 (example)

FAIR assessment

of a prototypical

PaN instrument at HZB / BESSY II

HMC | HELMHOLTZ METADATA COLLABORATION

Requirements

- Findable in metadata registries

    - with rich metadata

    - by humans via the web

    - by machines via API

- (Meta)data is indexed by PIDs

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

F2. data are described with **rich metadata** (defined by R1 below)

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

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

FINDABLE

RDA-F1-01M, RDA-F1-01D, RDA-F1-02M, RDA-F1-02D, RDA-F2-01M, RDA-F3-01M, RDA-F4-01M
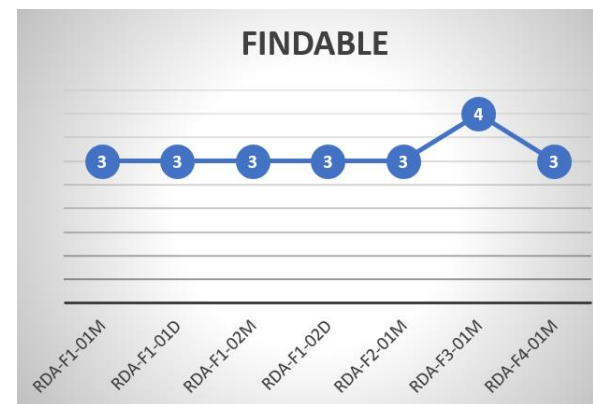
ICAT repository

- **Discovery metadata** in the ICAT catalogue

- **Embargo**ed (meta)data only for **authorized** users

- **No PID assignment** to (meta)data in the ICAT, yet

- PID records of **manual data publications** with DOI

NeXus format

- n.a.

Workflow at the instrument

- n.a.

Steps to improve Findability

- **Automatic assignment of PIDs** to data in the ICAT (w.i.p.)

- **Enable harvesting** by higher level services (B2FIND) (w.i.p.)

- Improve discovery in the ICAT by **specific metadata** (e.g., of **samples**)

Requirements

Accessible

- manually via the web

- via standardized protocols

- by resolving PIDs

- authentication
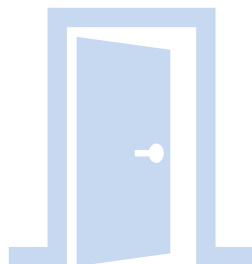
A1. (meta)data are **retrievable** by their **identifier** using a **standardized communications protocol**

A1.1 the protocol is **open, free, and universally implementable**

A1.2 the protocol allows for an **authentication and authorization** procedure, where necessary

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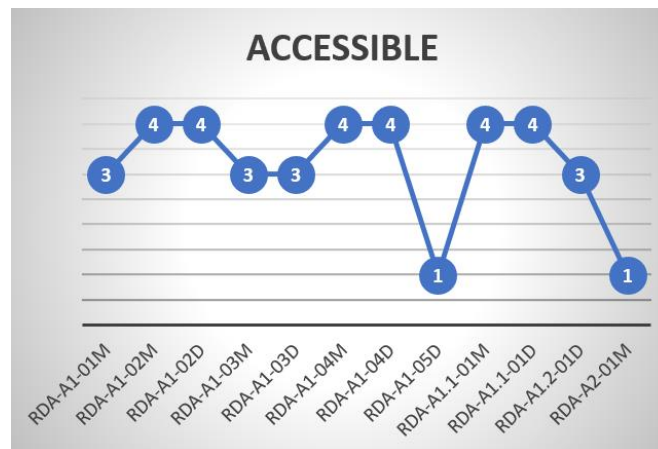
Accessible

ICAT repository

- Manual access to (meta)data via **ICAT web interface**

- Machine accessible via **https protocol**

- **Embargo**ed (meta)data only for **authorized** users

- **Authentication** with HZB-internal credentials (Keycloak authentication is w.i.p.)

NeXus format

- n.a.

Workflow at the instrument

- n.a.

Steps to improve Accessibility

- PIDs to resolve to a **metadata record** with access information

- Find a protocol to access **data from tape**

- Need option of **tomb stone records** (deleting data is currently not considered, yet)

Requirements

- Standards

- (FAIR) vocabularies and mappings

- Qualified references

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
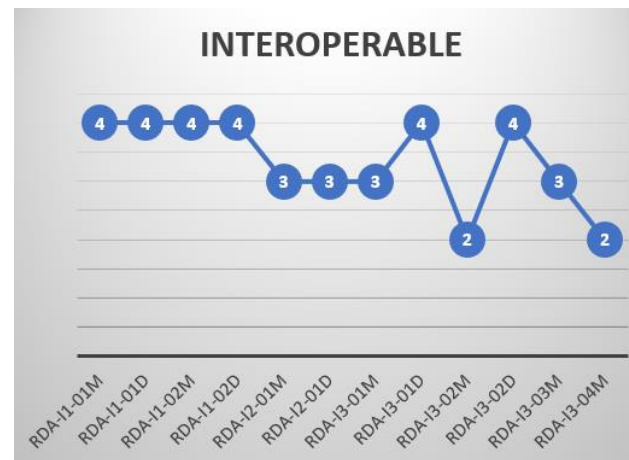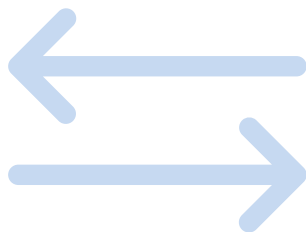
ICAT repository

- **ICAT schema** for discovery metadata (community standard)

- PID records of **manual data publications** use **DataCite** (cross-community standard)

NeXus format

- (Meta)data in **NeXus format** (community standard)

- **NeXus Definition Language** is machine-readable

- **Qualified references** to ORCID and instrument PIDs

Workflow at the instrument

- n.a.

