



Draft recommendations on metadata capture and specifications

Wiebke Lohstroh¹, Frank Weber², Sebastian Busch³, Heike Görzig⁴, Bridget Murphy⁵

¹ *Heinz Maier-Leibnitz Zentrum, Technische Universität München*

² *Institut für QuantenMaterialien und Technologien (IQMT), Karlsruhe Institut für Technologie (KIT);*

³ *German Engineering Materials Science Centre (GEMS) at MLZ, Helmholtz-Zentrum hereon GmbH;*

⁴ *Helmholtz-Zentrum Berlin; ⁵ Institut für Experimentelle und Angewandte Physik, Christian-Albrechts-Universität zu Kiel*

Version: 1.0

Date: 19.06.2024

Project Name: DAPHNE4NFDI

DFG project number: 460248799

DOI: 10.5281/zenodo.12169110

Table of Contents

1	Introduction	3
2	Current status: Recommendations from ExPaNDS.....	4
3	Current Status of metadata collection within the DAPHNE4NFDI community	6
4	Metadata – DAPHNE4NFDI comments and recommendations	7
4.1	Bibliographic / administrative data	8
4.2	Data sets	10
4.3	Sample	12
4.4	Record of experiment conduction.....	14
5	Acknowledgements.....	14
6	Appendix: use case examples	15
6.1	X-ray imaging and computed tomography (CT) technology	15
6.2	Triple Axis Neutron and Extended X-ray Absorption Fine Structure Spectroscopy	21

1 Introduction

DAPHNE4NFDI¹ is a consortium within the Nationale Forschungsdaten Infrastruktur² (NFDI) in Germany, dedicated to the development of data management tools and best practices for research data from Photon and Neutron sources. The consortium comprises 18 partners both from the user community conducting experiments at large scale facilities and the institutions operating the instrument suites at the facilities. The work program of task area 1 (TA1) of the consortium is centered around data and metadata capture during the experiment with the aim to develop the FAIRness^{3, 4} of data generated at the German Photon and Neutron large scale infrastructure facilities. Typically, scientists go to these facilities to complete experiments following a successful peer reviewed proposal. To reach this goal, TA1 aims to compile recommendations and set standards for data and metadata capture during the experiment, as well as to provide technical tools to foster live automatic capture of experimental relevant information. As such, TA 1 seeks to create a basis for the data acquisition that forms the basis for subsequent ingestion into data catalogues, the population of reference data bases (TA 2) or the software infrastructure for data analysis (TA3). Research with Photons and Neutrons is traditionally internationally well connected and DAPHNE4NFDI closely follows and participates on the national level with the committee for research with synchrotron and free electron science KFS⁵ and the committee for neutron science KFN⁶ and also with the European networks of User representation, ENSA⁷ and ESUO⁸ and the Large scale facilities (LSF) representations, LENS⁹ and LEAPS.¹⁰ In addition, two recent *European Open Science Cloud Projects* PaNOSC¹¹ and ExPaNDS¹² have laid a solid foundation for the development of the data management at Photon and Neutron sources that DAPHNE4NFDI will build upon and further develop with a strong focus on the needs of the user community to use and reuse data. This will improve the quality of science and build a strong framework for scientific collaboration.

The following document starts by summarizing the current status with respect to the FAIRness of data from Neutron and Photon sources, considering the final recommendation from the ExPaNDS project¹³ as a benchmark. A brief summary and assessment on the current general status at the facilities at the start of the project is given. Finally, the document formulates

¹ <https://doi.org/10.5281/zenodo.8040606>

² <https://www.nfdi.de>

³M.D. Wilkinson, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship, Scientific data. <https://doi.org/10.1038/sdata.2016.18>

⁴ <https://www.go-fair.org/fair-principles/>

⁵ <https://www.sni-portal.de/en/user-committees/committee-research-with-synchrotron-radiation>

⁶ <https://www.sni-portal.de/en/user-committees/committee-research-with-neutrons>

⁷ <https://www.neutrons-ensa.eu/>

⁸ <https://www.esuo.eu/>

⁹ <https://lens-initiative.org/>

¹⁰ <https://leaps-initiative.eu/>

¹¹ <https://www.panosoc.eu/>

¹² <https://expands.eu/>

¹³ Nicolas Soler, Abigail McBirnie, Alejandra Gonzalez-Beltran, Andrey Vukolov, Carlo Minotti, Heike Goerzig, Krisztian Pozsa, Brian Matthews, & Majid Ounsy. (2022). Final recommendations for FAIR Photon and Neutron Data Management (FINAL). Zenodo. <https://doi.org/10.5281/zenodo.6821676>

recommendations for relevant (meta)data fields that should be further developed in DAPHNE4NFDI. As such, the recommendations drafted here are a summary of the DAPHNE4NFDI TA1 regular meetings and workshops, where the user requirements and the facility services were discussed and evaluated in detail, both on a generic level as well as focused on use cases examples from a range of scientific fields. For these discussions, representatives of the user communities from universities, research institutes, and from the operating facilities worked closely together. With this approach, we are confident to address the breadth of scientific research areas using Photon and Neutron sources as well as to consider the technical and security requirements and boundary conditions of the operating facilities.

This document focuses on metadata capture during experiments at large scale facilities, i.e. it outlines the DAPHNE4NFDI recommendations for the captured, aggregated and stored metadata of experiments at PaN facilities. Recommendations concerning research data licenses, open data and access policies are not part of this document.

