An Evaluation of Metadata Tools and Methods to Assess the Impact and Value of Provenance Management in the PaNData community

PaN-data ODI

Deliverable D6.4

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	An Evaluation of metadata tools and methods to assess the impact and value of provenance management in the PaNData community
Deliverable Number	D6.4
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31 August 2014 (Month 35)
Actual Delivery Date	20 October 2014

The PaN-data ODI project is partly funded by the European Commission under the 7th Framework Programme, Information Society Technologies, Research Infra-structures.

Abstract

The aggressive update programme and continuous introduction of new generation of detectors and cameras across the Photon and Neutron (PaN) community have resulted in a rapidly rising awareness of data management problems across the data continuum of the large facilities. Today, facilities, synchrotrons and increasingly neutron facilities, are witnessing the arrival of big data experiments, that is, experiments that collect 10s of TBs of data within a few days of beam time. This has not only putting strain on the frontline "conventional" data services, e.g. data acquisition, storage and archiving, but also motivate the facilities to improve its data management and support services across the entire data continuum, from proposals, experiment, data analysis to (data and paper) publications. This is to ensure that the description of the data, in the form of metadata, can be better captured in the first instance. And subsequently, they can be better exploited, during and immediately after the data are collected, but also over time when the data is analysed by the people who conduct the experiment and by others who need to exploit the data in other context, for example, for validation, secondary analysis, or meta-analysis. To do that, it is of interest to the facilities, the scientists, and the science communities in general, to keep track of the data, i.e. experiment, analysed, and resultant data, across the data continuum. That is the problem of provenance management for scientific data in the facility science domain.

The aim of this deliverable is to assess the impact and value of provenance management in the PaN facilities in the PaNData-ODI project. It analyses the responses gathered from 11 *operational* large facilities in Europe regarding metadata capturing, storage, usage, and standardisation across the data continuum. Our survey has shown that there is still a long way to go in standarising metadata and metadata management across the data continuum within one facility and across different facilities. However, there are emerging efforts which aim at bridging the gap, including PaNKOS - an ontology for facility science, and ongoing work that leverages standardised metadata publishing and access protocol OAI-PMH to share metadata about experiment data. In particular, we show how PaNKOS can be used to encapsulate contextual metadata for experiment data and how it can be used to tag proposals to enhance the understanding of the science conducted in a facility.

Keyword list

Data provenance, tools for data provenance, data continuum, ontology for data provenance, research lifecycle, Linked Open Data, data publication, data sharing

Document approval

Approved for submission to EC by all partners on 20 October 2014.

Revision history

Issue	Author(s)	Date	Description
0.1	Erica Yang, Holly Zhen (STFC)	12 Aug 2014	First Draft
0.2	Holly Zhen (STFC)	28th Aug 2014	First draft PaNKOS
0.3	Erica Yang (STFC), Thorsten Kracht (DESY)	11th September 2014	3rd Draft with survey
0.4	Erica Yang, Sylvain Letreguilly (ILL), Jean-Franscois Perrin (ILL)	15 September 2014	Rewrite PaNKOS, survey, and tagging proposal with PaNKOS
0.4	Brian Matthews (STFC)	18th September 2014	Reformatted
0.5	Erica Yang (STFC)	3 October 2014	Revised survey
0.6	Vasily Bunakov (STFC), Alistair Mills (STFC)	3 October 2014	Added OAI-PMH and EUDAT
0.7	Brian Matthews (STFCC), Catherine Jones (STFC), Antony Wilson (STFC)	10 October 2014	Added Data Journal section.
0.8	Brian Matthews (STFC)	10 October 2014	Tidied up and added conclusion.
0.9	Erica Yang (STFC)	14 October 2014	Introduction, added institutions' comments to the survey analysis and final for the consortium

Table of contents

Page

1 INTRODUCTION 5

2 A SURVEY OF THE ROLE OF METADATA IN PANDATA COMMUNITY 6

2.1 DATA STORAGE 8
2.2 METADATA CAPTURE 9
2.3 METADATA STORAGE MECHANISM 10
2.4 METADATA STANDARDISATION 11
2.5 METADATA USAGE IN SERVICES 11
2.6 AN OPTIONAL QUESTION 12
2.7 DISCUSSION 12

3 PANKOS: AN UPDATE ON A COMMON VOCABULARY FOR THE PANDATA COMMUNITY 13

3.1 WHAT PANKOS DESCRIBES? 13
3.2 HOW PANKOS IS ORGANISED? 14
3.3 USING PANKOS 17
3.4 ACCESSING PANKOS 18

4 TOWARDS A UNIFIED PROVENANCE MANAGEMENT APPROACH WITH PANKOS 19

4.1 ENCAPSULATING CONTEXTUAL METADATA FOR EXPERIMENT DATA 19
4.1.1 Data Journal Functionality 20
4.1.2 Adding Context 21
4.1.3 Architecture 23
4.1.4 Further development 25
4.2 TAGGING PROPOSALS WITH PANKOS 26