Steps to improve Interoperability

- **Add qualified references** to external data (calibration data) in the ICAT / NeXus / PID record

- Add **NeXus terms** (sample info, instrument PIDs)

Reusable

Requirements

- Rich metadata

- Provenance

- Standards

- Licences
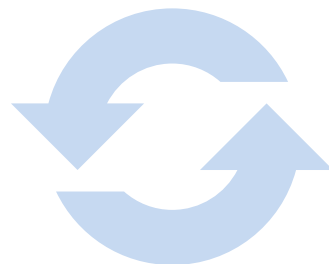
R1. meta(data) are **richly described** with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data **usage license**

R1.2. (meta)data are associated with **detailed provenance**

R1.3. (meta)data meet domain-relevant **community standards**
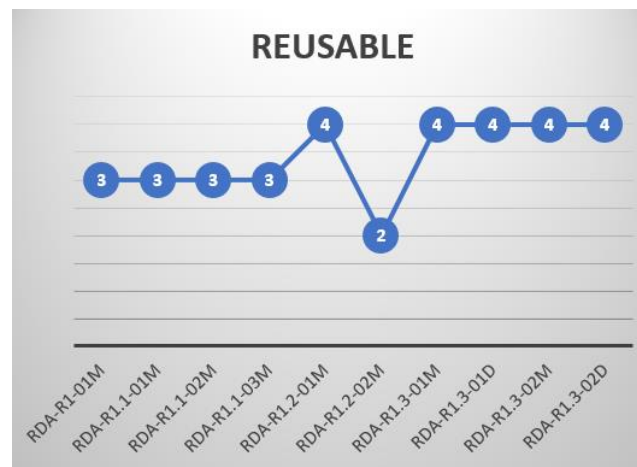
REUSABLE

Reusable

ICAT repository

- n.a.

NeXus format

- (Meta)data in **NeXus format** (community standard)

- Rich **metadata** for reusability

- **Lack of sample** + **calibration** metadata may prevent reuse

- **Minimal provenance** information in NeXus standard (contact, instrument, dates)

Workflow at the instrument

- (Meta)data **automatically ingested** to NeXus + ICAT

- Workflow needed to **reference samples** (PIDs)

- Manual **input via GUI** before writing NeXus files (sample, contact)

Steps to improve Reusability

- **Add** (meta)data on **sample (PIDs) and calibration**

- Implement **e-logbooks** to provide context

- Include **provenance** in cross-community **standard**

- Include links to machine-readable **reuse license**

Overall FAIR Metric

What level describes your digital object?

Level 0	Not FAIR
Level 1	FAIR essential criteria only
Level 2	FAIR essential criteria + 50 % of important criteria
Level 3	FAIR essential criteria + 100% of important criteria
Level 4	FAIR essential criteria + 100% of important criteria + 50% of useful criteria
Level 5	FAIR essential criteria + 100% of important criteria + 100% of useful criteria

FINDABLE   ACCESSIBLE   INTEROPERABLE   REUSABLE

Based on RDA FDMM WG's assessment sheet https://zenodo.org/record/3909563

Towards meeting FAIR criteria – work in progress

Findability:

- Discovery metadata
- Automatic assignment of PIDs
- Connection to higher-level services

Accessibility:

- (meta)data in ICAT repository
- Authentication / authorization

Interoperability:

- NeXus format
- ICAT schema (mappings w.i.p.)

Reusability:

- Sample + calibration (meta)data
- License metadata

Overview of available metadata vs. ExPaNDS D2.2 Draft recommendations

‹HMC› HELMHOLTZ METADATA COLLABORATION

Proposal

- principal investigator ✓
- co-investigators ✓
- instrument requested / used ✓
- sample description ✓
- facility where proposal is submitted
- proposal identifier ✓
- experiment description ✓
- proposed experiment conditions

Experiment

- visiting experimental team ✓
- experiment / measurement dates ✓
- sample information ✓
- instrument information ✓
- calibration information
- produced dataset information ✓

Processing

- data format after processing ✓
- processing information
- processing software packages
- original data link used for processing
- resulting dataset information ✓

Analysis

- resulting data format of the analysis
- file identifiers
- software package used for analysis
- original data link used for analysis
- resulting dataset information

Record

- resource identity ✓
- creator of the record ✓
- publisher of the record ✓
- publication year ✓
- release date (end of embargo) ✓
- title of the dataset ✓
- license for usage ✓