2 Current status: Recommendations from ExPaNDS

One of the results of the European Project ExPaNDS are the recommendations on the FAIR Photon and Neutron Data management, published in July 2022.¹⁴ The ExPaNDS report summarizes the current status, modalities and current practices at the large scale facilities inside and outside the ExPaNDS consortium concerning (meta)data collection, storage and exposure, and formulates recommendations for future adoption. The report also examines the commonalities between the framework at the PaN facilities with more overarching EOSC discovery platforms B2Find¹⁵ and OpenAire.¹⁶ DAPHNE4NFDI will build on these recommendations, developed on a European level, to foster further development and adoption at national large scale facilities as well as within the user communities.

The proposed metadata framework considering the research data life cycle for a PaN experiment from the proposal submission through the experiment and data analysis to publication for PaN facilities as developed by ExPaNDS is shown in Table 1. Careful consideration has been given to the level of priority of the collected information and which aspect of FAIRness it addresses.

¹⁴ Nicolas Soler, Abigail McBirnie, Alejandra Gonzalez-Beltran, Andrey Vukolov, Carlo Minotti, Heike Goerzig, Krisztian Pozsa, Brian Matthews, & Majid Ounsy. (2022). Final recommendations for FAIR Photon and Neutron Data Management (FINAL). Zenodo. <https://doi.org/10.5281/zenodo.6821676>

¹⁵ <https://b2find.eudat.eu/>

¹⁶ <https://www.openaire.eu/>

	METADATA FIELD	PRIORITY	ASPECT
Proposal	PI/Main proposer	P1	FA
Principal investigator	Co-investigators	P1	FA
User office	Instrument requested	P1	F
	Funding source	P2	F
	Sample description	P1	F
	Proposed experimental conditions	P1	F
	[Safety conditions]	P3	F
	Experiment description	P1	F
	Prior art (related publications, proposals)	P2	F
	Facility information	P1	F
	Proposal identifier	P1	F
	[Approval panel]	P3	/
	Sample safety assessment	P2	/
Approval			
User office	Allocated day & time on instrument	P2	FA
Instrument scientists	Scheduled visiting experimental team	P2	FA
Safety groups	<i>Safety Training data</i>	P3	/
Approval panel	Detailed experimental planning	P2	F
	Sample preparation	P2	FR
	[Sample reception]	P3	/
Scheduling			
User office	Visiting experimental team (user id)	P1	FA
Principal investigator	Experiment date	P1	FA
Instrument local contact	Sample information	P1	FR
	Instrument information	P1	FR
	Calibration information	P1	FR
	Experimental planning	P2	FR
	Environmental parameters	P2	FR
	Laboratory notebook	P2	FR
	Instrument scientist	P2	F
	[Experimental report]	P3	R
Experiment			
Experimental team	Persistent Identifiers (PIDs)	P1	FA
Instrument scientist	Preservation description information	P1	AR
Facility operations (DAQ)	Dataset information	P1	F
	File identifier	P2	AR
	[Representation information]	P3	IR
	[Instrument parameters]	P3	FR
Storage			
Facility operations (ITS)	Processing team (user ID)	P2	AIR
Data management team	Original data	P1	IR
	Data format (after processing)	P1	IR
	Dataset information	P2	AIR
	Processing information	P1	R
	Software package information	P1	R
Data processing			
Data processing team	Analysis team (user id)	P2	AIR
Instrument scientists (help)	Original data	P1	IR
Facility operations (for auto DP/DA)	Software package information	P1	IR
	Dependence tracking and workflow	P2	R
	Data formats (after analysis)	P1	IR
	Dataset information	P2	IR
	File identifier	P1	AIR
	[Instrument parameters]	P3	IR
	[Calibration information]	P3	IR
Data analysis			
Data analysis team	Authors / Coauthors (user ID)	P1	FA
Instrument scientists (help)	Proposal information	P1	FA
Facility operations (for auto DP/DA)	Publication information	P1	F
	persistent Identifier (PID)	P1	F
	[Supplementary data information]	P3	F
Publication			
Experimental team(s)	Resource identity	P1	FI
Instrument scientist	Related resource	P2	F
User office	Creator	P1	F
	Contributor	P2	F
	Title	P1	F
	Publisher	P1	FI
	Publication year	P1	FI
	Licence	P1	IR
	Release date	P1	IR
Data publication			
Data processing team			
Instrument scientists			
Facility operations (for auto DP/DA)			
Data management team			

Table 1: Representation of the metadata framework as suggested by ExPaNDS (Figure reproduced from [14]). The steps typical for data generated at large scale facilities (data continuum of PaNdata ODI D6.1 and ExPaNDS D2.2) are indicated on the left, the recommended metadata fields, as well as an assessment of the necessity to fulfil the FAIR criteria (P1 – essential, P2 – important, P3 – useful) and which aspect of FAIR (F - Findable , A - Accessible, I – Interoperable, R- Reusable) is addressed are given in the right columns.

3 Current Status of metadata collection within the DAPHNE4NFDI community

The experiment, storage and subsequent data processing¹⁷ is at the core of the provision of the service of any large scale facility, and its generated data are the starting point for the data analysis and subsequent publication of the scientific outcome. Here, we will primarily discuss metadata capture during and up to the completion of the experiment at the facility. Besides the scientific metadata, this also includes consideration for ownership of the research data, curation, archiving, and license for access, all usually regulated by the data policy of the respective facility. For most PaN facilities, updates of the data policies are in progress or have been recently prepared generally following the recommendations of the PaN-data Europe Strategic Working Group.¹⁸

Metadata connected to sample history not within beamtime, data analysis, software tools, data publications etc. are equally important but exceed the current scope of TA1 which is focused on the FAIRness of the data that is generated and aggregated during the experiment.