5 PUBLISHING PANDATA RECORDS USING OAI-PMH 28

6 A FINAL WORD ON PROVENANCE 30

APPENDIX A. A SURVEY QUESTIONNAIRE ABOUT METADATA CAPTURE, STORAGE, USAGE, AND STANDARDISATION 31

APPENDIX B. SECTIONS OF PANKOS DESCRIBING ISIS AND ILL, THE INSTRUMENTS, AND THE TECHNIQUES 32

APPENDIX C. WEBLINKS OF PANKOS 42

APPENDIX D. PANKOSREST APIS 42

APPENDIX E. OAI-PMH SERVERS CONSIDERED FOR PANDATA ENDPOINT. 45

APPENDIX F. THE MAPPING OF PANDATA ICAT FIELDS TO THE OAI-PMH SCHEMA AND CORRESPONDING FIELDS IN EUDAT METADATA SCHEMA. 47

APPENDIX G. THE HARVESTER, CONVERTER, AND PUBLISHER FROM ICAT TO QUALIFIED DUBLIN CORE IN OAI-PMH ENDPOINT. 49

1 Introduction

Like many large organisations, the photon and neutron facilities have used metadata extensively in many systems, from managing proposals, samples, experiment data, to publications. Increasingly, as more institutions develop data policies, the role of metadata is expected to become more eminent. However, there is still a great scope to extend the role of metadata for advancing scientific data management to

- Improve operational efficiency of the facilities;
- Accelerate data analysis;
- Facilitate data exploitation;
- Improve the use of metadata to increase the value of the experiment data taken at the facilities, through for example, linking up experiment data with publishers (e.g. IUCr), and community (resultant) data archives (e.g. CCDC);
- Provide scientists and facility managers better insights into the science and research taking place within a facility and across facilities; and finally
- Enable impact tracking throughout the lifespan of research lifecycle

Metadata is the founding stone underpinning provenance management, from describing proposals and data, to annotating the processes (e.g. algorithms, and software) that relate to the analysis of data, and to the publications of research results, be it conventional paper publications or publications of data. This deliverable presents the results from a study with the response from 11 operational large scale European national and international facilities regarding the *use* of metadata across the facility data continuum. This study represents the state of the practice in metadata management across the data continuum of the large facilities in Europe. Our survey has found that there is still a long way to go in standarising metadata and metadata management across the data continuum within one facility and across different facilities. One of the potential consequences, for instance, is that it creates a barrier for the users to search and link up the research data they collect from different facilities. This deliverable also describes an update of PaNKOS – Photon and Neutron Knowledge Organisation System. This is an ontology for facility sciences with the aim to standarise the vocabularies used to describe the techniques, facilities, and instruments from the facility science community. We demonstrate two use cases of PaNKOS in an attempt to unify provenance management: 1) how to encapsulate contextual metadata for experiment data to produce a data journal for a facility; and 2) how PaNKOS can be used to enhance the understanding of the science from plaintext proposals through a tagging mechanism. Finally, this deliverable presents an exploratory attempt stretching beyond the coverage of traditional data continuum of large facilities by describing a collaborative work between two large EU projects PaNData-ODI and EUDAT about publishing ICAT records using a standardised metadata publishing and accessing protocol OAI-PMH.

Now, let us dip into the survey about the role of metadata in the PaNData community.

2 A survey of the role of metadata in PaNData community

To develop a picture of the current usage of metadata across the data continuum in the European large facilities, we have conducted a survey among the PaNData-ODI partners, entitled “A survey about metadata capture, storage, usage, and standardisation”. This is the first time a survey of this nature conducted for this community. We have collected the responses from twelve large facilities in Europe. Thus, it represents the state-of-the-practice of metadata management in these large science facility operators.

This survey contains questions for probing the following aspects of the data continuum:

- Data storage: whether a facility stores experiment data, analysed data, and/or resultant data[1]
- Metadata capturing mechanisms for user information, proposal information, and experiment data
- Metadata storage mechanisms (database or file system)
- Adoption of metadata standardisation mechanisms (mandatory metadata, standardised metadata model, DOIs – Digital Object Identifiers[2])

A copy of the full survey questionnaire is presented in Appendix A. In the rest of this section, we describe the findings from this survey. The responses were collected between February and August 2014. Table 1 gives an overview of the results.

[1] Resultant data are defined as the data that are companion to the publications that are resulted from an experiment conducted at a facility.
[2] http://www.doi.org/

Table 1: An Overview of the Survey Results

#	Question	DESY	DLS	Elettra	ALBA	SOLEIL	CEA-LLB	ISIS	JCNS	ILL	HZB	PSI
1	Facilty Type	Photon	Photon	Photon	Photon	Photon	Neutron	Neutron	Neutron	Neutron	Both	Both
2	Store exp data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3	Store analysed data	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	No
4	Store resultant data	No	Yes	No	No	No	No	Yes	No	Yes	No	No
5	Capture metadata about users+proposals	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	Capture metadata about exp data	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
7	Use db to store metadata about exp data	Yes	Yes	planned	No	planned	No	Yes	Yes	Yes	Yes	Yes
8	Store exp metadata inside files	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
9	Use db to store metadata about analysed data	No	Yes	No	No	Yes	Yes	Yes	No	planned	No	No
10	Use db to store metadata about resultant data	No	Yes	No	No	No	No	Yes	No	planned	No	No
11	Use mandatory metadata	Yes	No	planned	No	No	No	Yes	planned	Yes	No	No
12	Use standard standard/file format to describe data	Yes	Yes	No	No	?	No	Yes	No	Yes	No	No
13	DOIs	planned	No	Yes	No	No	Yes	Yes	planned	Yes	No	No
14	Synchrotron metadata more diverse	-----	---	---	----	---	-----	No	---	----	Yes	No

