ExPaNDS ontologies

Document Control Information

Settings	Value
Document Identifier:	D3.2
Project Title:	ExPaNDS ontologies
Work Package:	WP3
Document Author(s):	S.P. Collins (Diamond), S. da Graca Ramos (Diamond), Daniel Iyayi (Diamond/ University of Oxford), Heike Görzig (HZB), Alejandra González Beltrán (UKRI), Alun Ashton (PSI), Stefan Egli (PSI), Carlo Minotti (PSI)
Document Reviewer(s):	Brian Matthews (UKRI), Daniel Salvat (ALBA), Sophie Servan (DESY)
Doc. Issue:	1.0
Dissemination level:	Public
Date:	04/06/2021

Abstract

We present ontologies for the domain of photon and neutron (PaN) science. With the primary goal of supporting PaN FAIR data catalogue services, we have developed three ontologies: PaN experimental techniques (PaNET), an ontology of NeXus definitions (NeXusOntology), and a semantic integration ontology for the PaN domain (PaNmapping). The ontologies are presented as initial versions, supported by community development workflows. The work represents deliverable D3.2 of the Horizon 2020 ExPaNDS project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Executive Summary

We report on the design and development of a set of ontologies to support FAIR data implementation in Photon and Neutron (PaN) data catalogues. Exploitation of FAIR principles (especially Findability) can be hampered by the lack of consistency in metadata used to annotate data records and search databases. This paper describes the development of several small ontologies to facilitate consistent semantics for terms within the PaN domain by providing global persistent identifiers, community agreed labels and synonyms, and human-readable definitions, annotations and references. Crucially, each ontology is supported by a community maintenance process to allow a managed and agreed approach to modification and extensions in future development. We present Version 1.0 of the ontologies as a viable starting-point for long-term use. The initial ontologies are very simple and designed with the data catalogue use-case firmly in mind. It is anticipated that future developments will allow increasing integration with the wider semantic web.

The ontologies developed so far, as work-package 3.2 of the ExPaNDS EU Horizon 2020 grant (https://expands.eu/), are:

PaNET: Photon and Neutron Experimental Techniques Ontology. This simple ontology provides a taxonomy of PaN techniques, with new techniques being defined as subclasses of multiple, more elementary, technique classes.

NeXusOntology: An ontology of NeXus format definitions (the dominant metadata model with the PaN domain). NeXusOntology is a representation of the formal NeXus definitions and is created automatically from NeXus definition files.

PaNmapping: A semantic mapping ontology for the Photon and Neutron Science domain. This ontology aims to map between the overlapping entities in the most common PaN metadata schemata NeXus format and CSMD/ICAT data schema, as well as integration with the DCAT v2 Ontology and DublinCore.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1. Background and Purpose of the Ontologies

1.1 Background

The work outlined in this document is part of the ExPaNDS project, carried out in close communication with the PaNOSC project (https://www.panosc.eu/), thus representing the majority of European Photon and Neutron sources in a coordinated activity to drive forward Findable, Accessible, Interoperable and Reusable (FAIR) facility data and EOSC services.

The specific task for this deliverable is as follows:

> *Task 3.2: Develop EU Photon and Neutron Ontologies*
>
> *Develop ontologies for main application domains of Photon and Neutron science to standardise the metadata used in metadata catalogues based on requirements defined in WP2. This will ensure that federated EOSC metadata catalogues are not only based on a common syntax, but also on a common semantics.*
>
> *The ontology itself will be provided as an EOSC service using existing tools to document and make ontologies accessible (e.g NeOn, Knoodle, Protégé, Swoop). Development of the ontology will be closely linked to the existing NeXus file format and its further developments (PSI, ISIS and DLS have leading roles in the specification and implementation of NeXus). Photon and Neutron-related ontologies provided by NeXus will be used and extended. In a similar way existing ontologies (such as those provided by NIST) should be taken into account.*

Deliverable related to this task:

> D3.2: Release V1.0 ExPaNDS ontology available as an EOSC online service

In order to hone in on a specific set of deliverables it is necessary to review the primary purpose of the ontologies. Indeed, the modern usage of the word ontology, and the standards, such as the Web Ontology Language (OWL https://www.w3.org/OWL/), used to express knowledge, cover a very wide range of semantic richness. In its simplest form, a controlled vocabulary sets out an agreed and managed set of terms. While this may be all that is required for some purposes, it is often highly advantageous to develop a taxonomy to facilitate classification of entities via standard pre-defined relationships such as ‘subclass’. Ultimately, OWL can be deployed to create a large web or graph of interrelated concepts, with new types of relationship to describe the properties of entities. Each new relationship enriches the semantic description of the entities, connects them to other concepts, and reduces the need for human-interpreted labels. This brings huge benefits to machine and human agents alike but at a very significant cost in terms of effort and complexity. The first step in such an endeavour is therefore to gain an understanding of the immediate and future potential uses of the ontologies.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

1.2 Purpose

The main purposes of the ontologies, alluded to in the original proposal and fleshed out by consultation with PaN community representatives, include:

- To provide controlled vocabularies in the domain of Photon and Neutron (PaN) science, with a globally-unique persistent identifier [1] (PID) for each concept, and a managed process of community maintenance.
- To provide a taxonomy of terms for catalogue tagging/annotation to facilitate FAIR data, including (where appropriate) class/subclass relationships and alternate labels. This will allow searches based, for example, on broader concepts than those used for tagging.
- To provide annotations for human and machine agents, helping newcomers to PaN science to understand the nomenclature of the domain.
- To utilize a framework that is both global and highly expressible in order to facilitate semantic mappings, both within and outside of the PaN domain, and to permit future semantic embellishment via new properties and relationships, allowing gradual integration into the semantic web.

While the initial work focuses on the primary goal of supporting PaN catalogue services, we have taken steps towards satisfying requirements for FAIR vocabularies [2] and have built structures that allow natural development in this area.

We note that the tasks outlined above are open-ended and involve long-term goals that are not, as yet, clearly defined. The specific deliverable of this component of the ExPaNDS project, ‘V1.0 ExPaNDS Ontology’ should be viewed as a ‘demonstrator’ to satisfy short-term FAIR data goals and provide a framework for future development within the PaN community.

2. Stakeholder Engagement

Active engagement with the European and global photon and neutron science community has played an essential role in this project. A brief summary of some of the main engagement activities is given below.

● *Engagement with CalipsoPlus project with a presentation of the Way for Light (https://www.wayforlight.eu) portal from the CalipsoPlus project manager (July 2020).*

The Way For Light portal is a Content Management System and has already an extensive list of photon instruments and experimental techniques. The presentation highlighted the importance of sustainability, governance and the establishment of processes for any changes in the controlled vocabulary used in their project. Each facility has its own representative who is responsible for updating the information present in the web portal.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

The presentation and all relevant documents are available in the ExPaNDS GitHub repository (https://github.com/ExPaNDS-eu/ExPaNDS/tree/master/WP3/20200701-WayForLight-Meeting).
The CalipsoPlus project is mainly focussed on the planning and scheduling of experiments rather than data catalogues. It was found that the taxonomy provided for the experimental techniques did not provide a description of the purpose and context of the experiment, and it was not represented as a FAIR vocabulary [2]. The work in this deliverable re-uses experimental techniques already available in CalipsoPlus and also includes the list of techniques provided by the PaNOSC and ExPaNDS facilities. In addition, it also organises the techniques into a hierarchy of experimental techniques to form a FAIR vocabulary.

● *Engagement with ExPaNDS WP2 in particular task 2.3 and participation in the deliverable D2.2 [3].*

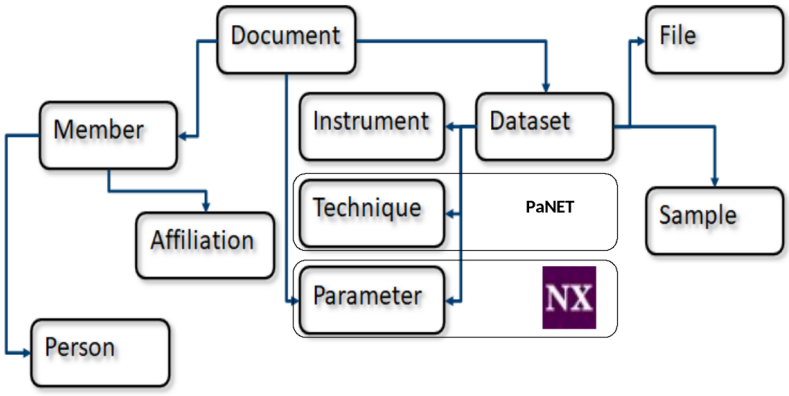
This deliverable complements the work performed on the recommendations for making metadata FAIR with a particular focus on scientific metadata and the required data annotations for the different entities being considered.

● *Survey on how to search for data in a data catalogue*

A significant stakeholder engagement was conducted through this survey. It included the actor type (data owner, non-data owner, future facility user, external Data Consumer, Funder/Policy Maker), the type of data (raw, processed, sample information, beamline, experiment/visit information), the search terms used and the reason for searching. Analysis has been performed and the results were presented in a joint WP3 ExPaNDS and PaNOSC presentation. The survey played an important role in confirming the main uses for data catalogue searches, e.g., data owners and non data-owners searching for both raw and processed data, and the importance of searching with physical metadata such as NeXus fields.

A full report is published in Appendix 1.

● *Coordination of ontology development with PaNOSC through regular meetings.*



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Figure 1: PaN search API Data Model

As the ontology will be implemented on the main data catalogues, close collaboration with the PaNOSC partners is important. Figure 1 shows the PaN search API data model (PaNOSC deliverable D3.1) and the anticipated role of the experimental techniques (PaNET) and Nexus ontologies within that API.

● *Engagement with NCBO BioPortal*

The ExPaNDS ontologies as an EOSC service will be provided by a link to the NCBO BioPortal ontology repository service, which will in turn be registered as an EOSC service. Indeed, our engagement with the University of Stanford led the program managers to agree to register themselves as a service provider to the EOSC and then onboard both BioPortal and Protégé. We believe this will be a valuable addition to the EOSC marketplace and will enable us to expose our ontology as an EOSC service.
The main advantage is the sustainability/re-use of the ontology as the IT infrastructure is provided by BioPortal. While NCBO Bioportal is a highly suitable repository for the present ontologies, due to their inclusion of a significant element of life sciences, it would be beneficial to exploit additional ontology services and repositories in the future.

● *Engagement with EOSC (European Open Science Cloud)*.

EOSC onboarding presentation was organised by ExPaNDS WP1 (August 2020) to determine which ontology repository to use and how the ontology would be integrated into EOSC as a service. BioPortal will start the EOSC onboarding process.

● *Engagement with NIAC (NeXus International Advisory Committee: https://www.nexusformat.org/NIAC.html)*

To discuss governance processes around the creation and sustainability of the Nexus ontology (section 4):

- Introduction HTML anchors for resolvable terminology at NIAC2020 meeting in October and first introduction of the NeXuS ontology.
- January 2021 Telco and March Telco talk about PIDs and usage of namespace nexusformat.
- April 2020 official request to use the namespace.
- May 2021 NIAC created the GitHub NeXusOntology to host the NeXus ontology after formal approval by the committee.

● *Presentation on the application of ontologies and PaNOSC WP3*.

One of the goals was to give an overview of the use of ontologies in other domains.

● *Presentation of the experimental techniques ontology at the Research Data Alliance (RDA) Persistent Identification of Instruments (PIDINST) Working Group session* [5]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

The experimental techniques ontology work was presented at the RDA VP17 PIDINST session 20/3/21 with the title: ‘An ontology for experimental techniques in the Photon and Neutron community’ [5].

● *Collaboration with Instruct-ERIC User Data Working Group on FAIR Data*.