The experiment, (short) term data storage and data processing are closely interlinked and currently done mostly at the time of the experiment on-site at the facility. Long-term storage or archiving is the responsibility of the principle investigator, nevertheless the facilities may act as custodian of the raw and processed data¹⁷ that have been collected at its premises as regulated by the data policies of the respective institute.

Typically, most or all of the metadata fields recommended by ExPaNDS for the proposal and approval (see Table 1), such as Principle Investigator (PI), proposed experiment, requested instrument, etc. are collected during the proposal submission and reviewing process, and are therefore collected from the user office software of the respective facility. Similarly, the user team and stay are organized via the user office systems hence are centrally captured. In contrast, experimental planning, sample description, preparation and pre-beam time characterization are typically not part of the documentation at the facility but are part of users documentation recorded in the home laboratory prior to or during the experiment.

The extent of the metadata fields and the ease of capture during the experiment differs quite substantially between facilities and beamlines. As for proposal submission stage, administrative metadata are typically available and easily recorded. However, everything that is required to concisely describe the work flow of the experiment relies on the aggregation of the facility supplied information on the instrument, instrument status, calibration, sample environment and environmental parameters as well as the user supplied information on sample and sample preparation and user supplied equipment.

¹⁷ Here, we refer to data processing to the process of generating data ready for scientific analysis from the collected raw data files and if applicable, the corresponding calibration and/or background runs, i.e. processed (or reduced) data are in meaningful scientific units and free from any instrument specific settings.

¹⁸ Common policy framework on scientific data: 10.5281/zenodo.3738498; PaNOSC data policy framework, 10.5281/zenodo.3826039.

As of today, this information is typically collected, but distributed over different storage locations, such as the users own laboratory notebook (either digital or hand written), the instrument logbook, facility electronic laboratory notebooks (ELNs), the facilities storage, etc.. For storage, the NeXus format¹⁹ is increasingly adopted (also from the DAPHNE4NFDI partners) but prevalence is not yet complete among all beamlines or instruments. In addition, the meta(data) collection is mainly focused on immediate needs of the PI and the research group in view of data processing and scientific analysis. Less in focus is the suitability and completeness of metadata in view of later findability, use and re-use, also by third parties. This also applies for facilities that have already on-line metadata catalogues and an open access policy in place, where typically, the raw and processed data belonging to a specific proposal/experiment can be obtained (under the license according to the respective data policy), but annotations concerning actual conduction of the experiment and the sample (and its history) are typically limited. Depending on research field, there are conventions and standards that are adopted by the experimenters during beam times, but a systematic capture of the basic sample properties for the digital record is missing at most facilities.

4 Metadata – DAPHNE4NFDI comments and recommendations

The core of the research of the Photon and Neutron community is the experiment at a specific instrument or instruments at the large scale facility. Here, the raw and processed data are generated that will be analysed and published in scientific publications. The captured metadata are thus especially important to further use and reuse the data.

Here, we consider not only the concept of a single scientist analysing the data but also that of a user consortium, and third party reuse either by human researchers but also for machine learning applications. Thus, the metadata captured and aggregated serve a number of different purposes, spanning from immediate processing in preparation for scientific analysis (e.g. to remove instrumental parameters) to the reuse at a later moment in time in a different context by a third party. Similarly, the 'provider' of the data ranges from machine registered data from the detectors and the instrument to 'analogue' information on e.g. details on the mounting a specific sample, or notes on any relevant event that might compromise data quality (such as e.g. heavy thunderstorms or heavy machinery nearby).

The challenge at the large scale facilities is to aggregate all these metadata in appropriate, electronic form, such that the data package that is generated during experimental beam time is as complete as possible and approaches the FAIR principles, and at the same time, serves the needs of the user community for their intended use and scientific analysis.

In the following, the DAPHNE4NFDI recommendations for metadata collection at the time of the experiment are outlined, thus a description of the minimum information that the DAPHNE4NFDI community considers as essential to be part of the aggregated data package

¹⁹ <http://www.nexusformat.org/>

that is generated at the end of an experimental beam time. In view of the breadth of the scientific domains and the variety of different techniques, this generic scheme covers the overarching communalities. Specific examples for use cases are presented in the appendix.

Within DAPHNE4NFDI, we consider a categorization into the following areas:

- Bibliographic and administrative metadata
- Metadata of the dataset (measurement): raw data, processed data
- Sample metadata
- Record of experiment conduction

Additionally, the recommended metadata fields serve different functions or purpose. On the one hand, different aspects of the FAIR criteria are addressed (see also Table 1), on the other hand, the captured meta(data) contains mandatory information for subsequent analysis steps. Thus, we have the cross cutting categories:

- metadata for findability, also for broader scientific community, adherence to FAIR criteria
- metadata for data processing and data analysis: to be included in standardized file formats, also for automated processing
- additional information on experiment conduction: data analysis and curation/quality control, electronic laboratory logbook
- metadata on sample information – findability, reusability and data analysis

4.1 Bibliographic / administrative data

Bibliographic and administrative metadata give provenance and context to the datasets created during beamtimes at the large scale facilities. The information is typically available from the facility user office system and available in digital form at the time of the experiment. Any handling of personal data is regulated by the GDPR (General Data Protection Rules), and the respective data policies / terms of reference of the user facilities.