Till 21 August 2014, twelve facilities responded to this survey. Among them, there are eleven operational facilities and one developing facility (MAX IV), which is under construction. Understandably, the answers from MAX IV are mostly 'undecided' or 'planned', except the question about storing information about users and proposals, the answer to which is positive. Among the eleven operational facilities, five are photon source, four neutron source, and two operate both neutron and photon sources.

In order to maximise the response rate, we have decided to keep the survey questions simple. It is up to the facilities to decide the scope for their answer and describe that scope in the response. When we designed the questionnaire, we felt that it is important to keep the amount of time taken to answer the questions short, hence the short questions. Similarly, we do not differentiate the 'yes' answers (to the same question) from different facilities. For instance, if a facility gives a 'yes' to question 2 – "Store experiment data 6 months after they are captured?", it makes no difference in our statistics whether the facility stores its data for *all* its instruments over 6 months or it only stores the data for *one* instrument. *So, even if there is only one instrument that stores the data for over six months, we would expect that the answer to be a 'yes'.* We have applied this rule to treat all the answers. It is also worthwhile to bear in mind what this survey does not tell you: this survey does not provide the coverage of the answers from the facilities.

The rest of this section will focus on the data from the 11 operational facilities on question 2 to 14. Question 1) is about the type of the facility, the answers to which is shown in Table 1.

2.1 Data storage

In this section, there are three questions, listed as follows.

Question 2. Store experiment data 6 months after they are captured? Yes/No/Planned

Question 2) Apart from what is presented in Table 1, a few facilities provided additional comments. ILL stores the data right after experiment and keeps them forever. But, not all neutron facilities keep all the data. For example, JNCS noted that they store all experiment data, except tomography.

SOLEIL, the French synchrotron does the same, but the exception for them is Macromolecular Crystallography (MX) data. The response, provided by PSI who operates both neutron and synchrotron facilities, indicates that they store neutron data for long term ("back to 1996 when SINQ started. Most in NeXus."), but, x-ray (synchrotron) data are deleted after 3 months (because of the cost involved). It is the users' responsibility to store and manage the x-ray data for the longer term.

The comment from HZB is that "Most captured neutron data is being kept long-term, photon data can be stored 6 month or longer." ISIS stores 25+ years of neutron (experiment) data and has made a commitment to store them for the long term[3].

[3] See ISIS data policy here for details of the meaning of the word 'long term': http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

Question 3. Store analysed data 6 months after they are captured? Yes/No/Planned

Question 3) 45% of the facilities answers 'Yes' to this question. Four out of the five synchrotrons indicate that they provide storage facilities for analysed data longer than 6 months. Only one out of the four neutron facilities say that it stores analysed data 6 months after they are captured. For ISIS, analysed data storage was offered "informally, but they started to offer it formally via their data analysis framework." The percentage of synchrotrons who store analysed data is a lot higher than that of neutron sources. This is unexpected because, generally, for the same technique, synchrotron facilities produce a lot more data than neutron facilities. One of the conjectures is that because synchrotron data is a lot more, synchrotron users tend to use the storage and compute infrastructure provided by the facilities more during an experiment. Hence, more synchrotron facilities are able to capture the analysed data.

ISIS commented that "for question 3), (analysed data are stored) informally and now (Mantid, 1 week ago) formally."

HZB noted that "There is no "coordinated" storage of analysed data. This is up to the scientists."

Question 4. Store supplementary data for papers? Yes/No/Planned

Question 4) 27% of the replies is 'Yes' to this question. Specifically, only one out of the five synchrotrons indicates that it stores supplementary data for papers. In contrast, two out of the four neutron facilities replied positively. The two institutions that operate both sources reply negatively to this question. To the best of our knowledge, none of the eleven facilities provides a mechanism to allow users to link up a publication to a dataset. Therefore, we conclude that a positive reply to this question is more likely an indication that the replying institutions capture the publications related to their experiments, but it is not necessarily an indication that the publications and datasets (or experiments) are linked.

The two institutions that operate both types of facilities indicate that the responsibility of managing analysed and resultant data is up to the scientists/users. SOLEIL also made the same note on their answers to questions 3) and 4).

ISIS commented that "(resultant data are stored for papers,) informally, but not yet made available for citation. See above answer (i.e. answer to 3)."

A note from HZB said that "At least not in direct connection to the data-sets. Information on publications is linked with the beamtimes, but there is no stringent link between data-sets and beam-times."

2.2 Metadata capture

The next category of questions relate to metadata capturing practice in the facilities. The questions are:

Question 5. Capture metadata about users and proposals? Yes/No/Planned

Question 6. Capture metadata about experiment data? Yes/No/Planned

Question 5) Without surprise, all institutions say they capture metadata about users and proposals. All replied positively to **question 6**.