Instruct-ERIC is a PaN-European distributed research infrastructure in structural biology, making high-end technologies and methods available to all European researchers. The Intruct-ERIC Data Working Group is interested in supporting FAIR data but it is mainly focused on the Macromolecular X-Ray Crystallography community. Due to the overlap with ExPaNDS activities, sharing information is important to ensure coordination with other similar projects.

3. Photon and Neutron Techniques Ontology

(IRI: http://www.purl.org/pan-science/PaNET/PaNET.owl)

3.1 Purpose

The primary purpose of the ontologies described in this paper is to support ExPaNDS Work Package 3: EOSC Data Catalogue Services for EU Photon and Neutron National RIs, along with the overall project goal of providing FAIR data for PaN facilities and the parallel PaNOSC work package. Specifically, the programme of developing a common interface to facility data catalogues requires a common vocabulary of terms used to annotate the catalogue and support advanced searches.

One of the key parameters used for labelling and searching of data is the experimental technique employed to collect datasets. The main purpose of the techniques ontology is to provide a common set of standard technique names with global persistent identifiers, along with common alternate names (labels), human-readable annotations and a rudimentary formal semantic description, sufficient to support catalogue services.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

3.2 Design Principles

The approach to ontology design was determined by the primary short-term goals (i.e. catalogue services), the large and growing number of experimental photon and neutron techniques, variants and alternate names currently employed within the domain, and available resources. There was also a strong desire to follow a process that would enable extending the ontology, not only in terms of the breadth of techniques, but also to support future relationships and mapping in anticipation of the increasing role of semantic web technologies in science. Conceptually, we view the experimental techniques as classes, defined by certain properties. For these reasons, we have adopted the OWL language instead of other alternatives such as SKOS (https://www.w3.org/TR/2009/REC-skos-reference-20090818/), although we make use of some SKOS entities in the ontology.

Here, we outline the design principles of the V1.0 ontology. Possibilities for future development are addressed at the end of this section. The key design decisions are as follows:

- The ontology contains only experimental techniques (i.e. every concept in the ontology is a ‘photon and neutron technique’).
- Currently, there is no upper ontology but this could be added in the future to facilitate mapping to other vocabularies.
- Each technique is represented by a owl:Class.
- The ontology focuses on the photon and neutron technique taxonomy, defining how the techniques are related in a hierarchy. Thus, the main relationships used at the moment are subclass (rdfs:subClassOf) and equivalent class (owl:equivalentClass), although the latter is used much less frequently.
- Each term (technique) has a primary label (rdfs:label) which is typically given by space-separated non-abbreviated lower-case name.
- Each term may have as many alternate labels (skos:altLabel) as required.
- Each term may have one or more definitions* (obo:IAO_0000115): ‘The official definition, explaining the meaning of a class or property’
- Each term may have a definition source* (obo:IAO_0000119): ‘Formal citation, e.g. identifier in external database to indicate / attribute source(s) for the definition.’

*We anticipate that additional new properties will be added in the future, extending and possibly replacing those outlined above.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

3.3 Semantic description by multiple named superclasses

The building blocks of our ontology are a small set (currently four) generic technique classes that each correspond to an implied relationship of a particular object property type. The basic classes are chosen to be as 'orthogonal' as possible, i.e. largely uncorrelated. These can be represented as the top levels of a class hierarchy, shown here by their labels:

```
OWL:Thing
      photon and neutron technique
            defined by experimental physical process
            defined by experimental probe
            defined by functional dependence
            defined by purpose
```

(Note that, in this section, we refer to the terms by their human-readable labels rather than their more obscure numerical identifiers).

Here 'defined by experimental physical process' is not the class of techniques that are based on a physical process (indeed, they all are!) but rather the class of techniques that are defined in an *essential* way by a particular physical process. For example, 'scattering technique' is the class of techniques that are based on the physical process of scattering. 'imaging', on the other hand, is a class of techniques that can utilize various physical processes (scattering, absorption etc) and is not, therefore, in the class 'defined by experimental physical process'.

The label 'defined by experimental probe' describes the class of techniques where the probe type is fixed, e.g. 'x-ray probe' is the class of techniques that always use an x-ray probe. Thus, 'x-ray diffraction' is a subclass of this class but 'diffraction' is not.

The class 'defined by functional dependence' refers to techniques where the measurement (not necessarily the final processed data) has a well-defined functional dependence, i.e. the measurement is made *versus* something (position, energy *etc*). The class of techniques 'x-ray spectroscopy' is a subclass of 'versus energy' because the measurement is always carried out as a function of energy, even though some x-ray spectroscopy techniques have a purpose of measuring atomic positions. (Note that we define 'spectroscopy' as being equivalent to 'versus energy').

Finally, 'defined by purpose' is the class of techniques that are defined by what they are for, such as 'obtain atomic structure', 'obtain electronic density of states' etc. For example, we define 'crystallography' as being equivalent to 'obtain crystal structure' which, in turn, is a subclass of 'obtain atomic structure'.

While these four first-level classes are by no means exhaustive, we find that most of the techniques identified can be given a reasonable if rudimentary description as subclasses of several of these. Moreover, they can be added to as and when required.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

We next create subclass trees of the basic classes to construct more specialised terms (i.e. subsets of techniques), and combine terms by asserting that a technique class is a subclass of multiple classes (i.e. intersection of sets of techniques). The techniques so-created can then be further refined by additional single or multiple subclass assertions thus using multiple inheritance.

Rather than displaying the entire class tree here, we show the second-level technique branches and one example of a complete description:

```
photon and neutron technique
      defined by experimental physical process
            absorption technique
            refraction technique
            force measurement
            emission technique
            dispersive technique
            scattering technique
            resonance phenomenon
            reflection technique
            interferometry technique
            propagation technique
            magnetism technique
            nonlinear interaction
       defined by purpose
            obtain spatial map
            characterize excitations
            chiral determination
            obtain atomic structure
            drug fragment binding
            obtain dynamics
            testing
            manufacturing technique
            medical application
            obtain internal field
            obtain electronic ground state properties
            therapy
      defined by experimental probe
            photon probe
            scanning probe
            solid probe
            neutron probe
            microfocussed probe
            muon probe
            pulsed probe
      defined by functional dependence
            versus polarization
            versus emission mass
            versus energy
            versus time
            versus momentum transfer
            versus position
            versus sample state
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

The example we chose to illustrate the entire subclass chains is ‘angle resolved photoemission spectroscopy’ (section 3.4):

```
photon and neutron technique

       defined by experimental physical process
                     emission technique
                            electron emission technique
                                   photoelectron emission
                                          angle resolved photoemission spectroscopy

       defined by purpose
                     obtain electronic ground state properties
                            obtain electronic band structure
                                   angle resolved photoemission spectroscopy

       defined by experimental probe
                     photon probe
                            photoelectron emission
                                   angle resolved photoemission spectroscopy

       defined by functional dependence
                     versus energy
                            versus emitted energy
                                   versus emission momentum
                                          angle resolved photoemission spectroscopy
```

The class trees are fairly straightforward and largely orthogonal, except for ‘defined by experimental probe’. Here there is a correlation between the physical process and probe, as ‘photoelectron emission’ is both a physical process and one that requires a specific probe type.

While the basic techniques are created from their building blocks, new techniques are increasingly defined as simple specializations of existing techniques. For example, while

```
‘nano angle resolved photoemission spectroscopy’
```

is a very complex technique, its description builds on existing definitions, *i.e.*

```
‘angle resolved photoemission spectroscopy’ AND ‘nanofocussed probe’
```

Here, ‘AND’ is essentially the intersection of the two sets (classes) of technique such that the new technique is defined as being a subclass of both of the parent techniques.

Similarly,

```
‘diffraction imaging’
```

is simply

```
‘diffraction’ AND ‘imaging’
```

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

These can be created very quickly and easily, while inheriting all of the properties of the super-classes in the hierarchy. It should be noted that there is no requirement for any new techniques to be formed by an intersection of subclasses of all the top four basic classes shown above. Many techniques use fewer and some use more than one subclass from a particular top-level class.

While we mainly use subclass relationships to define new techniques, it is sometimes extremely useful to use equivalent classes. As well as linking concepts to those in other ontologies, equivalent classes can be used to provide more search 'hits'. For example, by defining `'x-ray spectroscopy'` as being equivalent to `'x-ray probe' AND 'spectroscopy'` an ontology reasoner can infer that any techniques that is a subclass (child) of both `'x-ray probe'` and `'spectroscopy'` is also a subclass (child) of `'x-ray spectroscopy'` even though this was not asserted.

3.4 Example

As an example, we continue with a technique with a relatively rich definition: 'angle resolved photoemission spectroscopy'.

The entry (see section 3.6 for implementation) for this technique contains:

Alternate label(s): `'ARPES'`

Superclasses (definition of technique):
`subClassOf: 'photoelectron emission', 'versus emission momentum', 'obtain electronic band structure'`

Formal Citation (wiki page):
`https://en.wikipedia.org/wiki/Angle-resolved_photoemission_spectroscopy`

This leads to the following entry in the OWL ontology (shown for convenience in turtle format):

```
###  http://purl.org/pan-science/PaNET/PaNET01089
<http://purl.org/pan-science/PaNET/PaNET01089> rdf:type owl:Class ;
      rdfs:subClassOf     <http://purl.org/pan-science/PaNET/PaNET01058> ,
                          <http://purl.org/pan-science/PaNET/PaNET01075> ,
                          <http://purl.org/pan-science/PaNET/PaNET01269> ;
      obo:IAO_0000115 "Angle resolved photoemission spectroscopy"^^xsd:string ;
      obo:IAO_0000119 "https://en.wikipedia.org/wiki/..."^^xsd:string ;
      rdfs:label "angle resolved photoemission spectroscopy"^^xsd:string ;
      skos:altLabel "ARPES"^^xsd:string .
```

Here the identifiers PaNET01058, PaNET01075, PaNET01269 (deliberately made obscure for humans, as per accepted best practice!) represent 'versus emission momentum', 'obtain electronic band structure' and 'photoelectron emission', respectively.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

By inheriting the definitions of just three technique classes, this technique is defined to be a subclass of all of the following techniques, shown with their alternate labels:

```
'angle resolved photoemission spectroscopy'     'ARPES'
'obtain electronic band structure'
'obtain electronic ground state properties'
'defined by purpose'
'photoelectron emission'
'electron emission technique'
'emission technique'
'defined by experimental physical process'
'photon probe'
'defined by experimental probe'
'versus emission momentum'
'versus emitted energy'
'spectroscopy'
'versus energy'
'defined by functional dependence'
'photon and neutron technique'
```

A graphical representation of the class structure is shown in Figure 2.

While it is useful to consider the definition of the technique terms, our primary goal is the support of catalogue services. To this end, we note that a dataset tagged with ‘angle resolved photoemission spectroscopy’ or ‘ARPES’ should be found via a search based on any of the 17 labels and alternate labels above.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

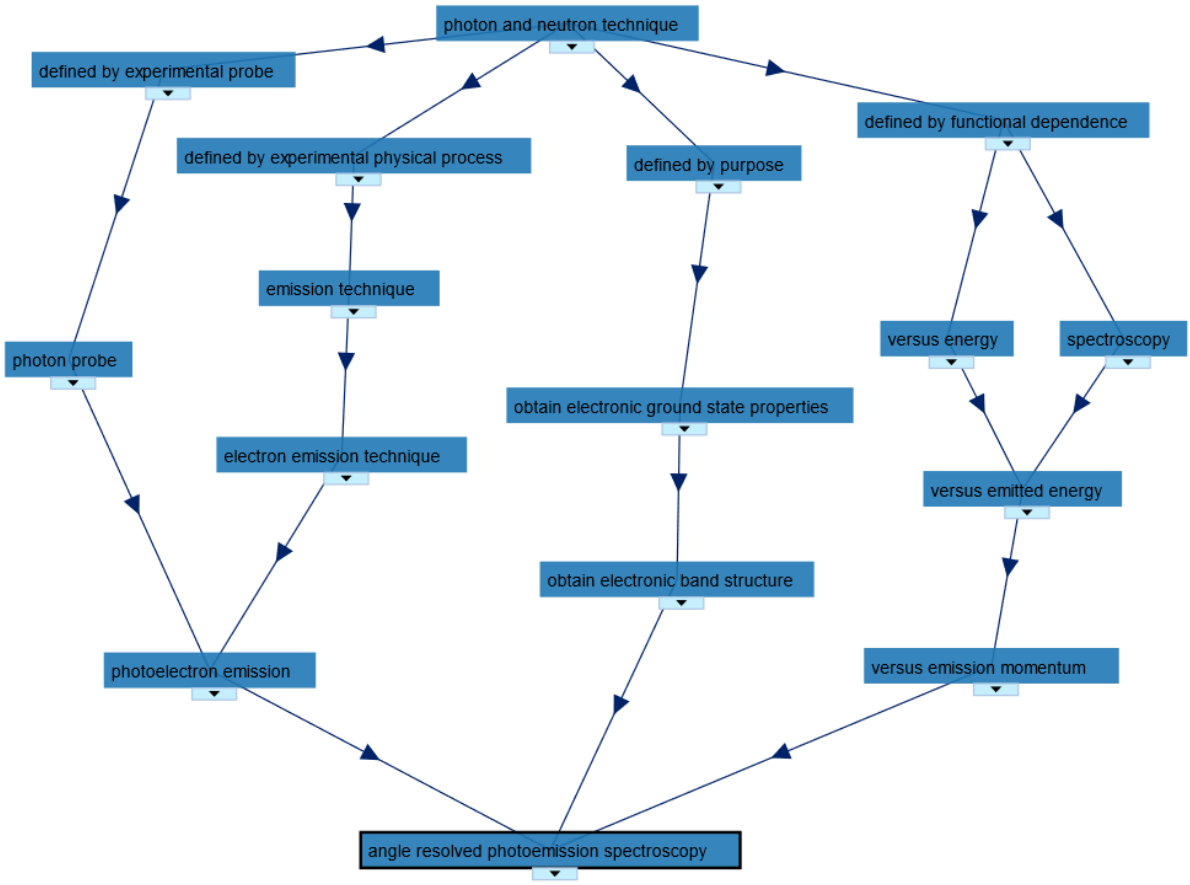


Figure 2: A graphical representation of the subclasses of the technique 'angle resolved photoemission spectroscopy' produced by NCBO BioPortal.

3.5 Namespaces and Identifiers

The requirement for globally unique persistent identifiers suggested the adoption of a purl (http://purl.org) namespace. This allows the specific location of the ontology to be changed without affecting identifiers.

The OWL file that represents the ontology has the Internationalised Resource Identifier (IRI):

http://purl.org/pan-science/PaNET/PaNET.owl

with individual terms (techniques) in the form:

http://purl.org/pan-science/PaNET/PaNETnnnnn

where nnnnn is an integer. The lowest numbers are reserved for the top-level classes, with gaps to allow top-level classes to be added later.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

3.6 Implementation

The ontology is expressed in OWL (Web Ontology Language). However, while standard tools for ontology representation and design (e.g. Protégé [4]) are extremely powerful, they represent a learning curve for most casual users. This, along with the fact that our ontology has a very simple structure, led us to adopt the ROBOT (Robot OBO Tool [6]) template to create and update the ontology, where the technique descriptions are entered in a spreadsheet. The ROBOT tool then creates the ontology OWL file, which is merged with a very small OWL ontology containing some basic ontology metadata. The spreadsheet is an excellent tool to discuss the terms and structure of the ontology with scientists.

The resultant OWL file, along with all input files, is maintained on a public GitHub repository: https://github.com/ExPaNDS-eu/ExPaNDS-experimental-techniques-ontology

The PaNET OWL file is also uploaded to the NCBO BioPortal ontology service, which is a service registered with EOSC. Registering the ontology with a suitable repository service is very important (see Section 6, Rule 8), as it allows the ontology to be discovered and mapped to other ontologies, as well as providing visualization and API functions. While BioPortal is aimed mainly at biomedical ontologies, it is well established and eminently suitable for the ontologies described in this work, which include a significant element of life sciences.

The ontology is openly available to registered users and can be found by searching for ‘PANET’ under ‘Find an ontology’. Current and development versions, along with version history, can also be found in BioPortal. The ontology can be explored using graphical tools (see Figure 2), showing mappings to syntactically-related concepts and exposing annotations such as Wikipedia entries describing the techniques. More importantly for our primary application, BioPortal provides a REST API and SPARQL endpoint to allow programmatic access to terms in the ontology.

Alternatively, the ontology can be accessed directly via its persistent identifier that redirects to the GitHub repository. Programmatic access can be achieved using a suitable library such as the owlready2 Python library [8].

As the primary purpose of the ontology is to support PaN catalogue services, immediate requirements might include:

- Find a technique identifier from a given technique label (or alternate labels).
- Find all the superclasses (identifiers) of a technique from a given identifier.
- Find all the labels (and alternate labels) of a technique from a given identifier.

These can be achieved, for example, with a few lines of Python code via the BioPortal API or the Github OWL file via owlready2 (a Python library for using OWL ontologies).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

For the ontology to be FAIR [2], it has to be findable and accessible on the web, and the individual ontology terms should be accessible via their IRI, i.e. typing a term IRI into a web browser returns some basic information about the term and links to documentation and the full ontology. For the PaNET ontology, we achieve this via publication of the ontology on the web, and redirection to its documentation produced automatically with the WIDOCO ontology documentation tool [9].

Finally, the PaNET ontology contains a very small dummy dataset as an example of use of the ontology for catalogue services. These are implemented as owl:NamedIndividuals which are 'tagged' with specific techniques using the PaNET:hasTechnique object property.

3.7 Update and maintenance workflow

For the V1.0 PaNET ontology we have included essentially all the techniques and alternate names (apart from some trivial variations such as capitalized names) that have been provided by the ExPaNDS/PaNOSC community, notably HZB, PSI and Diamond. While these cover most of the standard techniques, there will inevitably be new techniques to add. In particular, we have focused on PaN techniques, with little from muon and Electron Microscope (EM) facilities. We therefore provide a workflow for adding to, and modifying, the techniques in the ontology.

Figure 3 provides a release process workflow for any change requested to the PaNET ontology. An Excel spreadsheet is provided in the GitHub repository with the current version of the ontology. Additions and modifications to PaNET can be achieved by downloading the techniques spreadsheet from Github and amending it. The next step is to convert the Excel spreadsheet to an OWL format by using the ROBOT software package. Such modifications will be implemented by the custodians of the ontology, with requests processed via Github issues.

Apart from the main ontology generated by the Excel spreadsheet, it is convenient to merge the techniques ontology with a separate owl file containing ontology metadata. Since some of the class relationships are inferred rather than asserted, it is necessary to run an ontology reasoner to make these relationships explicit. As the ontology is rather simple, it is often convenient to run the reasoner before uploading the ontology so that the inferred axioms are accessible directly from the owl file. In order to perform these two tasks, one can follow the instructions below:

- Open PaNET_metadata.owl in Protégé
- Update metadata (version, created etc) and save
- Open PaNET techniques ontology created by ROBOT in current window
- Refactor/merge ontologies; select both ontologies; create new ontology; save file (ontology IRI: http://www.purl.org/pan-science/PaNET/PaNET.owl)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

- Start reasoner (e.g. HermiT)
- Export inferred axioms as ontology; tick all boxes and save to the final owl file; (Ontology IRI: http://www.purl.org/pan-science/PaNET/PaNET.owl)
- Select file name; save as RDF/XML
- Update the Widoco documentation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

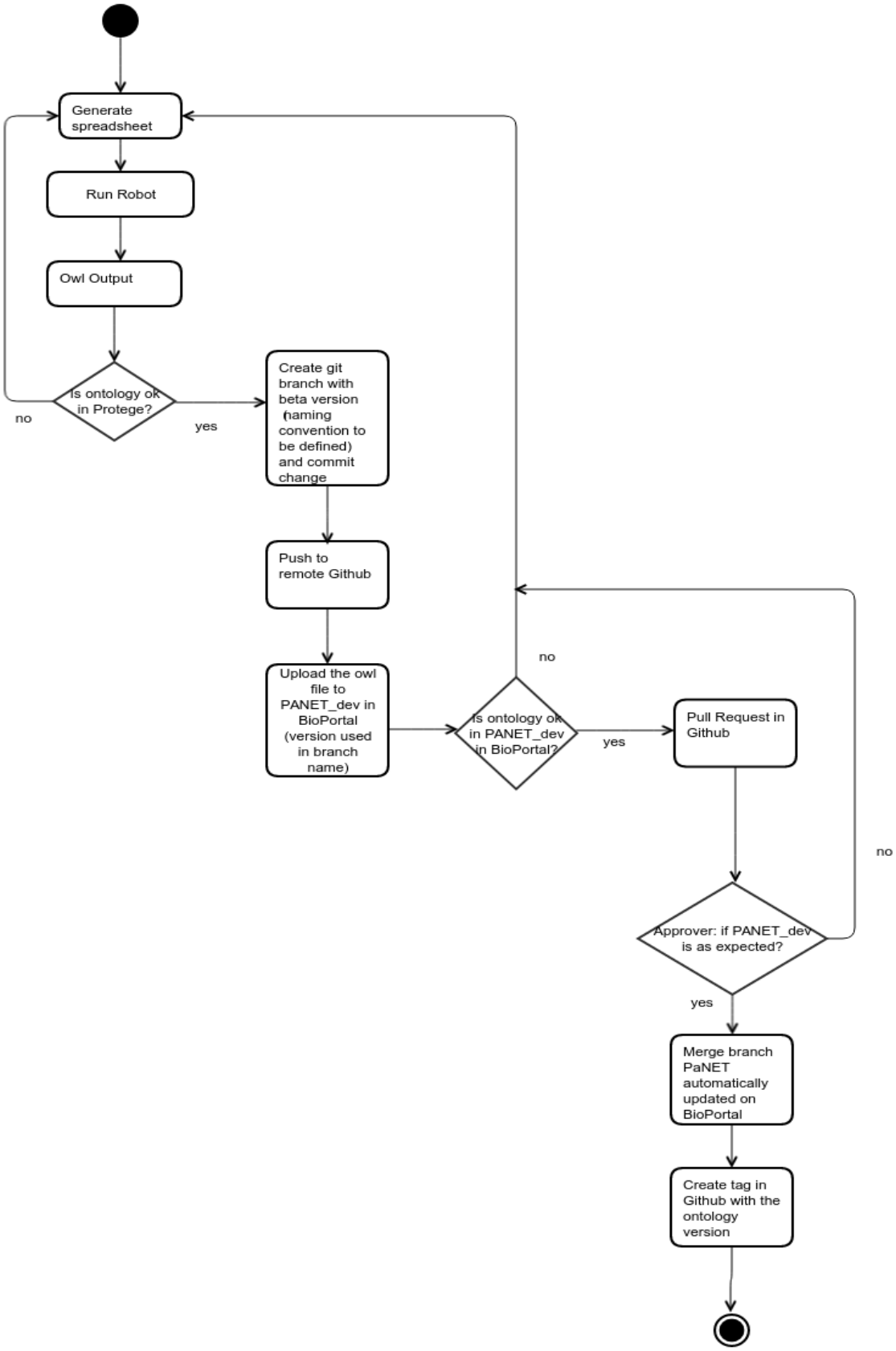


Figure 3: Release process workflow for the PaNET ontology hosted in GitHub.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

When all the steps above have been performed, the changes can be pushed to the GitHub repository by adding a pull request. At the moment, only one reviewer can approve and merge the changes to the master branch. Before any merge to the master branch, it is recommended to upload the new version of the OWL file to BioPortal to ensure that the file is correctly configured and displayed in BioPortal before any release. In each upload, BioPortal will check the validity of the file, convert it to other formats such as RDF and XML. After the merge to master, the PaNET ontology will automatically be updated on BioPortal over the night.

This workflow will be automated in due course using Continuous Integration pipelines in GitHub. Any change to the PaNET ontology will be managed by the PaNET custodians via requests received through the GitHub issue mechanism. For the duration of ExPaNDS, any WP3 member will be able to act as a reviewer and will check the release process workflow described in Figure 3. Before the end of the project, the review process will be finalized as part of the sustainability plans (Work Package 1, task 1.6).

3.8 Future development

We noted in section 3.3, that the top-level technique classes implied a property relationship of a particular type. A significant enhancement of the ontology can be gained by making these relationships explicit. Indeed, one can imagine the following object properties as defining the classes:

```
defined by experimental physical process:
      definedByProcess some experimentalPhysicalProcess

defined by experimental probe:
      definedByProbe some experimentalProbe

defined by functional dependence:
      definedByFunctionalDependence some physicalParameter

defined by purpose:
      definedByPurpose some experimentalPurpose
```

where the classes `experimentalPhysicalProcess` *etc* are further restricted by requiring that they have a role as an experimental evaluant (`'has role'` some `'evaluant role'`) *etc*. The adoption of nomenclature borrowed from Ontology for Biomedical Investigations (OBI) [7] is not accidental: there are close relationships between the current model and that of OBI. Future development of the underpinning object property relationships might profit from aligning with the OBI ontology.

We note that, while the 'domain' of these object properties is within the V1.0 ontology (i.e. 'photon and neutron technique'), the 'ranges' are not. Not only are new classes required, but as these are not within the PaN domain, constructing such relationships will require a very significant connection to external ontologies. This represents both an opportunity and a challenge.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4. NeXus Ontology

(url for the current version of the ontology, pending transfer to NeXus NIAC: https://raw.githubusercontent.com/ExPaNDS-eu/ExPaNDS-nexus-ontology/master/source/NeXus.owl)

4.1 Purpose

While the primary purpose of the NeXus ontology (NeXusOntology) is largely the same as the PaNET techniques ontology, i.e. to support EU photon and neutron data catalogues, its structure and logic are very different.

NeXus is a common data format for neutron, x-ray and muon science and includes an extensive and extensible metadata model [10]. While other data formats are in use in some science areas, NeXus is a generic data format designed specifically for the global PaN community, is very widely adopted, and most PaN research facilities either implement NeXus or plan to do so. We therefore need to include NeXus in any effort to enrich facilities data to make it more FAIR. NeXus enjoys a well-established governance structure in the form of the NeXus International Advisory Committee (NIAC: https://www.nexusformat.org/NIAC.html).

The NeXus metadata model is described in detail in the literature [10]. Salient features include the use of NeXus 'base classes' to describe concepts (largely instrumentation components but also 'users', 'transformations' and other abstract concepts). The base classes provide schemas for sets of key-value pairs, with defined names, forming a separate dictionary or namespace for each class. In order to manage the process of defining which classes must be represented in a particular dataset, NeXus 'application definitions' can be adopted if required. Application definitions aid interoperability by allowing datasets to be validated against particular workflows.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

An illustrative example of a small part of a hypothetical NeXus structure (a mixture of photon and neutron techniques, for the sake of impartiality!) is:

scan123:NXentry
    definition
    sample:NXsample
        chemical_formula
        temperature
        beam_at_sample:NXbeam
            incident_wavelength
            flux
    instrument:NXinstrument
        undulator:NXinsertion_device
            harmonic
            type
            beam_at_undulator:NXbeam
                flux
        moderator:NXmoderator
            temperature
        aperture1:NXaperture
        detector1:NXdetector
            x_pixel_size
    data:NXdata

An adequate description of the NeXus structure and metadata model is outside of the scope of this document, but we pick out a few key features. A NeXus file has a tree structure, with each branch containing named instances of NeXus base classes (of the form instance_name:NXclass_name). These can, in turn, contain other base class instances as well as NeXus fields.

The NeXus fields are key-value pairs, where the key name and its meaning are defined in the base class definition XML files, maintained by the NIAC. The values are either single data elements or arrays, of specified types (e.g. NXfloat) that have counterparts in standard data types. The latter are defined by NeXus attributes, which also give the 'unit category' such as NXenergy. The specific unit is given as a string and is not specified by the NeXus definitions.

As an example, 'incident_wavelength' may have a value 1.54, type 'NX_FLOAT', unit category 'NX_WAVELENGTH' and unit 'angstroms'.

In the example above, we have an NXentry (essentially a measurement), which contains three base class instances. NXdata contains the ‘plottable data’. NXsample and NXinstrument describe the sample and instrument, and each contain fields and other base class instances.

The field names are specific to the class, i.e. 'flux' is defined by NXbeam and has no meaning outside of NXbeam (that is, it is limited to the namespace of NXbeam). We see two

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

fields called temperature. One is defined within NXsample (the sample temperature) and the other within NXmoderator (the moderator temperature). In order to flatten the names into a single vocabulary, it is necessary to specify not only the name, but which class it belongs to.

Notice that there are two instances of NXbeam, giving a description of the beam at two places: the sample and the undulator. Here, the flux fields in each share the same semantics but the fact that there are two of them poses a challenge to database ingress. Non-standard field names can be used but they have no formal semantics.

Finally, we note that the NXentry instance contains the field 'definition'. This is the name of a NeXus Application Definition that specifies what the entry must contain in order to be successfully validated against the definition.

In terms of the FAIR data principles, NeXus serves mainly to support Interoperability and Reusability. To enhance all aspects of FAIRness (particularly Findability), there is a need to incorporate NeXus metadata in PaN data catalogues. Specifically, there is merit in exposing metadata that are common to many or all experimental instruments.

The purposes of the NeXusOntology can be summarized as:

- To provide a single controlled vocabulary of NeXus terms (base class and field names) by flattening and joining the separate namespaces of base classes in a consistent and reversible manner.

- To provide global persistent identifiers for each NeXus term.

- To describe the key NeXus concepts and relationships in a single ontology, linking to existing NeXus annotation and documentation, effectively providing a NeXus explorer tool.

- To reflect and formalize the intended semantic of NeXus but add no additional interpretation.

- To allow the ontology to be updated automatically following publication of new classes after approval of the NIAC

- To utilize a framework (i.e. OWL) that allows separate ontologies to provide mappings and relationships to terms in other vocabularies and ontologies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

4.2 Design Principles

NeXus supports a hierarchical structure of base classes. For example, the base class that describes a whole instrument – NXinstrument – contains within it base classes describing components of the instrument, e.g. NXmonochromator, NXmirror etc. In an ontology of an instrument one might utilize object properties such as 'hasPart' that describes the relationship between NXinstrument and NXmonochromator. Such a relationship is not explicit or formalized within NeXus, and although there is a common-sense understanding that certain concepts represent, for example, part of certain other concepts, employing such relationships would require additional interpretation and community agreement. As the goal of the current project is to represent the semantics as defined by the NIAC, and expressed as a set of NXLD XML files (in turn, structured by an XML schema) the only relationships in our ontology are those defined formally by the NeXus definitions. These concern the structure of the NeXus files and the annotation to describe its components.

NeXusOntology is not an ontology of an instrument, or a representation of a NeXus file, but an ontology of the NeXus definitions. A positive consequence of adding no new semantics to NeXus is the fact that it then becomes possible to create an ontology automatically from the NXDL definition files.

The design of the ontology is governed by the need to describe the logic of NeXus within the language of OWL, using OWL constructs in the way that best represents NeXus concepts. In some cases this mapping is obvious, while other aspects require explanation. To understand NeXusOntology one must first have at least a rudimentary grasp of NeXus.

The NeXus ontology does not utilize an upper ontology and the top-level classes are:

```
Owl:Thing
      NeXus
            NXobject
                  NeXusApplicationDefinition
                  NeXusBaseClass
            unitCategory
```

where we underline labels that are defined in NeXus. Here 'NeXus' is the class of all NeXus concepts. We find next, two classes: NXobject, and unitCategory which we will return to later. NXobject is defined as the 'base object of NeXus' and the starting point for defining NeXus base classes and application definitions. As NeXus base classes and application definitions both 'extend' NXobject, but are disjoint, we represent them as direct subclasses of NXobject.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

We next represent the complete set of application definitions and base classes as subclasses of NeXusApplicationDefinition and NeXusBaseClass, respectively. (Here we note that some application definitions and, in principle, base classes 'extend' other classes. This relationship is not currently reproduced in the class structure although it is made explicit in the annotation). The fifth and final level of classes within the ontology are therefore the base classes and application definitions themselves:

```
NXobject
      NeXusApplicationDefinition
            NXarchive
            NXarpes
            …
            NXxrot
      NeXusBaseClass
            NXaperture
            NXattenuator
            …
            NXxraylens
```

Here, we note that all subclasses of NXobject begin with capital 'NX' as required by NeXus. We thus have a complete list of NeXus base classes and application definitions, to which we assign global persistent identifiers. We reproduce the annotation from the NeXus definitions (NeXus .nxdl xml files from nexusformat GitHub). We also provide links to the NeXus html manual pages, employing NeXus html anchors to resolve to the relevant term.

Our next task is to describe the properties of the application definitions and base classes. Since both of these can contain NeXus base classes, described in the NeXus documentation as 'Groups cited', we introduce a 'citesGroup' object property with domain 'NXobject' and range 'NeXusBaseClass'. NeXus base classes and application definitions are then asserted to be subclasses of 'citesGroup some NeXusBaseClass'. For example, NXarpes is a subClass of:

```
citesGroup some NXdetector
citesGroup some NXinstrument
```

etc.

We now turn to the most challenging and important aspect of the ontology: representation of NeXus fields in base classes. Because each NeXus field (the key in key-value pairs) exists only within the namespace of the NeXus class in which it appears, it is not possible (at least, without additional semantic interpretation) to create a single flat list of field names with a guarantee of uniqueness. It is therefore necessary to create compound names by joining the NeXus class and field names. For example, the field 'flux' defined within the base class 'NXbeam' is given the full name 'NXbeam-flux' (label: 'NXbeam flux') and with it, a global persistent identifier:

http:/purl.org/nexusformat/definitions/NXbeam-flux

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

This entity, a subproperty of NeXusField, creates a relationship between a NeXus class and some kind of quantity or measurement, expressed as a NeXus unit category (unitCategory):

```
NXobject NeXusField some unitCategory
```

This is far from intuitive so we illustrate the principle with some examples. For NXbeam, we have

```
‘NXbeam distance’ some NX_LENGTH
‘NXbeam energy_transfer’ some NX_ENERGY
```

etc.

The first statement tells us that the NeXus base class NXbeam contains a field ‘distance’, to which we assign the label ‘NXbeam distance’. Furthermore, ‘NXbeam distance’ connects the base class to a type of unit category – in this case, NX_LENGTH. The unit categories describe a measure by having both a unit and value:

```
NX_LENGTH hasUnit some xds:string
NX_LENTH hasValue some datatype
```

Thus, a NeXus field is expressed as an object property that relates a base class to a measure (unit category), which in turn has both a unit string and data value. (Here datatype means some unspecified data type). It is worth pointing out that NeXus deals with units in a rudimentary way and does not include an ontology of units, conversions *etc*.

Finally, the ontology reproduces the documentation text from NXDL files and provides links to html manual pages. The NeXusOntology thus provides a useful NeXus definition explorer tool. Screen captures of a small part of the NeXus ontology in Protégé are shown in Figures 4 & 5.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

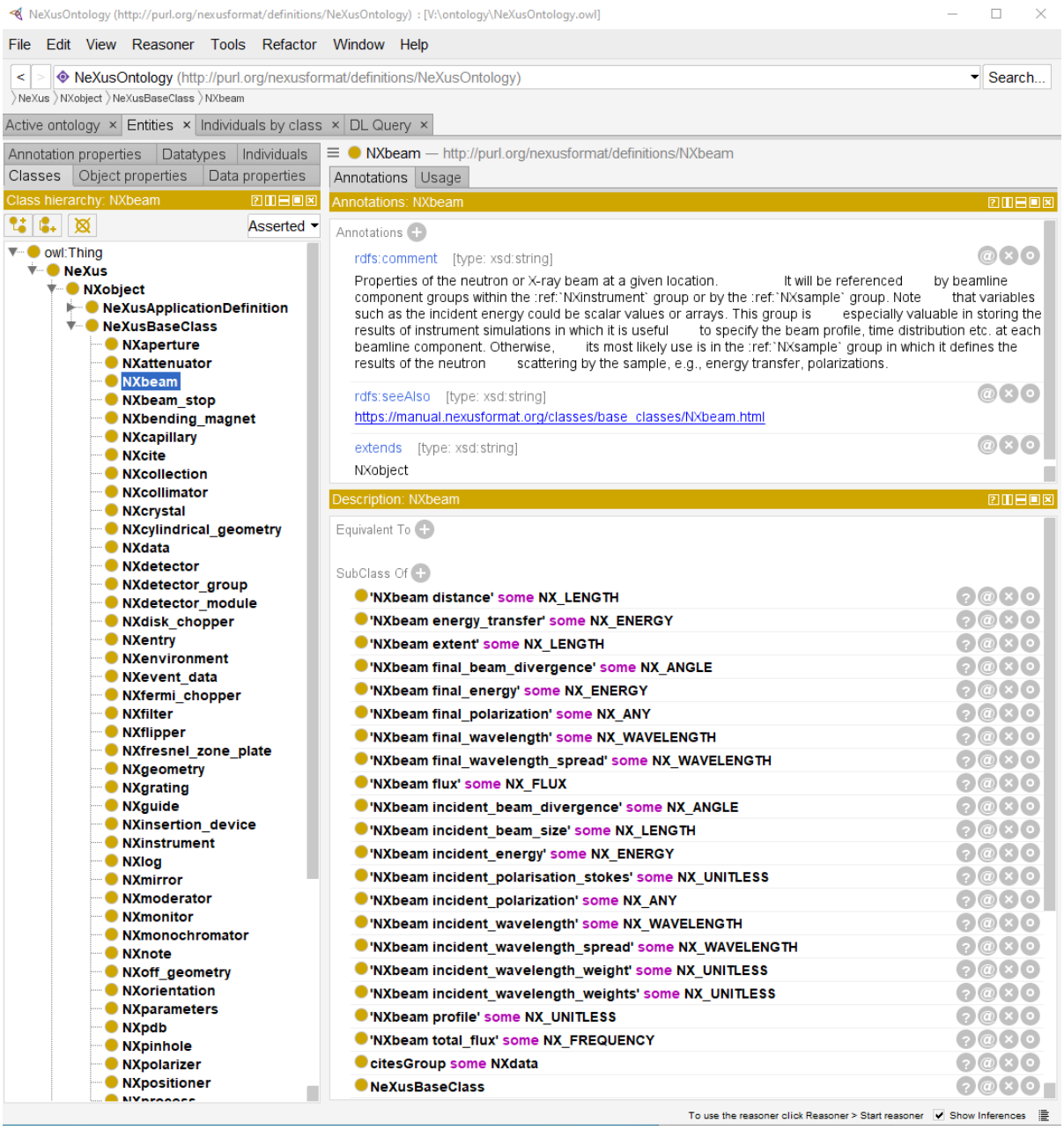


Figure 4. A screen capture of a small part of the NeXus ontology in Protégé. This view shows classes in the ontology. Here we see some of the NeXus base classes. NXbeam is selected and displays annotation (top-right): NeXus doc string, link to NeXus manual page *etc*. The class description is shown in the bottom-right. Here we find the NeXus fields and related unit category.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

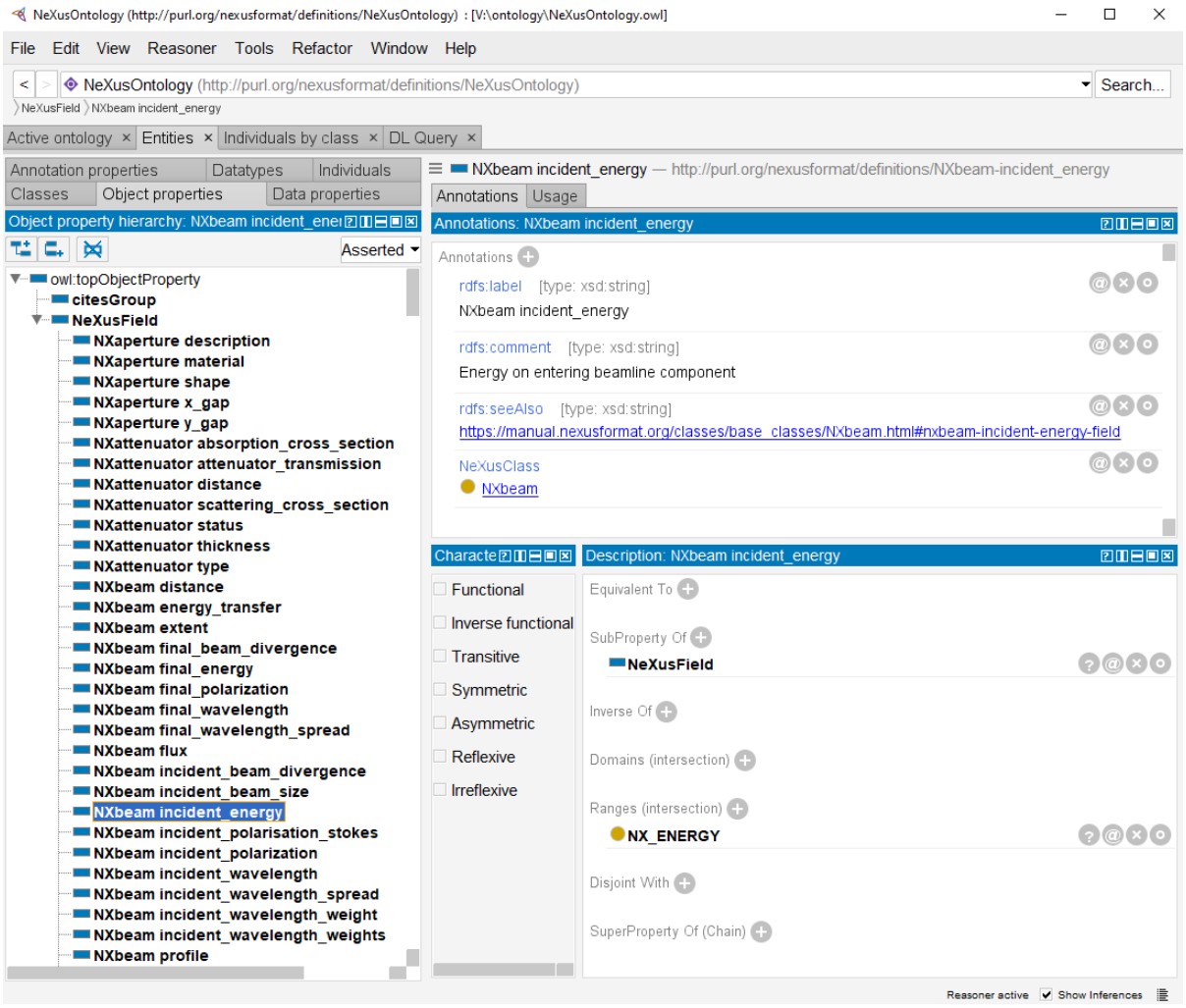


Figure 5. A screen capture of the NeXus ontology object properties in Protégé. Field labels (adjoined NeXus base class and NeXus field names) are shown on the left. On the right we see the purl PID for the field. Below this are annotations (here, the NeXus manual link navigates directly to the relevant part of the manual page by using xml anchors). The description (bottom right) indicates that the range of this property is the NeXus unit category NXenergy.

The NeXus ontology contains a very small dummy dataset as an example of use of the ontology for catalogue services. This is not considered to be part of the main ontology.

4.3 Namespaces and Identifiers

For consistency with the PaNET techniques ontology, we adopt purl as a global namespace. The NeXus International Advisory Committee (NAIC) will reserve the namespace nexusformat for NeXus-related use. The OWL file that represents the NeXusOntology will have the iri:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

http://purl.org/nexusformat/definitions/NeXusOntology.owl

with individual terms in the form:

http://purl.org/nexusformat/definitions/NXsample
http://purl.org/nexusformat/definitions/NXsample-density

etc.

4.4 Implementation

The OWL ontology is created automatically by a Python script that parses the NXDL files and converts it into an OWL file, with essentially no user input. There are two steps to this process. First, the GitHub API is used to obtain a listing of the urls of the NeXus NXDL files, which are then parsed using the Python minidom module, creating dictionaries of NeXus definitions. In the second step, the Python dictionaries are used to create the ontology OWL file using the owlready2 module [8].

The Python script will be available from the github site:
https://github.com/nexusformat/NeXusOntology

The community maintenance workflow is currently being refined in collaboration with the NIAC.

4.5 Future development

There are three key elements to future development. First, once the NIAC has approved and published new or modified NXDL definitions, these can be used to create a new ontology and new identifiers. The second potential area of development concerns the representation of the NeXus definitions in the ontology. Currently, some aspects of the definitions, such as subclassing via ‘extends =’ are not formally represented in the subclass structure (although they are shown in the annotation properties). Finally, the logic or structure of NeXus itself may evolve. This will require modification of the Nexus ontology in collaboration with the NIAC.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

5. Semantic integration

(https://github.com/ExPaNDS-eu/ExPaNDS-nexus-dcat-mapping-ontology)

Different communities have developed their terminology and are maintaining their metadata schemata or lists of vocabulary. In order to facilitate sharing of data across communities and allow collaboration between heterogeneous systems, interoperability of these vocabularies is required. In order to increase data interoperability and enable data integration, these vocabularies need to be semantically integrated with other vocabularies and their terms have to be aligned. This enables the development of a common shared vocabulary.

The NeXus format has been created to describe a measurement and some immediate processing steps in PaN facilities. In the NeXus format there are descriptions of concepts that are unique to PaN facilities. Discipline specific terminology elements and conditions are described that are required to create an instrument e.g. from the source to the detector and the sample environment, together with conditions e.g. pressure, temperature and the beam. Other concepts and items go beyond parts of instruments and are of a more general nature e.g., users/persons, geometries, materials, environmental conditions.

Normally an ontology is created to explain things that are part or connected to other things that might have been described in other ontologies. These ontologies might overlap, but each ontology has certain goals when being created therefore 100% overlaps are rare.

In NeXus, the base classes NXuser and NXentry contain fields with bibliographic and context specifications. The concepts of these specifications are required for filling demands concerning interoperability between repositories. There are a number of ontologies that play an important role for non-discipline specific integration and mostly representing bibliographic information. Prominent ones include:

Dublin Core (https://dublincore.org/specifications/dublin-core/dcmi-terms/),
FOAF (http://xmlns.com/foaf/spec/),
schema.org (https://schema.org/) ,
PROV-O (https://www.w3.org/TR/prov-o/)

and in the context of the EOSC and integration with data catalogues, DCAT v2 (https://www.w3.org/TR/vocab-dcat-2/) is gaining importance. DCAT v2 is meant for representing Data Catalogues, and integrates all the before mentioned ontologies. DCAT will be integrated in repositories as e.g. B2FIND, schema.org and DublinCore terms can be integrated in the landing pages of data catalogue publications.

On a disciplinary level there are a number of ontologies in chemistry and materials. For none of them could we find concrete usage scenarios where any integration on a larger level was an immediate demand from domain experts. There might be, and the demands will surely pop up in the next few years. But for now there was no outstanding ontology required for integration. Having the Nexus format defined in an ontology is the first step to allow semantic integration with other ontologies.

In order to achieve semantic integration of ontologies the usage of the Simple Knowledge Organisation System (SKOS) or the Web Ontology Language (OWL) and Resource Description Framework (RDF) vocabularies is one possible way to achieve this goal. Using OWL and RDF equivalent and sub-class relations between classes can be defined *owl:equivalentClass* and *rdfs:subClassOf*, SKOS can be used to create hierarchies by using *skos:broader/skos:narrower* or mappings using skos:mappingRelation, skos:closeMatch, skos:exactMatch, skos:broadMatch, skos:narrowMatch, or skos:relatedMatch.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

In order to integrate the NeXus format with non-disciplinary ontologies the NXuser and NXentry have been related to the classes in DCAT v2, DublinCore, FOAF and CSMD (http://icatproject-contrib.github.io/CSMD/).

NeXus Term	**Mapping predicat**	**DCAT, PROV-O, CSMD, schema.org or FOAF term**
NXuser	equivalentTo, skos:exactMatch	csmd:User
	subClassOf	prov:Person, schema:Person, foaf:Person
NXuser-address	equivalentTo	schema:address
NXuser-affiliation	equivalentTo	schema:affiliation
NXuser-email	equivalentTo	schema:email, foaf:mbox
NXuser-fax_number	equivalentTo	schema:faxNumber
NXuser-telephone_number	equivalentTo	schema:telephone, foaf:phone
NXuser-name	equivalentTo	foaf:name
NXuser-role*	equivalentTo	dcat:hadRole
NXentry	skos:broadMatch	dcat:Distribution, dc:Dataset
NXentry-title	owl:subDataPropertyOf	dcterms:title

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

NXentry-start_time	owl:subDataPropertyOf	dcat:startDate
NXentry-end_time	owl:subDataPropertyOf	dcat:endDate
NXentry-duration	owl:subDataPropertyOf	dcterms:PeriodOfTime
NXentry-experiment_identifier	owl:subDataPropertyOf	dcterms:identifier
NXentry-experiment_description	owl:subDataPropertyOf	dcterms:description
NXentry-collection_description	owl:subDataPropertyOf	dcterms:description
NXentry-collection_identifier	owl:subDataPropertyOf	dcterms:identifier
NXentry-entry_identifier	owl:subDataPropertyOf	dcterms:identifier

* in order to create the mapping between roles, SKOS Concepts had to be used. Which triggers the creation of individuals for each role. DCAT and NeXus are both having their own role scheme that could partly be mapped.

The related ontology is available at
https://github.com/ExPaNDS-eu/ExPaNDS-nexus-dcat-mapping-ontology
under the CC-BY license in order to comply with the “Ten Simple Rules”. We would ask the broader community to comment on the ontology and raise issues in order to develop it further and continue with the semantic integration.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

6. Adherence to FAIR Vocabulary and Ontology guidelines