_____

- + funding information ✓

Manually curated data publication

ExPaNDS D2.2: Draft recommendations for FAIR PaN Data Management https://zenodo.org/record/4312825

Lessons learned

- Use of **standards** and **PIDs**

- Degree of **granularity** is critical

- **Level of metadata** addressed by indicator

- **Manual** assessment can help to identify

  - **gaps** in workflows of specific instruments

  - general **community needs**

  - **community**-specific, essential **metadata**

  - **key activities** to address community needs

Outlook

- **Transfer** to other PaN instruments in Helmholtz

- **Communication** with consortia (NIAC / NeXus)

- **Comparison** with other approaches (e.g., F-UJI)

  https://www.f-uji.net
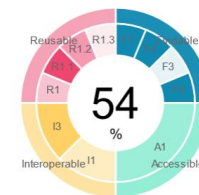
(example)

Community needs

- Facilities and scientists **need support** to go FAIR

- Devise concrete **tools, services and advice**

Thank you

Get in touch…

markus.kubin@helmholtz-berlin.de

gerrit.guenther@helmholtz-berlin.de

Group page

https://helmholtz-metadaten.de/en/matter/contact-us

Twitter

@helmholtz_hmc / #HMCMatter

Acknowledgements:

Heike Görzig, Rolf Krahl, Oonagh Mannix, Luigia Cristiano

Possible questions to be discussed ...

General
- **Which metadata** to be included **at which level**? (NeXus file-level / repository / PID record)
- **Granularity: Minimum set of metadata** required for different indicators and levels of metadata
- **Community-specific** recommendations, checklists, **good practices**

Findable
- **Author information** needs manual **curation** (DFG guidelines of good scientific practice)
- Minimal **threshold for "rich" metadata** (community specific)

Accessible
- Findability and Accessibility of (meta)data for **embargoed data**
- Generalized **authentication**
- Access protocol for **data on tape**

Interoperable
- **Mappings** of NeXus and ICAT Schema
- Criteria for **lower level interoperability** (data structures, formats etc)

Reusable
- How to deal with (legacy) metadata in **paper labbooks**?
- **At which level** should metadata hold **license** information and **provenance** information?
- Minimum **metadata for provenance**

Supporting Slides

FAIR	ID	Indicator	Priority	Level
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	Essential	3
F1	RDA-F1-01D	Data is identified by a persistent identifier	Essential	3
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	Essential	3
F1	RDA-F1-02D	Data is identified by a globally unique identifier	Essential	3
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	Essential	3
F3	RDA-F3-01M	Metadata includes the identifier for the data	Essential	4
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	Essential	3

1 – not applicable
2 – not being considered this yet
3 – under consideration or in planning phase
4 – in implementation phase
5 – fully implemented

Accessible   Assessment Details

FAIR	ID	Indicator	Priority	Level
A1	RDA-A1-01M	Metadata contains information to get access to the data	Important	3
A1	RDA-A1-02M	Metadata can be accessed manually	Essential	4
A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	Essential	4
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	Essential	3
A1	RDA-A1-03D	Data identifier resolves to a digital object	Essential	3
A1	RDA-A1-04M	Metadata is accessed through standardised protocol	Essential	4
A1	RDA-A1-04D	Data is accessible through standardised protocol	Essential	4
A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	Important	1
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	Essential	4
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	Important	4
A1.2	RDA-A1.2-01D	Access protocol supports authentication and authorization	Useful	3
A2	RDA-A2-01M	Metadata remains available after data is no longer available	Essential	1

Interoperable   Assessment Details

FAIR	ID	Indicator	Priority	Level
I1	RDA-I1-01M	Metadata uses representation in standardized format	Important	4
I1	RDA-I1-01D	Data uses representation in standardised format	Important	4
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	Important	4
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	Important	4
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	Important	3
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	Useful	3
I3	RDA-I3-01M	Metadata includes references to other metadata	Important	3
I3	RDA-I3-01D	Data includes references to other data	Useful	4
I3	RDA-I3-02M	Metadata includes references to other data	Useful	2
I3	RDA-I3-02D	Data includes qualified references to other data	Useful	4
I3	RDA-I3-03M	Metadata includes qualified references to other metadata	Important	3
I3	RDA-I3-04M	Metadata include qualified references to other data	Useful	2

FAIR	ID	Indicator	Priority	Level
R1	RDA-R1-01M	Plurality of accurate and relevant attributes allow reuse	Essential	3
R1.1	RDA-R1.1-01M	Metadata includes information about the licence for reuse	Essential	3
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	Important	3
R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	Important	3
R1.2	RDA-R1.2-01M	Metadata has provenance information in community standards	Important	4
R1.2	RDA-R1.2-02M	Metadata has provenance information in cross-community language	Useful	2
R1.3	RDA-R1.3-01M	Metadata complies with a community standard	Essential	4
R1.3	RDA-R1.3-01D	Data complies with a community standard	Essential	4
R1.3	RDA-R1.3-02M	Metadata is in machine-understandable community standard	Essential	4
R1.3	RDA-R1.3-02D	Data is in machine-understandable community standard	Important	4