Suggested metadata fields are as follows

Principal Investigator /Co-Proposers / Experimental team		
Last Name		
First Name		
Email		
Affiliation		
Organisation ID	ROR ID ²⁰	
User ID	ORCID / ResearcherID	
Role	e.g. principle investigator / experiment conduction / data analyst / sample preparation / support laboratory personnel / Local contact other	
Instrument / Beamline		

²⁰ <https://ror.org/>

Name		
Instrument PID	PID	PID of instrument or publication describing the state of the instrument at the time of measurement
Operator	organization ID	operator of the beamline, might be different from the facility
Facility		
Name		
Organisation ID	ROR ID	of the facility where the measurement was done
Facility contact		functional email, e.g. data steward
Proposal		
Proposal name		
Proposal description		
Proposal / Experiment Identifier	Proposal ID	e.g. proposal identifier
Start time	date	allocated dates of scheduled Experiment
End time	date	allocated dates of scheduled Experiment
Public research	public or proprietary	
Funding identifier		
License	e.g. CC0, CC-BY, CC-BY-SA, see ²¹	according to the data policy of the facility
Expiry of embargo period	date	according to the data policy of the facility

Purpose of the suggested metadata fields:

	FAIR	Provenance	Data processing and analysis	Quality control	Sample information
Principal Investigator / Co-Proposers / Experimental team	F				
Instrument / Beamline	FR	x			
Facility	F	x	x		
Operator	F				
PID of instrument of beamline	F				
Facility contact	FR				
Proposal/ Experiment Identifier	FA	x			
Start time	F	x			
End time	F	x			
Public research	F				
Funding identifier	F	x			

²¹ <https://creativecommons.org/licenses/by-sa/4.0/>

License	AR				
Expiry of embargo period	AR				

4.2 Data sets

Data set (for raw and processed data) are captured at the time of the experiment and are supplied by the facility or the user through automatic capture or manual input. Due to the variability of techniques and research areas, the metadata fields that will serve the needs to the user community will also be varying. In short, the captured metadata should provide information on:

- Scientific context of the data set ('what is the measurement about')
- Instrumental setup and used measurement technique
- Parameters that are required for data processing and analysis
- Parameters that are required for quality control

Relevant keywords or a short description should be included to easily understand the purpose of the measurement for both persons and machines, e.g. 'tomography of a part of dinosaur tooth'.

The data set metadata should also include references to the instrument configuration, such as the used sample environment and any installed additional equipment, as well as the identification of the measurement technique. In case, the data is not available via the instrument control system, structured and easy-to-use tools for manual input should be made available such that the used instrument setup is documented.

The environmental parameters and instrument settings should include any parameter that is needed for data processing and analysis. This might include instrument specific data needed to correct for any instrument related part in the collected detector output, e.g. sample-to-detector distances, monitor outputs, energy of the incoming beam, temperatures, etc..

During the experiments, time-logs for instrument actions, as well as instrument parameters are helpful for quality assurance and validation of the collected data. These might include e.g. temperature controls, motor/encoder positions, beam intensity, etc.. On an optional level, parameters that are required for other tasks such as digital twinning, instrument maintenance or prediction of instrument maintenance requirements can be included. This data is likely to be of higher relevance to the facility rather than to the users. It is none the less very useful to record this data.

The metadata fields below should be collected for any individual scan during an experiment beam time. Typically, one parameter will be varied in a one dimensional or more than one in a multi dimensional scan or set of nested scans and the detector output collected at each point. The level at which metadata needs to be refreshed will depend strongly on the individual requirements. It is recommended that a high level is adopted to avoid needless repetition. Here some flexibility is required and cannot be prescribed, as there is no one single solution that fits all users needs.

Suggested metadata fields are as follows:

Identifier/Name	data set name	Human readable
start time	Data/time	of the measurement/run: Applies to file
end time	Data/time	of the measurement/run: Applies to file
Sample ID	PID of sample	PID (external of local ID)
Data format		ideally NeXus format is used
Title/Description	free text, describing the scan	
Instrument: Configuration used, installed options	choice of different instrument specific options by keyword, including installed detectors and monitors Additional free text, if applicable. Also PIDs of the components can be used if available.	description of installed options and links to any further information on the used instrument setup
Measurement technique(s)	PaNET ontology ²²	if possible or free text,
Environment conditions and instrument settings	key value pairs 'name, value, unit'	as many key value pairs (or time-log series) as needed, also for automatic data processing and analysis, e.g. Temperature, pressure, sample-detector distance, used wavelength, energy, beam parameters, etc.
Measurement Mode	keywords	identifying the free variable that is varied during the scan or measurement: EnergyScan, TemperatureScan, WavelengthScan, Snapshot, TimeScan, MagneticfieldScan
Sample environment: installed options	choice of different sample environment options by keyword, or free text, sample environment PID	
Purpose of data set:	choice of different options: measurement, calibration / alignment/ background / or free text	
References	link to related information/datasets e.g. calibration data	
Comments	free text: to add comments after the measurement (e.g. to note mistakes failures etc.)	
Quality Metrics		data Quality Metrics: number given by the user to rate the dataset
Validations Status		defines a level of trust, e.g. a measure of how much data was verified or used by other persons