6.1 PaNET

In this section we describe how the “Ten Simple Rules for Making a Vocabulary FAIR” [2] have been applied to the development of PaNET.

Ten Simple Rules for Making a Vocabulary FAIR	**PaNET application of the rule**
Rule 1. Verify that the legacy-vocabulary license allows repurposing, and agree on the license for the FAIR vocabulary	Both the spreadsheet and the OWL file have been assigned a CC-BY 4.0 license and are accessible via the IRI and Github repository.
Rule 2. Determine the governance arrangements and custodian of the legacy vocabulary	The vocabulary additions and modifications will be maintained in the spreadsheet, and new versions of the OWL file will be produced by an automated process based on the spreadsheet. Diamond Light Source is the current custodian of the spreadsheet. The community can submit changes or new terms through the GitHub repository, in which we have provided appropriate issue templates. The custodians will address comments and act upon the requests.
Rule 3. Check term and definition completeness and consistency in the legacy vocabulary	All the terms have a label, and some have definitions and definition sources and alternative labels. For those definitions that are missing, they will be added continuously in the development process. In the future, we will also add a definitive reference to the experimental techniques.
Rule 4. Establish a technical maintenance environment for the FAIR vocabulary	The ontology development and maintenance processes are established using GitHub.
Rule 5. Assign a unique and persistent identifier to (a) the vocabulary and (b) each term in the vocabulary	Using purl.org for the ontology and term IRIs.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

Rule 6. Create machine readable representations of the vocabulary terms	See table below.
Rule 7. Add vocabulary metadata	Ontologies contain metadata: License, Version, Creators/Contributors, Creation date
Rule 8. Register the vocabulary	The vocabulary has been registered in the NCBO BioPortal to make it findable.
Rule 9. Make the vocabulary accessible for humans and machines	The vocabulary is accessible via the GitHub interfaces and PURL redirection.
Rule 10. Implement a process for maintaining the FAIR vocabulary	Issue templates have been created to enable the community to submit

Rule 6 - Create machine-readable representations of the vocabulary terms

Identify terms	We have identified most of the techniques being used at the relevant facilities.
Encode term labels and synonyms	We have included term labels and defined synonyms when relevant.
Add textual definitions	Work in progress
Add notes or comments for clarifications	Work in progress
Add codes and symbols	Not relevant
Add notes or comments for clarification	Work in progress
Add per-term metadata if available	Work in progress
Define the hierarchy of terms	Done
Encode relationships between terms	We are only considering a taxonomy for now, and other relationships will be considered later.
Define subsets	Not needed.
Define and document the whole vocabulary	Done.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

6.2 NeXusOntology

In this section we describe how the “Ten Simple Rules for Making a Vocabulary FAIR” [2] have been applied to the development of NeXusOntology

Ten Simple Rules for Making a Vocabulary FAIR	**NeXus Ontology application of the rule**
Rule 1. Verify that the legacy-vocabulary license allows repurposing, and agree on the license for the FAIR vocabulary	The NeXus definitions are available under licence: http://www.gnu.org/licenses/fdl-1.3.txt
Rule 2. Determine the governance arrangements and custodian of the legacy vocabulary	NeXusOntology is auto-generated from NeXus NXDL definition files. These are maintained by the NeXus International Advisory Committee (NIAC). The ontology adds nothing to the semantics or definitions as these are governed by the NIAC.
Rule 3. Check term and definition completeness and consistency in the legacy vocabulary	The term definitions refer to NeXus documentation and annotation. No additional term annotation is provided by the ontology.
Rule 4. Establish a technical maintenance environment for the FAIR vocabulary	NeXusOntology is auto-generated from NeXus NXDL definition files by a Python script. A new ontology will be created by the NIAC when the definitions are changed on the nexusformat Github site. Changes to the conversion of the definitions to the ontology will be carried out by modifying the Python script that generates the ontology. This will also be maintained by the NIAC, at the site: https://github.com/nexusformat/NeXusOntology
Rule 5. Assign a unique and persistent identifier to (a) the vocabulary and (b) each term in the vocabulary	Using purl.org for the ontology and term IRIs.
Rule 6. Create machine readable representations of the vocabulary terms	See table below.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

Rule 7. Add vocabulary metadata	This includes version information, creator, licence etc.
Rule 8. Register the vocabulary	The vocabulary may be registered in the NCBO BioPortal to make it findable. It is currently not published in BioPortal.
Rule 9. Make the vocabulary accessible for humans and machines	The vocabulary is accessible via the GitHub interfaces and PURL redirection.
Rule 10. Implement a process for maintaining the FAIR vocabulary	Updates with new versions of NeXus definitions will be automatic and triggered by the NIAC releasing a new version. Requested changes to the ontology conversion will be via github issues.

Rule 6 - Create machine-readable representations of the vocabulary terms

Identify terms	All terms in the NIAC-adopted definitions are included.
Encode term labels and synonyms	Labels are either copied directly from NeXus or formed by concatenating NeXus names.
Add textual definitions	Automatic from NeXus definitions
Add notes or comments for clarifications	Automatic from NeXus definitions
Add codes and symbols	Not relevant
Add notes or comments for clarification	
Add per-term metadata if available	Only metadata from original definitions are added.
Define the hierarchy of terms	Not relevant
Encode relationships between terms	Only relationships defined by NeXus are included.
Define subsets	Not needed.
Define and document the whole vocabulary	Done.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

7. Other relevant Ontologies

7.1 PaNKOS

The PaNKOS (Photon and Neutron Knowledge Organisation System) Ontology (see https://github.com/ral-facilities/pankos) developed under the PaNdata EU grant, provides a framework for organizing knowledge within the PaN domain. While the current use-case and requirements are more focussed (i.e. PaN data catalogues), the present work builds on PaNKOS by providing a much larger set of terms in more restricted area (e.g. experimental techniques). The structure of PaNET and PaNKOS techniques are a little different, with PaNET describing techniques in terms of their classes while PaNKOS focuses on individuals. However, it is anticipated that important PaNKOS object properties will be used to connect ExPaNDS ontology terms to other vocabularies.

For example, the ‘supportsTechnique’ property (along with its inverse, ‘techniqueOf’) will likely be used to connect PaN instruments to PaN techniques. The PaNKOS#NexusParameters class has an obvious mapping to NeXus fields in NeXusOntology (although the latter is expressed as an object property).

7.2 Science Subject

Two important developments for the PaN domain are the Domain Resource Application Ontology (DRAO, https://github.com/FAIRsharing/domain-ontology) and Subject Resource Application Ontology (SRAO, https://fairsharing.org/bsg-s001177/). DRAO builds on the BFO upper ontology and connects many domain ontologies. One of the most relevant is the Ontology of Biomedical Investigations (OBI) which builds on OBO foundry and includes an experimental ‘assay’ class that has a significant overlap with the PaNET ontology described in this paper, as discussed previously.

A fine-grained description of the subject of an experiment is likely to play a central role in FAIR data, allowing data in very specific science areas to be discovered. PaN techniques are specialist techniques within the PaN community, making the PaN community obvious custodians of the relevant taxonomies. However, the same is not true for the subject of most experiments (unless they are directly in the field of the fundamentals of photon-matter interactions, for example) and so it is appropriate to adopt and extend ontologies developed for specific science subject domains. It is highly desirable to use and extend the SRAO ontology by proposing new terms or entire branches via the SRAO github issues mechanism. We see no merit in developing a new PaN science subject ontology.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

7.3 Sample description

The final essential aspect of the physical description of an experiment concerns the sample. A sample description is part of the NeXus metadata in the form of an NXsample class. While this class includes fields such as ‘name’ (a descriptive name), ‘type’, and ‘chemical_formula’ (using CIF standards), the description is rudimentary and optimized for relatively simple sample types such as crystals. A taxonomy of sample types, each with relevant properties, would be highly advantageous. This could be achieved within the NeXus framework or via a new or existing ontology, if a suitable solution can be identified.

7.4 Instruments

