FAIR
Photon and Neutron (PaN) data

ORSO, 21 June 2023,
Heike Görzig

heike.goerzig@helmholtz-berlin.de

FAIR data principles

https://www.go-fair.org/fair-principles/

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

Accessible

Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Findability and accessibility

- Data repositories with data catalogs

https://data.esrf.fr/public
https://rodare.hzdr.de/record/2311
https://discovery.psi.ch/datasets

https://data.panosc.eu/search/?q=%27%27&facility=PSI&technique=tomography
https://b2find.eudat.eu/dataset/?groups=pans

https://zenodo.org/record/5760882

Interoperability, Reusability

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for

Use file and data formats. Formats, variables, and data structures that others understand and can be used with standard software and workflows.

for

ow

es to

other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Enough information that third parties, after a long period of time can use the data (also for things you have not thought about).

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

EOSC ExPaNDS
(11.2019-2.2023)

ExPaNDS WP2 Deliverables

Draft Recommendations for FAIR Photon and Neutron Data Management (Dec 2020)

https://doi.org/10.5281/zenodo.4312825

Final Recommendations for FAIR Photon and Neutron Data Management (Mai 2022)

https://doi.org/10.5281/10.5281/zenodo.6821676

Used RDA FAIR Data Maturity Model indicators to prioritise metadata for the 4 elements of FAIR (i.e. F,A,I,R) …

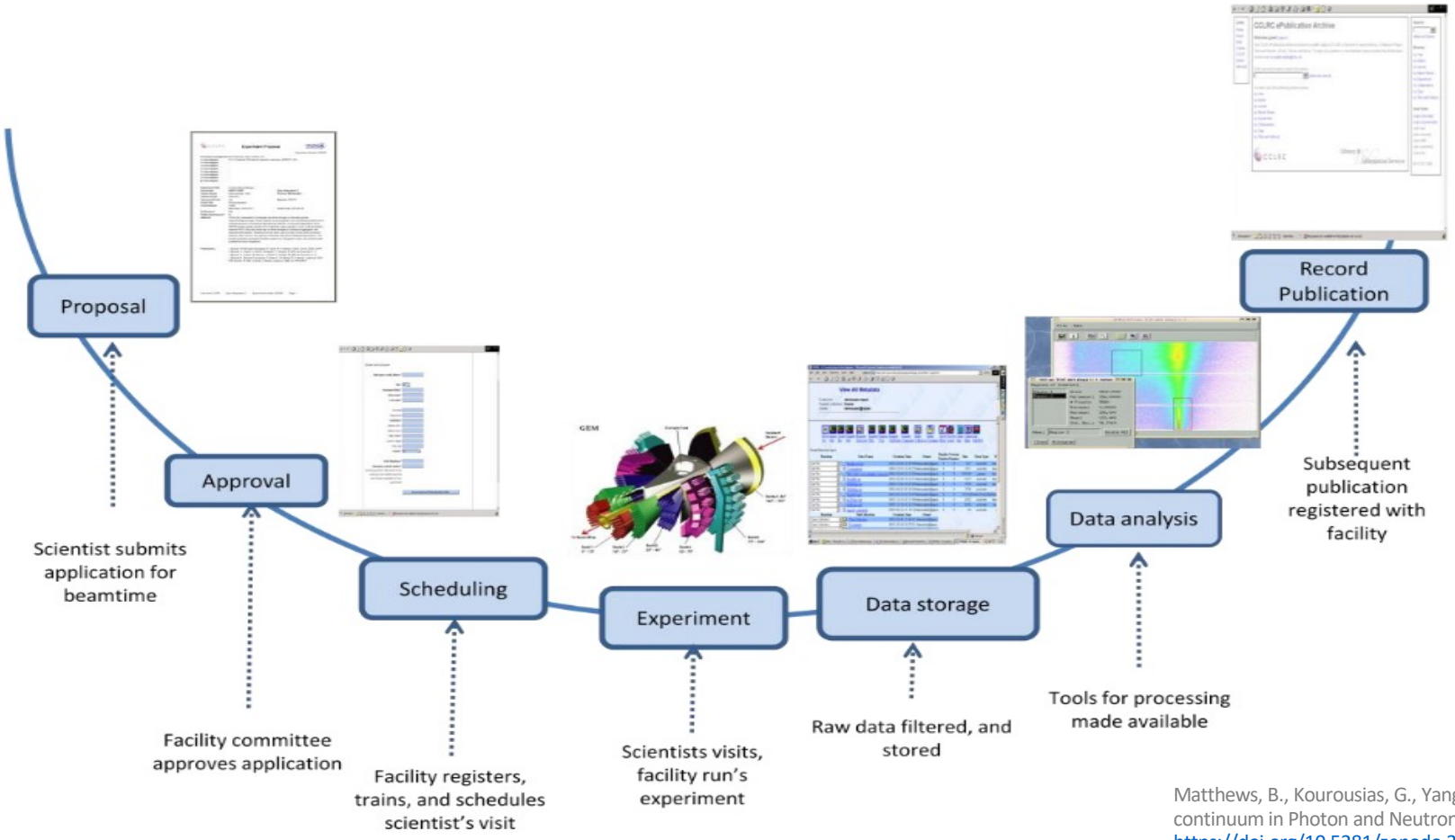
- **Essential:** such an indicator addresses an aspect of the utmost importance to achieve FAIRness under most circumstances, or, conversely, FAIRness would be practically impossible to achieve if the indicator were not satisfied.