In its response to 5), HZB noted that "Such data is being stored but not directly linked to the data sets." For 6), HZB commented that "Some experiments do in their data sets, some only in the lab-book."

For 6), PSI commented that "for neutrons, not sure about x-ray".

2.3 Metadata storage mechanism

This section has four questions:

Question 7. Use databases to store information about experiment data? Yes/No/Planned
Question 8. Store experiment metadata inside data files (file format is _irrelevant_ in this context)? Yes/No/Planned
Question 9. Use databases to store information about analysed data? Yes/No/Planned
Question 10. Use databases to store information about supplementary data about papers? Yes/No/Planned

Question 7) Eight out of the eleven facilities use databases to store information about experiment data. Three are in a planning stage. For a detailed state-of-the-practice use of database (in particular, ICAT[4]) for cataloguing experiment data, please read Section 2.4 Metadata Catalogue ICAT-4 in the PaNData-ODI deliverable D5.3, "Report on the implementation of the three vir-tual laboratories".

PSI commented that "Neutrons: Yes, being ported to ICAT; X-ray: no".

Question 8) Ten out the eleven facilities also store metadata inside files.

PSI commented that "Neutrons: Yes, X-ray: no".

Question 9) Two out of the five synchrotrons use databases to store information about analysed data. Two out of four neutron sources do the same although one has just started (though it is unclear about the up-take).

ISIS made an comment that "Mantid now supports this via ICAT (1 week ago!). There is as yet no take up."

Question 10) Only one synchrotron and one neutron source indicate that they use databases to store information about supplementary data of papers.

The answer from PSI and HZB to questions 9) and 10) is negative.

---

[4] http://icatproject.org/

2.4 Metadata standardisation

Question 11. Use a set of mandatory metadata to describe the data gathered at your facility, regardless of which beamline/instrument they are collected? Yes/No/Planned
Question 12. Use metadata standard to describe data gathered at your facility? Yes/No/Planned

Question 11) One synchrotron indicates that they use mandatory metadata to describe their data for at least one beamline. Two neutron sources indicate that they use Nexus file format. However, it should be pointed out that using the Nexus file format doesn't necessarily mean that a facility adopts a set of mandatory metadata to describe experiment datasets from that facility. The Nexus file format describes how data (e.g. images) and information (e.g. metadata like camera parameters for images) is organised inside a Nexus file and but, Nexus is a flexible file format in that it is able to accommodate virtually any metadata.

Both PSI and HZB reply negatively to this question. Particularly, PSI made a comment that "Neutrons: there are a certain number of fields which are in all files. But there is no officially defined mandatory set. X-rays: no organised storage".

For Diamond Light Source, it should be noted that although they don't claim that they apply mandatory metadata for their datasets. But, "the user office provides the who, what, which beamline, when automatically". These metadata get into the archive with the datasets during the data cataloguing and archiving process and in most cases, carry through to the data files produced by the beamlines.

SOLEIL noted that "Some fields are in all files, but we can't say that there is officially a mandatory set".

Question 12) Two out of the four synchrotron facilities reply positively to this question. Only two neutron sources explicitly indicate that they use Nexus file format and employ metadata from DataCite (for DOI landing pages). Both PSI and HZB reply negatively to this question.

In response to questions 11) and 12), ILL noted that "standardisation came from Nexus and Datacite DOI metadata".

ISIS commented that "We use Nexus and ICAT."

2.5 Metadata usage in services

Question 13. Use DOIs in your facility? Yes/No/Planned

Question 13) Only one synchrotron uses DOIs; whilst three out of the four neutron sources use them. Neither of the remaining two institutions employs DOIs to describe their data. Altogether, 36% of the facilities use DOIs for their data. Interestingly, although one facility indicates that they store supplementary data for papers (question 3), but they don't use DOIs for the data. On the other hand, there are two facilities that use DOIs, but they don't store supplementary data. These suggest that DOIs are primarily used for experiment data, rather than analysed or supplementary data.

ISIS made a note indicating that "Issue them (DOIs), typically at the experiment level."

2.6 An Optional Question

Question 14. In general, synchrotron facilities collect more diverse range of metadata than neutron facilities. Is it true? Yes/No

Question 14) this is an optional question. Only three intuitions reply. One said 'Yes' and two said 'No'. It is interesting that the two institutions who operate synchrotron and neutron sources gave an opposite answer: one 'Yes' and one 'No'. The other comments are: "Neutron data is as diverse as x-ray data. The x-rays just have a harder job: because of data rates, because of commercial equipment with each company having her own 'standard', higher turnaround of users." And "Not particularly but synchrotrons store much higher volumes of files and hence repeated metadata".

ISIS made a note indicating their view on this question – "Facilities vary more than synchrotron to neutron."

2.7 Discussion

First of all, in large facilities, the support for analysed and supplementary data reduces significantly when users leave a facility: the number of facilities supporting experiment, analysed, and resultant data drops from 11, to 5, then to 3, respectively. However, the number of facilities that has adopted DOIs is 4, with 2 are planning to adopt them. This suggests that the facilities are interested in the tracking facility outputs, although they have much less resources to manage the other types of data down the data continuum.