It is highly advantageous to make use of an ontology of instruments, where the instrument types are defined by particular properties, facilitating a search for appropriate instruments. Here, PaNKOS provides a very useful structure, where specific instruments are expressed as individuals of instrument types, whose properties include the facility that they are ‘in’, for example. Building on PaNKOS to develop a more extensive set of instruments would therefore seem appropriate.

At the level of PIDs, the PIDINST (persistent identifiers for scientific instruments) working group of the Research Data Alliance (in which ExPaNDS is represented) has created a schema for registering instruments and providing PIDs. The relationship between PIDINST [5] and PaNKOS requires development. Similarly, the DICE project (https://www.dice-eosc.eu/) includes the task to develop an instrument PID registry that we will follow.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

References

[1] Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, et al. (2017) PLOS Biology 15(6): e2001414. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2001414 https://doi.org/10.1371/journal.pbio.2001414

[2] Ten Simple Rules for making a vocabulary FAIR. Simon J D Cox, Alejandra N Gonzalez-Beltran, Barbara Magagna, Maria-Cristina Marinescu https://arxiv.org/abs/2012.02325

[3] Draft recommendations for FAIR Photon and Neutron Data Management. Daniel Salvat et al. 10.5281/zenodo.4312825

[4] M.A. Musen, The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

[5] Research Data Alliance Plenary 17 (2021). PIDINST Adoption session. PIDINST Adoption | RDA (rd-alliance.org); https://www.rd-alliance.org/groups/persistent-identification-instruments-wg

[6] ROBOT: A tool for automating ontology workflows. R.C. Jackson, J.P. Balhoff, E. Douglass, N.L. Harris, C.J. Mungall, and J.A. Overton. BMC Bioinformatics, vol. 20, July 2019.http://robot.obolibrary.org/

[7] The Ontology for Biomedical Investigations A. Bandrowski et al. PLoS One. 2016 Apr 29;11(4):e0154556. doi: 10.1371/journal.pone.0154556. eCollection 2016. http://obi-ontology.org/

[8] Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. J. B. Lamy.Artificial Intelligence In Medicine 2017;80:11-28 https://bitbucket.org/jibalamy/owlready2

[9] WIzard for DOCumenting Ontologies (WIDOCO), D. G. Verdejo DOI: 10.5281/zenodo.4320190

[10] The NeXus data format. M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, S. I. Campbell, B. Clausen, S. Cottrell, J. U. Hoffmann, P. R. Jemian, D. Männicke, R. Osborn, P. F. Peterson, T. Richter, J. Suzuki, B. Watts, E. Wintersberger and J. Wuttke J. Appl. Cryst. (2015). 48, 301-305 https://doi.org/10.1107/S1600576714027575 https://www.nexusformat.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Appendices

Appendix 1: Survey results analysis

(Task 3.2)

1. Introduction

In December 2019, ExPaNDS WP2 and WP3 conducted a Data Landscape survey to establish a baseline on the current state of the 10 participating institutions on FAIR data policies and data management practices. Some facilities are currently evaluating the existing data catalogues while others have limited metadata available. As a consequence, we decided to do a follow-up survey targeted at instrument scientists, to understand the minimum metadata required in a data catalogue that allows scientists to find data across Neutron and Photon facilities. The goal was to build a set of use cases by considering all facilities. The responses to this survey guided us in assessing priorities for task 3.2.

This is work is highly relevant to the PaNOSC deliverable on the search API, and also ExPaNDS Task 2.3 for defining the recommendations on metadata standards. For the scientific metadata, the ontology will be closely linked to the existing Nexus file format.

2. Survey presentation

The survey presented four simple questions to be easily and quickly filled by the user. The idea was to have open questions (especially for the parameters used in the search) in order to better understand how a defined user will be using a data catalogue when trying to find their data. Having very defined questions could potentially lead to biased answers. Here is the detailed of the question:

- Who is searching for the data: what is their role and motivation (e.g are they the data owner looking to find and reprocess their data or are they the non-data owner wanting to look for additional work). The six possibilities provided are:
  - Facility User (Data Owner)
  - Facility User (Non-Data Owner)
  - Future Facility Owner
  - External Data Consumer
  - Funder/Policy Maker

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

    - Other
- What is being searched for? Is it processed data or some information about a sample? The seven possibilities provided are:
    - Raw Data (e.g Experimental Nexus data file)
    - Processed Data
    - Raw and Processed Data
    - Sample Information
    - Beamline/Facility
    - Experiments/ Visit Data
    - Other
- How are they searching for it? What are the search criteria and, importantly, how might the user refine their search? If they are searching on the sample type, then what sample parameters might they search for? If they are searching by experimental technique, then what information might be most relevant?
- Why are they doing the search? What do they hope to find?

In order to facilitate the analysis, the facility of the user was a mandatory field while the email and the comments were optional. Microsoft Forms was used to design and distribute the survey. An email with a link to the survey and a set of use cases as example was sent to the all-expands mailing list and PaNOSC Project Coordinator. Furthermore, to include data miners in our use cases, this survey was distributed to the EOSC community, more specifically:

- to INFRA-EOSC 5 Projects,
- the EOSC liaison platform
- the EOSC Service and Research Product Catalogues Interest Group.

3. Survey results analysis

In total, the number of responses were 52. The respondents were mainly instrument/beamline scientists, but we had two external responses as well. Table A1 summarizes the number of responses per affiliation type. Synchrotrons facilities are largely represented in the survey responses (81%), this may be explained by the fact that many ExPaNDS contributors are associated with this facility type.

Affiliation Type	**Number of responses**
Synchrotron	42
Neutron/Muon	3
FEL	5
External User (University)	2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Table A1: Number of responses by affiliation type

The survey was mainly directed to scientists as they will be the primary user of the system. Furthermore, in this task we were mainly interested in the scientific metadata that could potentially be extracted from the NeXus file. However, the Communication teams from various facilities have shown interest on data catalogues in order to:

- Link data with journal publications
- Perform “business” analysis that could be useful for Communication and User training
- Apply for funding.

1. *Who*

Scientists could be divided in various categories depending on their role in the research lifecycle. A scientist could be involved or participate for example:

- in an experiment proposal (Data owner or Data creator).
- on supporting an experiment (non-data owner or Data contributor).
- in the submission of new proposals to help their research (future user).
- in what other facilities are doing to compare with results from their own research (External consumer).
- in applying for research funding. In this case, they have experience on requirements from Journal Editors or Funder/Policy Maker.

User Type	**Number of responses**
Facility User (Data Owner)	28
Facility User (non-Data Owner)	17
External Consumer	6
Future User	5
Funder/Policy Maker	3
General Public	2
PhD student	1
Journal Editor	1

Table A2: Number of responses by user type

The Facility User type (both Data Owner and non-Data Owner) are largely represented in the responses. In EOSC, it is expected that those users (Data creators and contributors) will be mainly using the services provided by the federated metadata catalogues as they already

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

know the data collected and the context of the experiment. The main advantage is that they will be able to search and analyze data collected in multiple facilities using a single interface.

External consumers, funders, journal editors and members of the public will be looking for public data provided for example by a data repository such as Open Aire or through a DOI from a journal article.

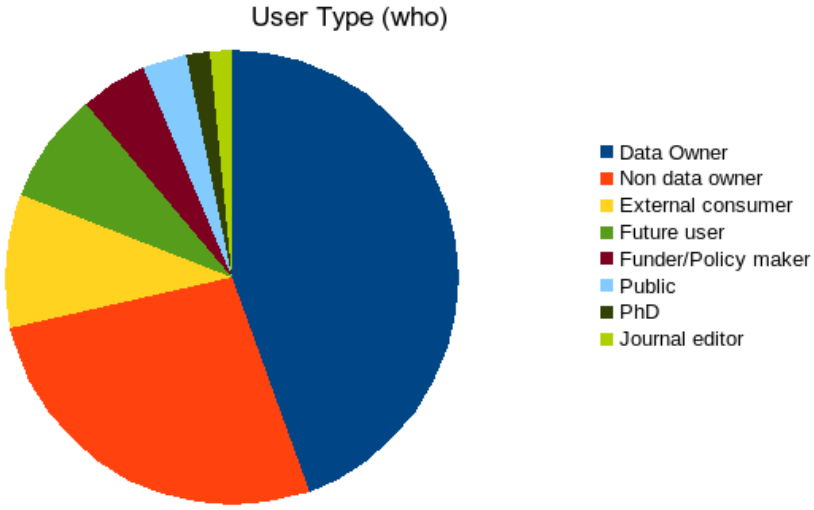


Figure A1: Pie chart representing the proportion of user types in the survey responses.

2. *What*

According to the survey, the main reason for using a data catalogue is to access raw and processed data. A few users would like to access information on beamline/facility, experiment/visit or sample. Here, both raw and processed data are equally important.

What a user is searching for	**Number of responses**
Raw data	33
Processed data	36
Beamline/Facility	4
Experiment/Visit	4
Sample information	4

Table A3: Number of responses per item that a user is searching for

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

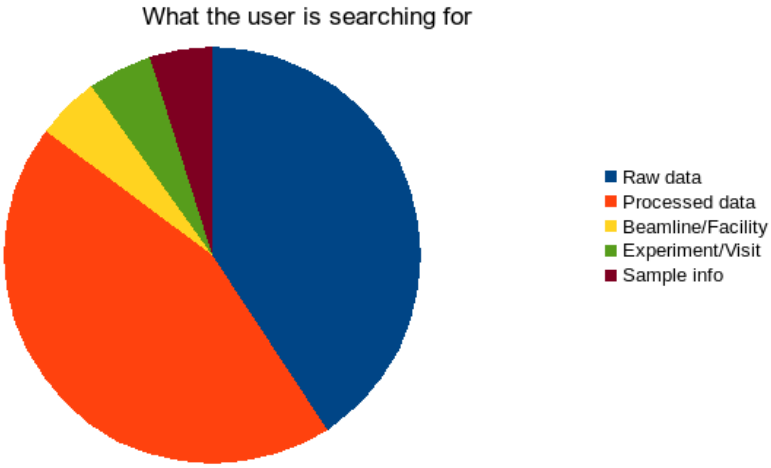


Figure A2: Pie chart representing the proportion of data/metadata that users are looking for in the survey responses.

3. *How*