- **Important:** such an indicator addresses an aspect that might not be of the utmost importance under specific circumstances, but its satisfaction, if at all possible, would substantially increase FAIRness.

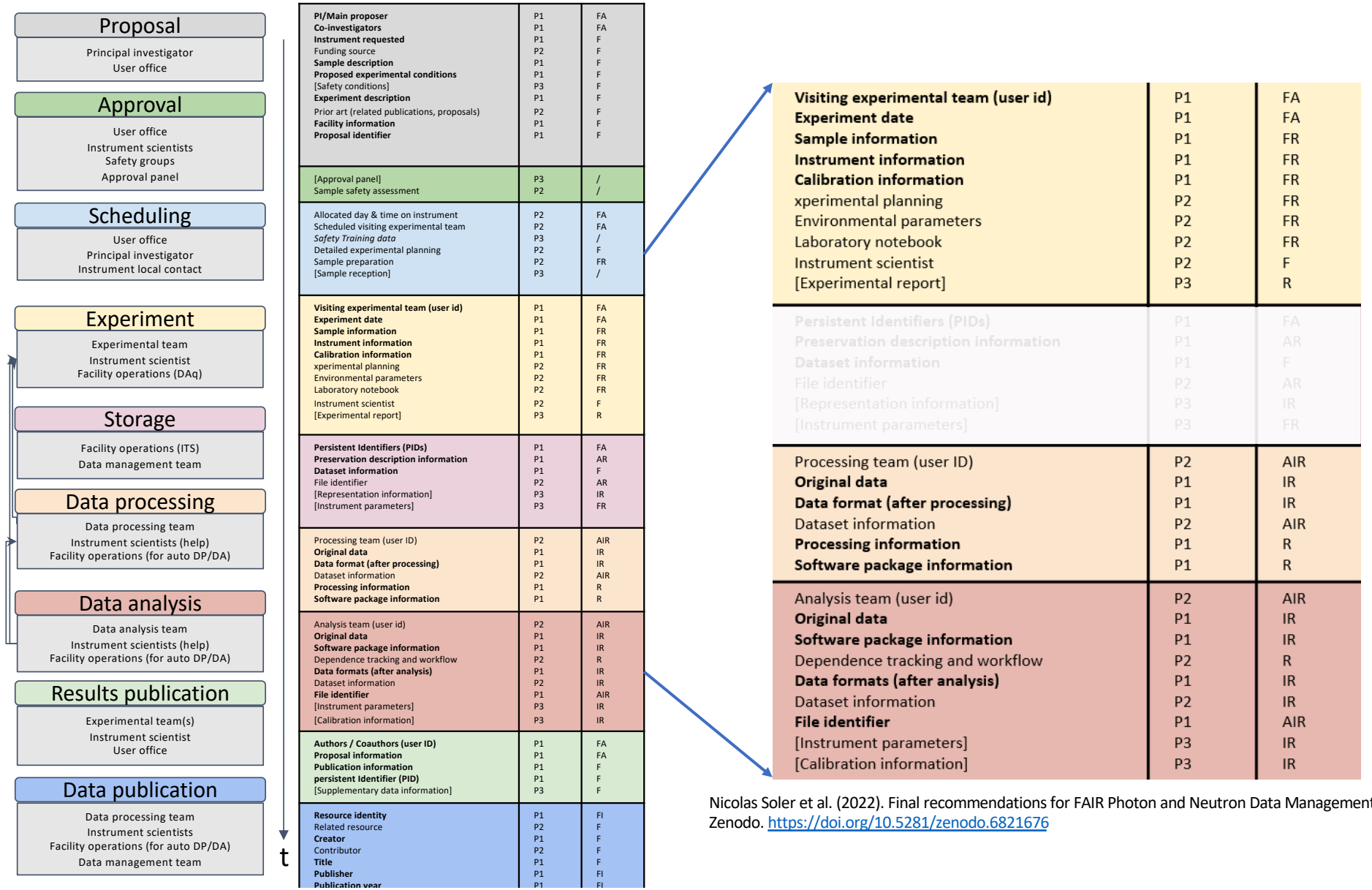
- **Useful**: such an indicator addresses an aspect that is nice-to-have but is not necessarily indispensable to achieve FAIRness