On the other hand, the answers to question 7) and 8) indicate that the facilities are transitioning from file-based metadata storage to a more organised and managed form of metadata storage using database technologies. Some facilities go one step further to systematically collect and manage metadata for analysed and supplementary data. This is encouraging news for the users, in particular, the big experiment data users, e.g. imaging and MX users, who are facing major difficulty of analysing such data. However, the fact that there are already 4 out of 11 facilities (half photon and half neutron, the answers to question 9) who are managing analysed data for users is a probably an indication that users increasingly depend on the compute infrastructure provided by the facilities.

However, when it gets to supplementary data, that is, the data used in a publication, only two facilities responded that they manage such data. This indicates that it remains to be a challenge to gather such data from users, despite that 6 facilities have already adopted (or planned) to use DOIs for their experiment data.

The answers to questions 11) and 12) indicate that there is a lack of metadata standardisation across the facilities. Less than half of the facilities say that they adopt some form of metadata standards (27% for mandatory metadata or 36% for file format) within their institution. No respondents indicate whether their mandatory metadata is consistent with those from other institutions.

Given that the adoption rate within one institution is low, standardising the metadata for experiment data across institutions is expected to be even more challenging. For the photon

and neutron user community, this may be a worrying message. Because, even with the same sample with the same technique on the same type of instruments, the data from different facilities could be described differently using different types or names of the metadata, and/or different file format. This seems to suggest that some degree of data handling (e.g. converting from one image format to another), is still needed by facility users.

3 PaNKOS: An update on a common vocabulary for the PaNData community

3.1 What PaNKOS describes?

PaNKOS stands for *Photon and Neutron Knowledge Organisation System*. It is an ontology describing three main categories of things related to facility sciences and their relationships. These three categories are **Facility**, **Instrument** and **Technique**.

PaNKOS is written in OWL[5] (The Web Ontology Language) using an open source tool called Protégé. It describes various neutron and synchrotron facilities from all over Europe, with information regarding their instruments and techniques used. It could be used for tagging and annotating units of information, for example it could be incorporated inside ICAT which is a data cataloguing tool developed by STFC.

In OWL, everything is a subclass of *Thing* which is the superclass. *Facility*, *Instrument* and *Technique* are classes themselves and they are all subclasses of *Thing*. A class is an abstraction of things in the real world. A Turtle document is a textual representation of an RDF graph. The class representations in both OWL and Turtle format are as followed:

OWL	Turtle

`<SubClassOf>`
`<Class IRI="#Facility"/>`
`<Class abbreviatedIRI="owl:Thing"/>`
`</SubClassOf>`
	`PaNKOS:Facility rdf:type owl:Class ;`
	`rdfs:subClassOf owl:Thing ;`
	`owl:disjointWith PaNKOS:Technique , PaNKOS:Instrument .`

`<SubClassOf>`
`<Class IRI="#Instrument"/>`
`<Class abbreviatedIRI="owl:Thing"/>`
`</SubClassOf>`
	`PaNKOS:Instrument rdf:type owl:Class ;`
	`rdfs:subClassOf owl:Thing ;`
	`owl:disjointWith PaNKOS: Technique ,`
	`PaNKOS:Facility .`

`<SubClassOf>`
`<Class IRI="#Technique"/>`
`<Class abbreviatedIRI="owl:Thing"/>`
`</SubClassOf>`
	`PaNKOS:Technique rdf:type owl:Class ;`
	`rdfs:subClassOf owl:Thing ;`
	`owl:disjointWith PaNKOS:Technique ,`

[5] http://en.wikipedia.org/wiki/Web_Ontology_Language

	PaNKOS:Instrument .

Here is a simple example. PaNKOS describes the following things (but not limited to): ISIS, ENGIN-X, and Diffraction.

- ISIS is a neutron **Facility**
- ENGIN-X is an **Instrument**
- Neutron Diffraction is a **Technique**

For these three things, they have the following relationships:

- ISIS has an instrument called ENGIN-X.
- ENGIN-X supports a technique called NeutronDiffraction.

In the turtle format, we describe these relationships as follows:

```
PaNKOS:ISIS rdf:type  PaNKOS:NeutronSource ;
            PaNKOS:preferredName  "ISIS"^^rdfs:Literal ;
            PaNKOS:hasInstrument  PaNKOS:ENGIN-X .
```

```
PaNKOS:ENGIN-X rdf:type PaNKOS:NeutronDiffractometer ;
               PaNKOS:preferredName "ENGIN-X"^^rdfs:Literal ;
               PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .
```

3.2 How PaNKOS is organised?

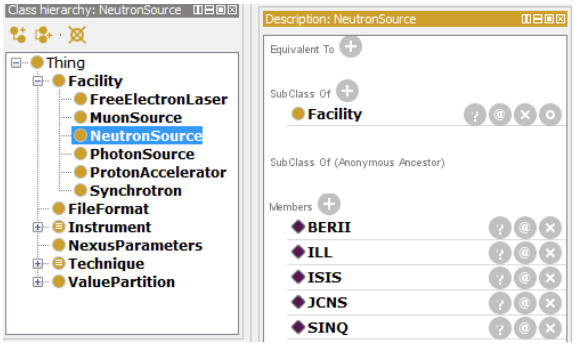
PaNKOS[6] ontology mainly consists of Individuals, Properties and OWL Classes which can then be divided into two different components, namely the Abox and Tbox component. A TBox is a "terminological component" which is a conceptualization associated with a set of facts, known as an ABox. An Abox is an “assertion component” which is a fact associated with a terminological vocabulary within a knowledge base.