²² Expands Ontologies V1.0". Zenodo, June 4, 2021. doi:10.5281/zenodo.4806026; <https://expands-eu.github.io/ExPaNDS-experimental-techniques-ontology/index-en.html>

Purpose of the suggested metadata fields:

	FAIR	Data processing and analysis	Quality control	Sample information
Identifier	FA	x		
start time	F		x	
end time	F		x	
Sample ID	FR			x
Data format	IR	x		
Title/Description	FR			x
Instrument: Configuration used, installed options	FR	x	x	
Measurement technique(s)	FIR	x		
Scan axis (more than one option)	IR, aggregated value also for F	x		
Sample environment: installed options	FR			
Scientific metadata:	FIR	x	x	x
purpose of data set:	FIR	x	x	
Comments	FIR		x	

4.3 Sample

Sample: The sample is at the core of scientific experiment, and any collected data at the PaN facility is only meaning full in the context of the information provided on the actually probed sample. DAPHNE4NFDI recognizes the sample identification and description as especially important to be included in the aggregated data package, especially for findability after the experiment and potential reuse of data. Appropriate tools to connect the aggregated data package e.g. with a pre-registered sample PID, or the manual input of most basic sample description parameters to enhance later findability and reusability is of outmost importance. The samples measured at PaN facilities cover a broad range, from ‘stable’ individual samples or single crystals that are taken to the facility to be measured at varying conditions to samples that are prepared just at the time of the experiment on the beamline, for instance as for the study of soft matter monolayers, prepared from a buffer solution.

The metadata of a sample is as diverse as the disciplines related to the sample. Datasets created during the lifetime of a sample can be used as a description of the sample and display its changes and alteration. Ideally, a sample PID is used to aggregate all these datasets that are created during the lifetime of a sample and to follow its history.²³ Therefore, the use of sample PIDs is highly recommended. At the time of the experiment, a minimum of

²³ In case the sample only generated in the moment of the measurement at the beamline and does not exist after the measurement no PID might be needed.

information should be included in the metadata of the collected dataset, including a human readable name.

Name	free text, human readable	e.g. 20160104_sampletA_run_SmithE_v1
Local ID or PID		Ideally registered PID, e.g. IGSN, otherwise local ID
registration schema		if externally registered
Agency		if externally registered
key words	e.g. fossil, compound, artifact, cathode materials	following e.g. PhySH
Description	free text	description of sample
Chemical composition		e.g. CAS number, SMILE, or InCh
Common name		if applicable, free text
State	liquid, powder, single crystal, solid, thin film, other	
Mass		if applicable
Crystal structure		if applicable
Unit cell parameters		if applicable
Preparation date	date	if applicable
Supplier / creator		
Parent batch	parent sample ID	if applicable
additional fields		any additional information

The purpose of these fields can be categorized as:

	FAIR	Data processing and analysis	Quality control	Sample information
Name	F			x
Local ID or PID	F			x
registration schema	F			x
Agency				x
key words	F			x
Description	F			x
Chemical composition	F	x		x
Common name	F			x
State	R	x		x
Mass	R	x		x
Crystal structure	R	x		x
Unit cell parameters	R	x		x
Preparation date	FR		x	x
Supplier / creator	FR		x	x
Parent batch	FR		x	x
Sample stored after experiment			x	x
additional fields				x

4.4 Record of experiment conduction

Besides the instrumental time log of any (automatically captured) action of the instrument, the accurate record of the workflow of the experiment forms an essential part of the aggregated data package. This includes the documentation of all steps of the experiments and links to the data set. Typically, it consists of a mixture of manual inputs (e.g. documentation of sample mounting, intermediate steps of sample manipulation, parameters for in-situ treatment or preparation), information on data acquisition from the instrument or beamline, as well as the documentation of intermediate results that are used to decide upon the next step of the experimental plan. It is also important that the workflow documentation includes the sample preparation, prior characterization and treatments along with all other activities carried out on the sample during beamtime. This should be recorded in a laboratory notebook, which is then part of the aggregated data package generated during the experiment.

Electronic Laboratory Labbook (ELN)	Link	User record, link to find ELN, or *pdf export
ELN format	ELN format	
Instrument time log	Link	Automatically captured instrument and sample environment status (as function of time), link where to find
Source beam time log	Link	Time log of source parameters (status of beam in ring etc.)

The requirements for the technical solution for the electronic user logbook are discussed in a separate white paper of the DAPHNE4NFDI consortium.²⁴

5 Acknowledgements

This work was supported by the consortium DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortium is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460248799.

²⁴ In preparation by TA1, DAPHNE4NFDI

6 Appendix: use case examples

In this section, we give some examples and application for the suggested metadata collected for different use cases. It focuses solely on examples for metadata fields that are specific for the research area and the used techniques as the more general terms, such as administrative data etc., have already been discussed in the main part. These will not be repeated here, but of course they will be part of the aggregated data sets.