Bahim, C., Casorrán-Amilburu, C., Dekkers, M. et al (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Science Journal, 19(1): 41
http://doi.org/10.5334/dsj-2020-041

(Idealised) PaN Experimental Lifecycle



Matthews, B., Kourousias, G., Yang, E., Griffin, T. (2012). Model of the data continuum in Photon and Neutron Facilities. PaN-data ODI Deliverable 6.1. https://doi.org/10.5281/zenodo.3897910



Nicolas Soler et al. (2022). Final recommendations for FAIR Photon and Neutron Data Management (FINAL). Zenodo. https://doi.org/10.5281/zenodo.6821676

Metadata for experimental data

Metadata record	Prio.	Asp.	Type	Source
Visiting experimental team (user id)	P1	FA	administrative	user office
Experiment date	P1	FA	administrative	user office
Sample information	P1	FR	scientific:sample	sample DB or user
Instrument information	P1	FR	scientific:beamline setup	data acq. and control system
Calibration information	P1	FR	scientific:beamline setup	beamline scientist or user
Experimental planning	P2	FR	scientific:context	principal investigator
Environmental parameters	P2	FR	scientific:physical	data acq. and control system
Laboratory notebook	P2	FR	scientific:context, sample	user
Instrument scientist	P2	F	administrative	user office
[Experimental report]	P3	R	scientific: context	user

- Visiting experimental team
  - E.g. NeXus roles: “local_contact”, “principal_investigator”, and “proposer”
  - ORCID, name, affiliation, …
- Sample information (suggestion)
  - What it is, is under discussion
  - In measurement file relevant information for basic understanding of the dataset
  - PID refers to sample synthesis and processing workflow
- Instrument information
  - PID, name, hosting institution
- Calibration information
  - Essential calibration information for processing the data
  - Depends very much on the measurement technique
  - Other calibration information might be useful in other context
  - Refer to calibration information via PID or directly in file

Example RIXS discussion

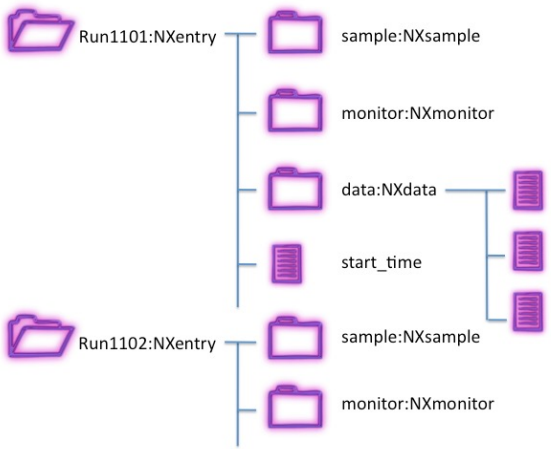
- Calibration files are required for data analysis
  - File location:
  - PID, path to data in file
- Detector pixel size
- Sample can be anything:
  - Lable, description, and PID
  - → we can't do much here