This question in the survey was quite open in order to understand patterns in scientist searches. We have divided the responses in eight main categories: administrative, sample/state, physical (T, P …), measurement type, beamline setup, material use/context/project, publication/PDB, data size/resolution. Table A3 summarizes the metadata parameters researchers would use when performing a search in the data catalogue. The number of responses will consider the number of occurrences of each metadata item in the survey.

How to search for data (Category)	**Typical responses**	**Number of responses**
Admin	Investigator/PI name, Affiliation, Proposal title, Experiment ID, Date of the experiment, Instrument name, Grant number Time of acquisition Scan/run reference number	42
Sample/State	Sample name/protein name Sample composition Sample Magnetic State Battery Charge State Mineral name	46

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

	Sample description Structure information Electronic information Feature Type (for example single peak) **MX:** protein sequence Protein class Enzyme type Atomic coordinates of a protein **Crystallography (including MX):** Space Group Chemical Formula Unit cell parameters PlotType Sample Rocking Curve Width Reciprocal space (hkl) **Spectroscopy:** Edge and sample compound name Emission line/Absorption Edge Oxidation State	
Physical (T, P, …)	Sample Temperature Magnetic field applied Energy **FEL:** Bunch pattern Beam intensity Electron bunch charge and energy	19
Measurement Type	Experimental technique	18
Beamline setup	Scan command (can contain detector and motor used) Calibrant File Gas Dosing Status Scan duration/length Sample position (for example Top) Detector Type **FEL:** Operation Mode Detector intensity Focusing lenses used Monochromator	12
Material use/context/project	Project Name Material/Sample Type name	12

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Publication/Community standard databases (PDB …)		4
Data sizes/resolution		1

Table A3: Number of responses per metadata category that users used during a search

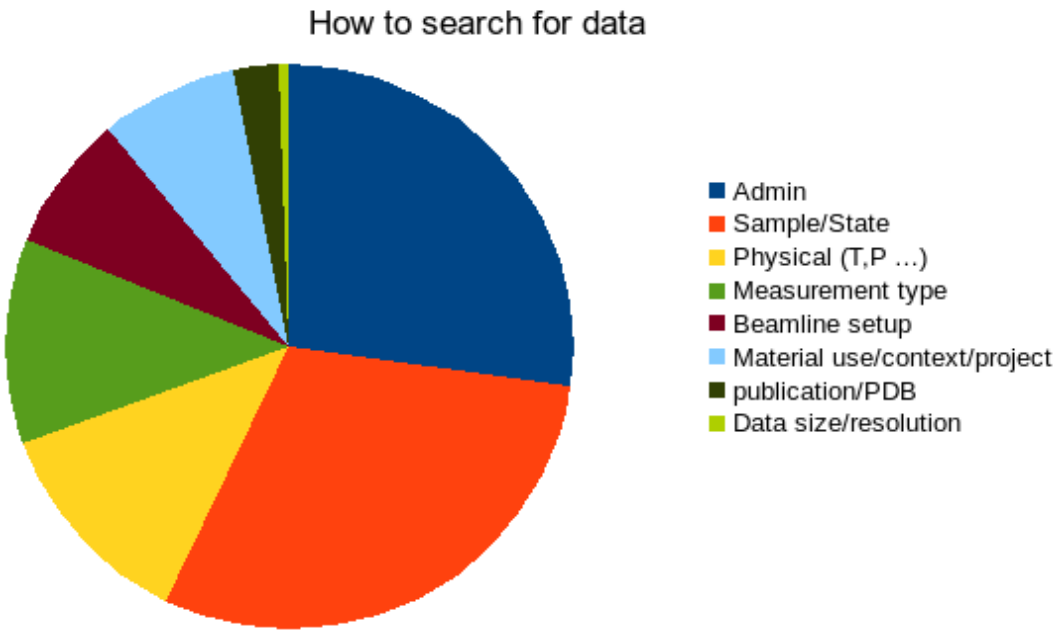


Figure A3: Pie chart representing the proportion of metadata type that users are looking for in the survey responses.

Figure A3 shows that scientists will be mainly using administrative and sample metadata when searching for data. In the comments, we can understand that their search criteria are based on their experience with journal publication and/or scientific community relevant databases based on samples such as Protein Data Bank (PDB) or Inorganic Crystal Structure Database. The MX community has suggested recording the PDB access code within the data catalogue and upload raw data from the facility during the deposition of a new structure (with associated DOI).

The beamline setup and physical parameters metadata can be obtained automatically from the control and data acquisition system during an experiment. However, it represents only 20% of the responses. Many sample metadata must be manually introduced by a user: sample name, sample type, material name, space group, chemical formula …

Surprisingly, the experimental technique was not explicitly referred to as a search criterion. However, looking closely at the responses we can observe that the sample properties/parameters used for search are related to specific techniques. Furthermore, most of the respondents are beamline scientists within their respective facility, generally local data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

catalogue will give them access to their own beamline/instrument and not all data collected by the facility. Consequently, they have limited need to use experimental techniques for search.

Looking simultaneously at the “what” and “how” responses, researchers will be using experimental parameters such as Sample Temperature for example to access processed data. It is suggested that ideally the data catalogue should provide a link between raw and processed data. The Nexus processed data files can provide external file links (virtual datasets) to raw data as well as calibration files.

For synchrotron facilities, the scientific communities more willing to fill the survey were: Crystallography/Diffraction (including MX) and Spectroscopy. It was interesting to see very similar requirements for a specific scientific community. In particular, the Spectroscopy group shows the necessity to link experimental data with standard spectra. The Nexus format provides a good set of sample metadata for Crystallography while there will be a need to add specific sample information for Spectroscopy.

In this survey, we had only three responses from the Neutron facility. They mainly use administrative metadata for their search. This can have at least two possible interpretations:

- the data rate and volume are smaller than for synchrotrons, consequently they may not need any refined search based for example on sample properties.
- Users are already using the data catalogue for many years using those parameters.

However, it is difficult to draw more conclusions due to the limited number of responses.

Finally, the two external users responses were particularly interesting. One of them mentioned that the corresponding datasets are too large to download, and suggested adding a thumbnail to help choose the data. Furthermore, enabling a more refined search within the dataset might help finding data.

4. *Why*

Like the previous question “How”, this one was open as well. We have divided the responses in eight main categories: Compare with published data, Data to process, Search admin and sample data, find publications/data published by other facilities, software development, machine learning, find best data and map phase diagram. Table A4 summarizes the reasons why a scientist will be searching for data. The number of responses will consider the number of occurrences of each metadata item in the survey.

Why users are searching for data (Categories)	**Number of responses**
Compare with published data	20

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857641.

Data to process	18
Search admin and sample data	10
Find publications/ data published by other facilities	9
Software development	5
Machine learning	4
Find best data	2
Map phase diagram	1

Table A4: Number of responses per reason why a researcher will be searching for data.

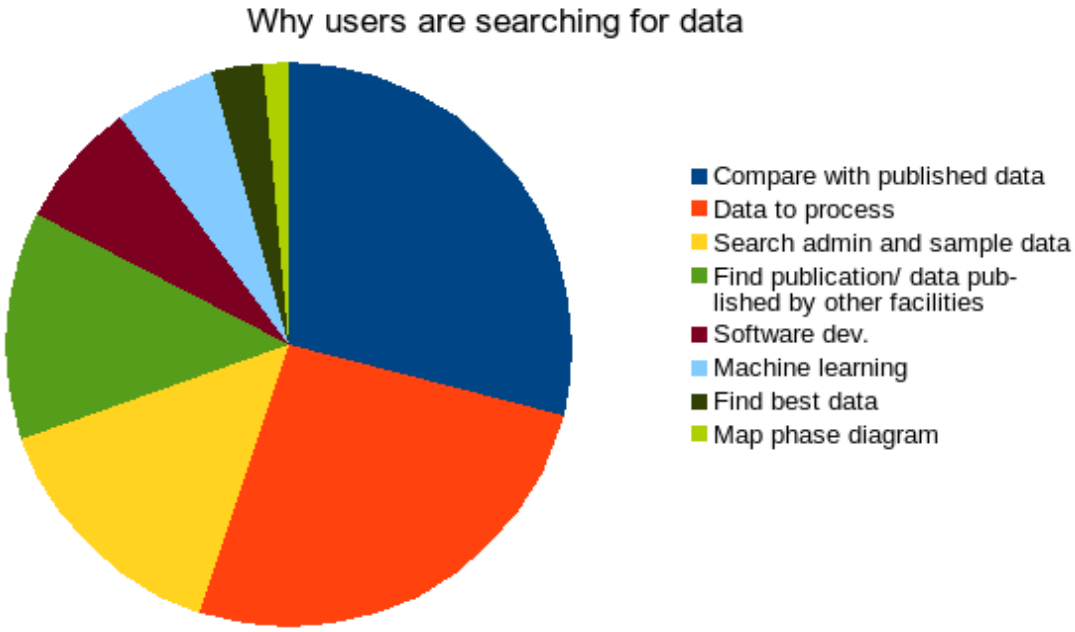


Figure A4: Pie chart representing the number of responses per reason a researcher will be searching for data.

The survey shows that scientists will be searching for data for the following main reasons: compare collected data with published data, process data, search for admin and sample data and find journals or data published by other facilities. It is clearly in the scope of the EOSC infrastructure i.e providing data analysis as a service and a data repository where researchers can find public data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

5. *When and Where*

The two questions where and when were not part of the survey as it is more related to implementation aspects. The “when” aspect will consider the research lifecycle from proposal to journal publication and maybe beyond if users keep reanalyzing their data and uploading it to the data catalogue. It will be difficult to extract information when a user leaves the facility. We consider for task 3.2 that the metadata is essentially provided by the information provided by the User Office or by the Nexus files written by the control/data acquisition or the automated processing pipelines during the user session at the facility. More details are presented in the deliverable D2.2 the Recommendations for FAIR Photon and Neutron Data Management [3] as part of task 2.3.

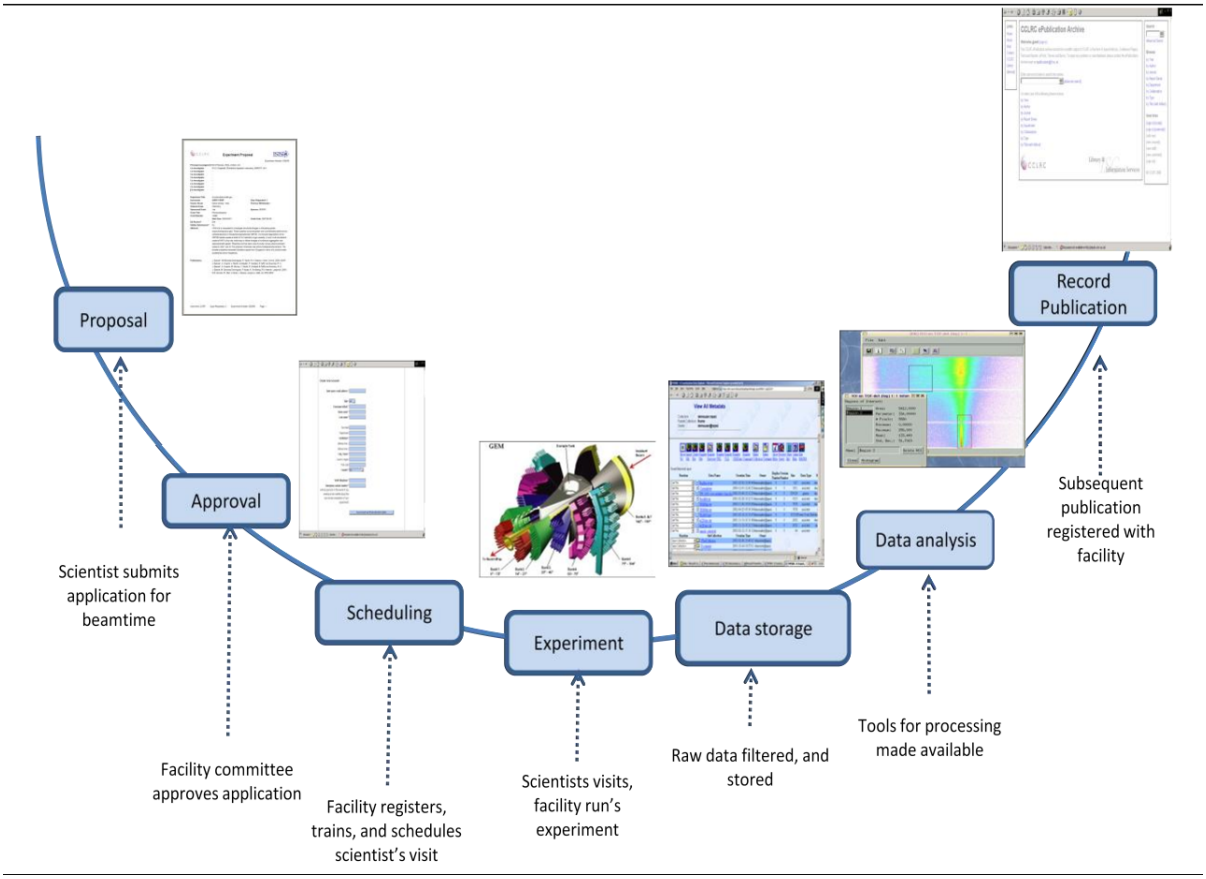


Figure A5: Research lifecycle

Looking at the responses related to “How users will be searching for data”, we try to summarize in Table A5 when each metadata type could be potentially collected within the facility research lifecycle. Here each facility may have their own databases, but we consider that the User Office is providing information before an experiment starts. Information about Physical parameters, measurement technique and beamline setup will be more likely related to raw Nexus files collected during an experimental session. The Sample information/properties is more challenging as it involves the overall facility research lifecycle from proposal to publication.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

How to search for data (Category)	When
Admin	User Office: Investigator/PI name, Affiliation, Proposal title, Experiment ID, Date of the experiment, Instrument name, Grant number During the session/experiment: Time of acquisition Scan/run reference number
Sample/State	All phases of the research lifecycle: from proposal, sample preparation, raw and processed data and interpretation of the data when users leave site.
Physical (T, P, …)	During the session/experiment: Raw Nexus file.
Measurement type	During the session/experiment: Raw Nexus file
Beamline setup	During the session/experiment: Raw Nexus file
Material use/context/project	User Office: for Material name and project name For the Experiment context: User Office (proposal, experiment report) or elog book for context within the session.
publication/Community standard databases (PDB …)	After the experiment: link with DOIs
Data sizes/resolution	During the session experiment: Raw Nexus file?

Table A5: Link between metadata type used in searches and the facility research lifecycle

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.