6.1 X-ray imaging and computed tomography (CT) technology

Paola Coan¹, Hafiz Fahad¹, Markus Osterhoff²

¹Ludwigs-Maximilian Universität München, ²Georg-August Universität Göttingen

The advances in synchrotron imaging and computed tomography (CT) technology and experimental setups has opened new avenues in scientific research by offering levels of sensitivity and specificity in the investigation of materials that overcome the performance of standard laboratory methods. Phase-contrast synchrotron X-ray computed tomography have ushered in an era of non-invasive, three-dimensional visualization, allowing researchers to explore imaging at the micro- to nano-scale for applications that span from materials science, biomedicine and cultural heritage, among others. As a result, a large amount of data is produced, up to several petabytes per year. Effective management of such a substantial data influx requires tailored solutions for electronic data capture, coupled with robust data management and storage capabilities, as well as for the repository of processed (reduced) data and analysis code to go with each publication in order to maximise data reuse and transparency.

At the time of measurement, metadata are collected that are specific for this field of research, the used instrument and the type of conducted measurement. For X-ray computed tomography with the focus on biological matter, the key information to be collected during the experiments for data processing and analysis include:

Dataset

Instrument: configuration used, installed options:

- Synchrotron source type: undulator with gap or wavelength shifter with field
- Monochromator type: double Si crystal, Bragg, Laue mirror (PID, if available)
- installed detector (PID if available)

Environmental parameters/Settings: key value pairs:

- {EnergyResolution, pink/white/mono, setting}
- {Energy, value, unit}
- {SampleToDetectorDistance, value, unit}
- {Detector binning, [number_pixel_w, number_pixel_h], [pixel, pixel]}

- {Detector cropping, [number_pixel_w, number_pixel_h], [pixel, pixel]}
- {Detector field of view, [number_pixel_w, number_pixel_h], [pixel, pixel]}
- {Detector pixel size, [size_pixel_w, size_pixel_h], [unit, unit]}
- {Magnification, value, scalar}
- {ExposureTime, [time1, time2, ..., timeN], [seconds, seconds, ..., seconds]}
- {NumberOfProjections, value, ""}
- {NumberOfFlatFields, value, ""}
- {NumberOfDarkFields, value, ""}
- {NumberOfPositionalReferences, value, ""}
- {monitorCounts, [m1, m2, ... ,mN], [counts, counts, ..., counts]}
- {TimestampImage,[time1, time2, ..., timeN], [date,date, ..., date]}"

Measurement Mode :

- CTAcquisition
- CTFullAcquisition
- CTHalfAcquisition
- ImageOrdering_flat_dark_radiosequence

Sample environment: installed options:

- in gas
- liquid solution
- cryo (PID, if available)

Further information:

- Rotation center: value, estimated
- estimated spatial resolution: value, unit
- Acquisition CT-Marco: {string}"

Sample information:

Sample preparation metadata captures critical details about the experimental samples and its preparation for the measurement campaigns. Each sample is identified by its name and local ID or PID (Persistent Identifier). Parameters such as state and mass provide insight into the physical properties of the sample essential information for inclusion in the catalogue for the user. Features like part of a batch, parent batch sampleID, and time-dependent samples are essential aspects of sample preparation that merit inclusion in the catalogue. A range of additional fields allows flexibility to meet specific experiment-related requirements. The groups specification indicates the category to which a sample belongs in a study involving groups of samples prepared in different ways or of different origins. The metadata table also includes the sample preparation technique, embedding materials, along with the possibility

of using labelling materials (e.g. staining, contrast agents), providing specific information on the type of material if utilized. All these parameters are crucial information for the user for further data analysis; as different embedding materials, preparation techniques or groups may determine different image analysis steps and processing pipelines for the datasets of a specific sample. In addition, these details may assist in quality control by ensuring that samples are correctly assigned to their respective groups. This comprehensive set of metadata ensures a detailed understanding of each sample's characteristics and the methods employed in its preparation for experimental purposes, making it imperative for inclusion in the catalogue.

Additional fields might be:

- SampleGroup identifier
- sample state: alive / dead
- PreparationNotes: methods (incl. DOI, used materials for cutting)
- labeling material: none / iodine / gadolinium / nanoparticles / etc
- embedding: Paraffine, formalin: concentration, ethanol: concentration, agar, etc

Metadata Field	Type	Comment 1	Usecase Example
Identifier	uuid		
Identifier/Name	data set name	human readable name	
start time	Data/time	of the measurement/run: Applies to file	
end time	Data/time	of the measurement/run: Applies to file	
Sample ID	PID of sample	PID (external of local ID)	
Data format		ideally NeXus format is used	
Title/Description	free text, describing the scan		
Instrument: Configuration used, installed options	Keywords of free text	description of installed options and links to any further information on the used instrument setup	<ul style="list-style-type: none"> - Synchrotron source type: undulator with gap or wavelength shifter with field - Monochromator type: double Si crystal, Bragg, Laue mirror (PID, if available) - installed detector (PID if available)
Measurement technique(s)	PaNET ontology ²⁰	if possible or free text,	
Environment conditions/Settings	Key value pairs ; 'name, value, unit'	as many key value pairs (or time-log series) as needed, also for automatic data processing and analysis, e.g. Temperature, pressure, sample-detector distance, used wavelength, energy, beam parameter,	<ul style="list-style-type: none"> - {EnergyResolution, pink/white/mono, setting} - {Energy, value, unit} - {Detector binning, [number_pixel_w, number_pixel_h], [pixel, pixel]} - {Detector cropping, [number_pixel_w, number_pixel_h], [pixel, pixel]} - {Detector field of view, [number_pixel_w, number_pixel_h], [pixel, pixel]} - {Detector pixel size, [size_pixel_w, size_pixel_h], [unit, unit]} - {Magnification, value, scalar} - {ExposureTime, [time1, time2, ..., timeN], [seconds, seconds, ..., seconds]} - {NumberOfProjections, value, ""} - {NumberOfFlatFields, value, ""} - {NumberOfDarkFields, value, ""} - {NumberOfPositionalReferences, value, ""} - {monitorCounts, [m1, m2, ... ,mN], [counts, counts, ..., counts]} - {TimestampImage,[time1, time2, ..., timeN], [date,date, ..., date]}