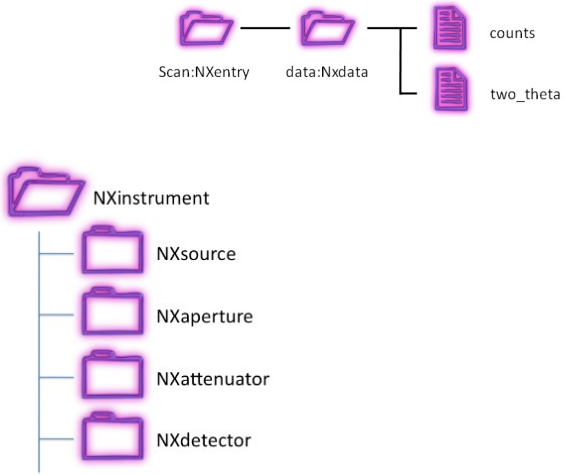
NeXus experimental for interoperability and re-usability

NeXus is a common data format for neutron, x-ray, and muon science. NeXus is defined in base class definitions, application definitions, and design principles.

Base class definitions:	Application definitions:
• define the *complete* set of terms that *might* be used in an instance of that class • give structure and semantics	• define the *minimum* set of terms that *must* be used in an instance of that class • prescribes requirements for a specific application

Design principles:






```
verysimple.nx5 : NeXus data file
  @default = "entry"
  entry:NXentry
    @NX_class = NXentry
    @default = "data"
    data:NXdata
      @NX_class = NXdata
      @signal = "counts"
      @axes = "two_theta"
      @two_theta_indices = [0]
      counts:int32[15] = [1193, 4474, 53220, '...', 1000]
        @units = "counts"
        @long_name = photodiode counts
      two_theta:float64[15] = [18.9094, 18.9096, '...', 18.9122]
        @units = "degrees"
        @long_name = "two_theta (degrees)"
```

Result at one beamline

```
NXrixs:NXsubentry
    @NX_class = "NXsubentry"
    @default = "DATA"
    definition:NX_CHAR = [b'NXrix']
    start_time: --> /entry/experiment_info/start_time
    title: --> /entry/experiment_info/label
    DATA:NXdata
        @NX_class = "NXdata"
        @axes = ["data"]
        data: --> /entry/instrument/detector_1/data
        incident_energy: --> /entry/instrument/monochromator/energy
    instrument:NXinstrument :
        source: --> /entry/instrument/source
        beam:NXbeam
            incident_energy: --> /entry/instrument/monochromator/energy
        detector:NXdetector
            data: --> /entry/instrument/detector_1/data
            x_pixel_size:NX_FLOAT = size[0]
            y_pixel_size:NX_FLOAT = size[1]
            mode:NX_CHAR = DataType
    calibration_data:NXdata
        @NX_class = "NXdata"
        calibration_1:  --> ../calibration/rixsCucalcold_R0001.nxs | entry/instrument/detector_1/data/data
            @pid = "https://hdl.handle.net/21.11151/xxx1-yyy1"
        calibration_2:  --> ../calibration/rixsCucalcold_R0001.nxs | /entry/data
            @pid = "https://hdl.handle.net/21.11151/xxx2-yyy2"
        energy_calibration_type:NX_CHAR = [b'y_pixel']
    SAMPLE:NXsample
        @NX_class = "NXsample"
        name: --> /entry/experiment_info/label
```

DataCite bibliographic information for findability

Mandatory fields

- Identifier
- Creator
- Title
- Publisher
- PublicationYear
- ResourceType

Recommended fields

- Subject
- Contributor
- Date
- Publisher
- RelatedIdentifier
- Description
- GeoLocation

→ DataCite DOI also relevant for citing data and getting credits.

https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf

PIDs to consider by facilities

- Richer information context for facility experiments,
- Better (more structured and less ambiguous) data provenance,
- Role-based credits to various participants of a facility research lifecycle,
- Advanced reasoning over a facility research impact:
- Not necessarily reduced to one or two aggregated metrics but based on connections between various elements of a facility research discourse and, potentially, on some graph-based algorithms

PIDs for referencing