Individuals represent objects in the domain which are interested. Individuals are also known as *instances.* They can be referred to as being ‘instances of classes’.

Examples:

The Class NeutronSource includes BERII, ILL, ISIS, JCNS and SINQ which are the individuals, as below:

[6] See Appendix B for selective sections of PaNKOS for the ISIS and ILL information.



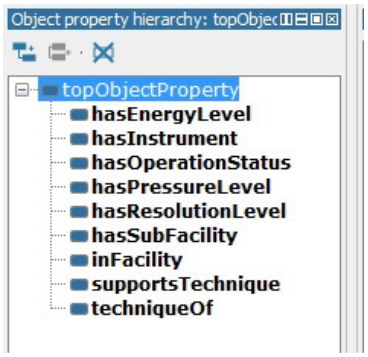
They are represented in turtle as followed:

```
:BERII rdf:type :NeutronSource .

:ILL rdf:type :NeutronSource .

:ISIS rdf:type :NeutronSource .

:JCNS rdf:type :NeutronSource .

:SINQ rdf:type :NeutronSource .
```

OWL classes are interpreted as sets that contain individuals. They are described using formal description that states precisely the requirements for memberships of the class.

Examples

OWL properties represent relationships. There are two main types of properties: Object and annotation properties. **Object properties** are relationship between two individuals.



Example of an ISIS instrument SXD

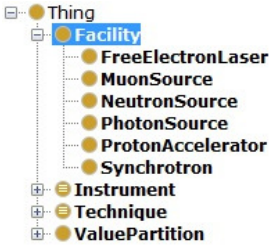
```
:SXD rdf:type :NeutronDiffractometer ,
              :SingleCrystalDiffractometer ;

     :preferredName "SXD"^^rdfs:Literal ;

     :inFacility :ISIS ;
```

```
      :supportsTechnique   :NeutronDiffraction
;

      :hasOperationStatus :Operational ;

      :supportsTechnique
:SingleCrystalDiffraction
```

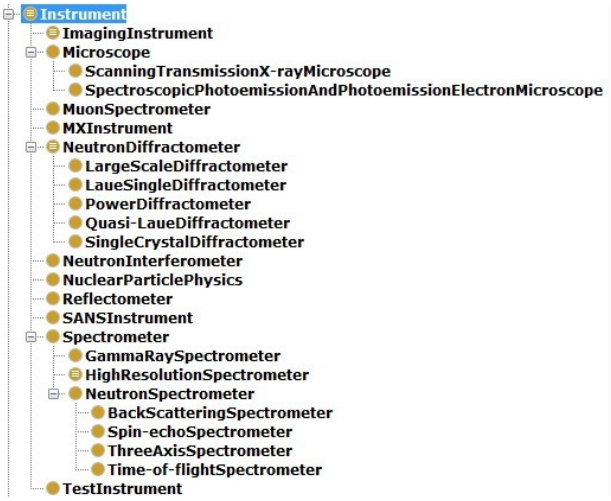
Annotation properties can be used to add metadata to classes. A new annotation property *preferredName* has been added to the ontology to specify a preferred term. It is represented in OWL as

```
<AnnotationProperty rdf:about="&PaNKOS;preferredName"/>
```

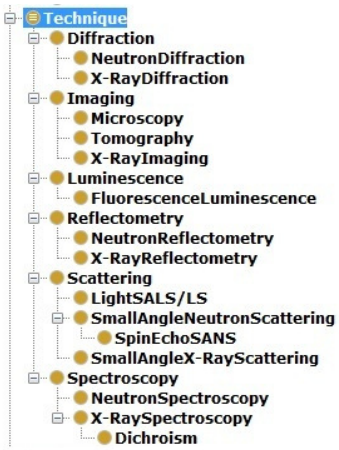
Facilities are organised by facility type, whether it is a free electron laser (definition), Muon Source, Neutron Source, Photon Source, Photon Accelerator and Synchrotron.



Instruments are organised by the types of techniques used. Instruments use diffraction are types of diffractometer, and instruments use spectroscopy are types of spectrometer and so on.



The **Technique** class contains a list of experimental techniques that instruments commonly use. There are currently six top level techniques including Diffraction, Imaging, Luminescence, Reflectometry, Scattering and Spectroscopy. Each of these top level techniques can be further expanded to include more sub levels as shown as below.



3.3 Using PaNKOS

This section briefly outlines two scenarios that PaNKOS can be used in real world situations.

First of all, PaNKOS can be used as a set of *controlled vocabularies*, i.e. a *standardised* set of predefined keywords to describe things. For example, the experimental techniques captured by PaNKOS, once they are standardised, can be used to describe the techniques an instrument employs in a facility. If and when all the facilities adopt the same vocabulary to describe the experimental techniques they offer, it will be beneficiary for the users to comprehend the range of science capabilities that one facility supports and for the facility science community as a whole to compare and assess the overall provisioning of science capabilities across the facilities. In terms of applications, controlled vocabularies can be used for document (or unit of information) tagging/annotation, indexing, and search or information retrieval.