Measurement Mode	keyword	identifying the free variable that is varied during the scan or measurement: EnergyScan, TemperatureScan, WavelengthScan, Snapshot, TimeScan, MagneticfieldScan	suggested values e.g.: CTAcquisition, CTFullAcquisition, CTHalfAcquisition ImageOrdering_flat_dark_radisequence
Sample environment: installed options	Choice of different sample environment options by keyword, or free text, sample environment PID		e.g. - {in gas / liquid solution / cryostream}
Purpose of data set:	choose one of: measurement, calibration / sample / background / alignment / ...		
References	e.g. calibration data		
Comments	free text: to add comments after the measurement (e.g. to note mistakes failures etc.)		- Rotation center: value, estimated - estimated spatial resolution: value, unit - Acquisition CT-Marco: {string}
QualityMetrics Validations Status		Data Quality Metrics is a number given by the user to rate the dataset.	

Sample

Metadata Field	Type	Comment 1	Usecase example
Name - human readable	free text	human readable name	
Local ID		local ID	
ext PID	link		e.g. IGSN
registration schema		if externally registered	

Agency		if externally registered	
key words		following e.g. PhySH	examples might be: fossil, compound, device, artifact, cathode material, etc see also PhySH.org
Description	free text	Description of sample	
Chemical composition		e.g. CAS number, SMILE, or InCh	see also core scientific metadata Model http://icatproject-contrib.github.io/CSMD/
Common name		if applicable, free text	
State	liquid, powder, single crystal, solid, thin film, in situ		
Mass	value, unit pair	if applicable	
Crystal structure		if applicable	
Unit cell parameters		if applicable	
Preparation date	Data	If applicable	
Supplier / creator			
Relationships	Parent sample ID	if applicable	Link to related samples (e.g. parent / child relations), see also as for datasets- Samplegroup identifier
additional fields		any additional information	e.g. - {alive / dead} - {PreparationNotes: 'methods (incl. DOI, uses materials for cutting')} - {labeling material: 'none / iodine / gadolinium / nanoparticles / etc'} - {embedding: 'Paraffine, formalin: concentration, ethanol: concentration, agar, etc}

6.2 Triple Axis Neutron and Extended X-ray Absorption Fine Structure Spectroscopy

Yuliia Tymoshenko¹, Sebastian Paripsa², Frank Weber¹, Astrid Schneidewind³, Christoph Herb⁴

¹Institut für QuantenMaterialien und Technologien (IQMT), Karlsruhe Institut für Technologie (KIT) ; ²Fk. 4, Physik, Bergische Universität Wuppertal; ³Jülich Centre for Neutron Science (JCNS) at Heinz Maier-Leibnitz Zentrum (MLZ), Forschungszentrum Jülich GmbH; ⁴Heinz Maier-Leibnitz Zentrum, Technische Universität München

Neutron and X-ray spectroscopy methods are important tools to understand the structural and dynamical properties of materials in both applied and basic science. While spectroscopy typically involves a time or energy dependence of the measured signal, we note that there are in fact two distinct applications concerning the purpose and the results of the measurements. One application focuses on the identification of characteristic signatures ('footprints') in the absorption spectra as the energy of the incoming beam is scanned. This footprint is characteristic for the element or isotope composition or to the oxidation state of the comprising atoms. A second application determines the energy transfer (and potentially also the momentum transfer) between the incoming probe beam and the sample, as typically measured in neutron spectroscopy applications. Depending on the energy and momentum transfer range, information on binding states, lattice vibrations and magnetic excitations are obtained. Typical neutron spectroscopy applications include excitations in e.g. H₂ storage materials, MOFs, high-temperature superconductors, quantum magnets, spin liquids, etc.. Typical X-ray spectroscopy applications, especially in Extended X-ray Absorption Fine Structure (EXAFS), include the examination of local atomic structures around selected absorbing atoms. This method is crucial for studying the coordination environment, interatomic distances, and disorder effects in a variety of materials, from catalysts and batteries to environmental samples and biological systems. EXAFS provides unique insights that are invaluable for materials science, chemistry, and even cultural heritage preservation.

Here we propose a set of metadata for

- a typical (unpolarized) neutron triple axis instrument where a set of crystal monochromators is used to determine the neutron energy before and after the scattering (at mostly single crystalline) samples.
- an EXAFS setup, utilizing synchrotron radiation and monochromators to control the energy absorbed by target atoms. This setup captures essential data on atomic distances and coordination environments.

We try to define a set of metadata which can be generically used for these kinds of spectroscopy, including the information for typical samples, that we consider necessary for future user and reuse. The first three columns of the following tables are generic for both use case examples, whereas the last one gives specific examples which might be relevant for a neutron TAS or EXAFS or both techniques.

Dataset

Metadata Field	Type	Comment 1	Usecase Example
Identifier	uuid		
Identifier/Name	data set name	human readable name	
start time	Data/time	of the measurement/run: Applies to file	
end time	Data/time	of the measurement/run: Applies to file	
Sample ID	PID of sample	PID (external of local ID)	
Data format		ideally NeXus format is used	
Title/Description	free text, describing the scan		if applicable
Instrument: Configuration used, installed options	Keywords of free text	Description of installed options and links to any further information on the used instrument setup	<ul style="list-style-type: none"> - Single or multianalyzer/ detector system - Monochromator type: PG(002), Ge(311), Si(111), Si(311), - Focusing mode: 'flat', 'horizontal', 'vertical', 'double', ... - Analyzer type: Ge(311), PG(002), ... - Focusing mode: 'flat', 'horizontal', 'vertical', 'double', ... - Scattering sense: (-1, 1, -1) (mono, sample, ana) - Collimation
Measurement technique(s)	PaNET ontology ²⁰	if possible or free text,	'inelastic neutron spectroscopy', 'neutron scattering', 'neutron diffraction', 'x-ray absorption spectroscopy (including EXAFS and x-ray absorption near edge structure (XANES)', 'x-ray diffraction', 'x-ray fluorescence'
Environment conditions/Settings	Key value pairs ; 'name, value, unit'	as many key value pairs (or time-log series) as needed, also for automatic data processing and analysis, e.g. Temperature, pressure, sample-detector distance, used wavelength, energy, beam parameter,	<ul style="list-style-type: none"> - {"Ei": Initial energy Ei for constant ki mode, value, meV} - {"Ef": Initial energy Ef for constant kf mode, value, meV } - {"ki": Initial wavevector ki for constant ki mode, value, AA-1} - {"kf": Initial wavevector kf for constant kf mode, value, AA-1} - {"constE": constant Energy transfer E (Q-scan), value, meV} - {"constQ": constant momentum transfer Q (E-scan), vector, r.l.u.} - {Electric field vector in reciprocal space, vector, V/m} - {"element": Chosen target element for analysis, value, a.u.} - {"absoEdge": specific absorption edge analyzed, value, a.u.} - {"maxkRange": upper limit of the wavevector (k) considered during the analysis of the EXAFS oscillations, value, Å-1} - {"eResolution": energy resolution for EXAFS scan, value, eV}

Measurement Mode	keyword	Identifying the free variable that is varied during the scan or measurement: EnergyScan, TemperatureScan, WavelengthScan, Snapshot, TimeScan, MagneticfieldScan	<ul style="list-style-type: none"> - constant Q scan with fixed k_i, - constant Q scan with fixed k_f, - constant E scan with fixed Q (rocking scan), - constant E scan with fixed Q/ Q (longitudinal scan), ... - TOTAL ELECTRON YIELD: energy scan capturing surface electron emission as energy varies. - FLUORESCENCE: energy scan measuring x-ray induced fluorescence for bulk analysis.
Sample environment: installed options	Choice of different sample environment options by keyword, or free text, sample environment PID if available		<ul style="list-style-type: none"> -polarization analysis (separate discussion) -slits (Positions and Openings) -magnet -electric Field device -pressure Cell -cryostat -high temperature furnace
Purpose of data set:	choose one of: measurement, calibration / sample / background / alignment / ...		<ul style="list-style-type: none"> - measurement - background - alignment ... - CALIBRATION
References	e.g. calibration data		link to calibration data, link to logfiles, link to instrument control logbook, link to jobfiles or measurement scripts, publications, links to pictures of mounted sample,...
Comments	free text: to add comments after the measurement (e.g. to note mistakes failures etc.)	can also be provided by Nicos via the Remark() function	<ul style="list-style-type: none"> - Scan command: 'qscan((0, 0.46, 0, 0), (0, 0, 0, 0.025), 18, kf=1.3, mon1=6*mpm)', - 'measurement during cooling', 'temperature was rising', 'sample fall down', 'shutter was closed', ...
QualityMetrics Validations Status	free text	Data Quality Metrics is a number given by the user to rate the dataset.	<ul style="list-style-type: none"> - Temperature unstable - sample was moving - bad data, ...

Sample

Metadata Field	Type	Comment 1	Usecase example
Name - human readable	free text	human readable name	
Local ID		Local ID	

ext PID	link		e.g. IGSN
registration schema		if externally registered	
Agency		if externally registered	
key words		following e.g. PhySH	examples might be: keywords similar to paper (more is better): e.g. vanadate, hermatit, iron phosphate, superconductivity, helimagnet, ... (see also PhySH.org)
Description	free text	Description of sample	
Chemical composition		e.g. CAS number, SMILE, or InCh	NiO, ZnCr ₂ Se ₄ ...
Common name		if applicable, free text	
State	liquid, powder, single crystal, solid, thin film, in situ		liquid, powder, single crystal, solid, thin film, pressed pellet, ...
Mass	value, unit pair	if applicable	mass (g)
Crystal structure		if applicable	space group number according to IUCR
Sample Size (along crystal axis approximate is fine)	vector	if applicable	sample volume (mm ³)
Unit cell parameters	vector	if applicable	vector [a, b, c] (Å), vector [alpha, beta, gamma] (deg)
Preparation date	Data	if applicable	
Supplier / creator	free text	if applicable	supplier, crystal grower at institution, natural, ...
Relationships	Parent sample ID	if applicable	link to related samples (e.g. parent / child relations), datasets
additional fields	free text	any additional information	sample mosaic, twinning