Secondly, PaNKOS can be used as taxonomy, i.e. a classification of instruments, techniques, and facilities. In PaNKOS, all these things are organised in a hierarchal structure. For example, a **Technique** consists of one or several **SubTechnique**s. A type of instruments, i.e. an **InstrumentType** can consist of one or more **SubInstrumentType** and so on. In library and information science, taxonomies are used to classify and annotate subject headings.

Thirdly, but not lastly, as a RDF graph, PaNKOS can be used to establish further relationships between PaNKOS things through inference techniques by exploiting the transverse and inverse property relationships defined in PaNKOS. For example, the very simple RDF triples describing the instruments (LOQ from ISIS, and IN6 from ILL):

```
PaNKOS:LOQ rdf:type PaNKOS:SANSInstrument ;
           PaNKOS:preferredName "LOQ"^^rdfs:Literal ;
           PaNKOS:supportsTechnique PaNKOS:SmallAngleNeutronScattering .

PaNKOS:IN6 rdf:type PaNKOS:Time-of-flightSpectrometer ;
    PaNKOS:preferredName
   "Cold neutron time-focussing time-of-flight spectrometer "^^xsd:string ;
  PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .
```

Along with the triples describing other instruments and techniques in PaNKOS, can be used to infer further relationships between the PaNKOS “things”. In the future, this could lead to a very powerful way of exploring and exploiting the proposals, data, and publications across the facilities (assuming the content is publically available).

```
PaNKOS:SmallAngleNeutronScattering
        a   owl:Class ;
        rdfs:subClassOf PaNKOS:Scattering ,
                        PaNKOS:Technique,
                        PaNKOS:SmallAngleScattering ;
        PaNKOS:hasSubClass PaNKOS:VerySmallAngleNeutronScattering,
                           PaNKOS:VSANS ,
                           PaNKOS:USANS ,
                           PaNKOS:UltraSmallAngleNeutronScattering ,
                           PaNKOS:SpinEchoSANS ,
                           PaNKOS:Time-of-FlightSANS ,
                           PaNKOS:SpinEchoSmallAngleNeutronScattering ;
        PaNKOS:hasSubTechnique
                          PaNKOS:SpinEchoSmallAngleNeutronScattering ,
                          PaNKOS:VerySmallAngleNeutronScattering ,
                          PaNKOS:Time-of-FlightSANS ,
                          PaNKOS:USANS ,
                          PaNKOS:VSANS ,
                          PaNKOS:SpinEchoSANS ,
                          PaNKOS:UltraSmallAngleNeutronScattering ;
        PaNKOS:preferredName
               "Small Angle Neutron Scattering"^^rdfs:Literal ;
        PaNKOS:subTechniqueOf  PaNKOS:SmallAngleScattering ;
        PaNKOS:techniqueOf
                  PaNKOS:LOQ , PaNKOS:IN6 , PaNKOS:IN15 , PaNKOS:FIGARO ,
                  PaNKOS:D22 , PaNKOS:D11 , PaNKOS:IN11 , PaNKOS:S18 ,
                  PaNKOS:D17 , PaNKOS:SuperADAM , PaNKOS:IN5 ,
                  PaNKOS:SANS2D , PaNKOS:ZOOM , PaNKOS:D33 ,
                  PaNKOS:IN4C , PaNKOS:D7.
```

3.4 Accessing PaNKOS

PaNKOS is written in OWL, expressed as RDF triples. A popular way of storing these triples is through RDF triplestores. We have chosen to use a popular, free, and open source Java based framework, called Jena, to store the triples and to perform inference over the

PaNKOS. The query language for retrieving information from the store is called SPARQL, equivalent to SQL for databases.

To retrieve information from the triplestore, two access mechanisms are available:

- *Fuseki*[7]: a SPARQL end-point accessible over HTTP, this is an interactive web application allowing users, who are able to use SPARQL, to directly interact with PaNKOS data. This is a full access to all the data. An instance of the Fuseki server serving PaNKOS data is offered at: http://PaNData.org:8009/.
- *PaNKOSREST APIs*: a set of REST style APIs allowing users and applications to get json representations from PaNKOS data with a predefined set of SPARQL queries. A restricted set of PaNKOS data is offered via this method. Behind the scene, the APIs build upon the Fuseki server to deliver the json messages. This is implemented as a web application based on the Java Jersey REST framework[8]. It can be deployed and tested in Apache Tomcat 7 at RAL. According to the Jersey website, it should also work on Glassfish and other servlet containers. See Appendix D for more details about these APIs.

4 Towards a unified provenance management approach with PaNKOS

In this section we describe two demonstration prototypes of using PaNKOS and provenance management to provide enhanced data discovery, exploration and analytic methods.

4.1 Encapsulating contextual metadata for experiment data

In deliverables D6.2 we considered how to extend the metadata model which is used within the ICAT experimental catalogue in order to represent provenance information. In D6.3 we discussed how this can be used in conjunction with other contextual information to provide a complete picture of the research arising from a facilities experiment as an *Investigation Research Object* (IRO). In this section, we describe how this notion has been instantiated and used to develop the notion of a Data Journal, providing a record of a facilities research[9].

The concept of a "Data Journal" is analogous to a Technical Report Series as it contains records of experiments and their associated data, formally published/provided by a facility; for the prototype, ISIS was used as the subject facility. The data journal could provide the production service to provide the experiment's DataCite DOI landing page. When developing the Data Journal, the current ISIS DOI landing pages for investigations were used as a starting point, see https://data.isis.stfc.ac.uk/doi/INVESTIGATION/24089729/. The basis of the IRO is the information which is already present in ICAT, the ISIS data management sys-

[7] Part of the Jena suite of RDF tools https://jena.apache.org/index.html
[8] https://jersey.java.net/
[9] This work has been undertaken in conjunction with the European Project SCAPE on scaleable preservation. The aim of SCAPE is the preservation of large scale and complex data and has concentrated on those aspects relating to preservation, which are omitted in this discussion.

tem. The IRO Builder harvests Investigations from the ICAT Data Management system, generates IROs, which are stored in an RDF Triple Store.

However the focus of this work is to enable the enhancement of the automatically generated IRO to explore how to preserve the complex links. Thus all that has been provided within the Data Journal functionality is a fairly basic browse capability which is based on the structure of the process of running the neutron spallation source. There are additional filters to enable the browser to limit the result set to a particular instrument, or type of data (experiment, test etc.). It is also possible to search for a given unique investigation number, known as the RB number. However for a fully functional search interface more consideration would need to be given to how end-users, who are not familiar with the ISIS processes, might want to discover the data and using some of the additional context being added, such as terms from an ontology which describes the experimental techniques, would enhance the searching experience.

4.1.1 Data Journal Functionality

Figure 1 illustrates the starting page of the Data Journal. Within the Data Journal, Investigations are grouped into cycles, where a cycle represents a time period, during which ISIS was operational, typically two to four months. The cycle data is obtained via a triple store query. A cycle can be expanded to show all of the investigations in that cycle. The cycles and investigations are presented in chronological order, based on start date of the cycle/investigation, with the most recent cycle/investigation at the top of the page.

Clicking on an investigation on the contents page brings up information about a specific investigation, as illustrated in Figure 2. Additional functionality, beyond the current production DOI landing page, has been provided through a number of tabs and buttons across the top of the page.

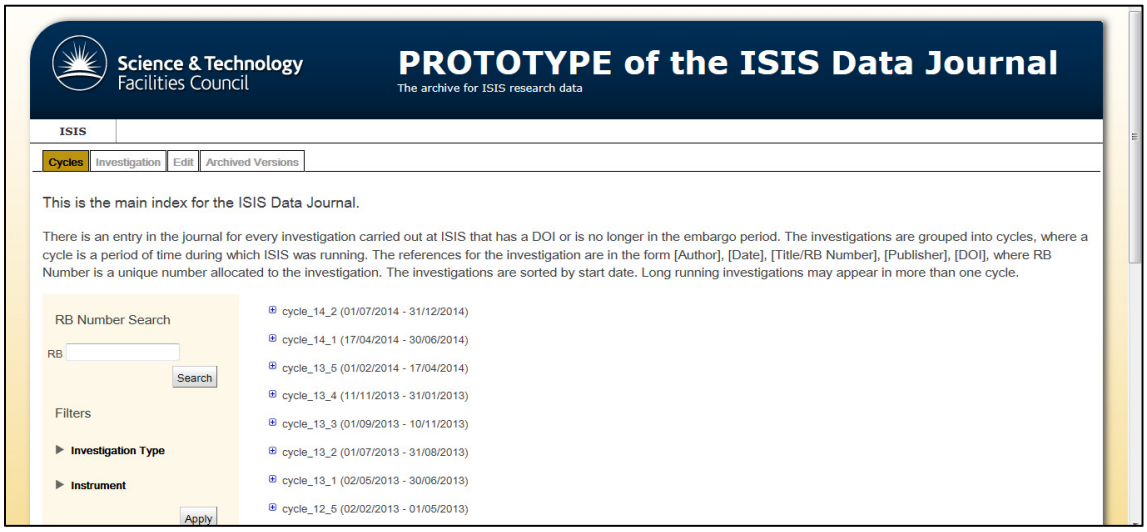


Figure 1: Starting page for the Data Journal showing cycles and browse filters

The tabs are:

- **Cycles**: returns to the view of all of the cycles. N.B. this does not reload the data so if you expand any cycles that have previously been expanded you will get the cached data (refreshing these data is a very expensive query)
- **Investigation**: this is the DOI landing page and contains all of the available information about the IRO
- **Edit**: add or remove links to resources for the current IRO
- **Archived Versions**: view and download archived versions of the current IRO
- **Download**: download the actual dataset behind the metadata (this is only available when security is enabled)

The navigational buttons are:

- **Previous Investigation** and **Next Investigation**: which scroll though investigations in order of the start time of the investigation
- **Archive**: trigger an archive action for this IRO
- **RDF**: download the RDF of the current IRO with the option to select the format

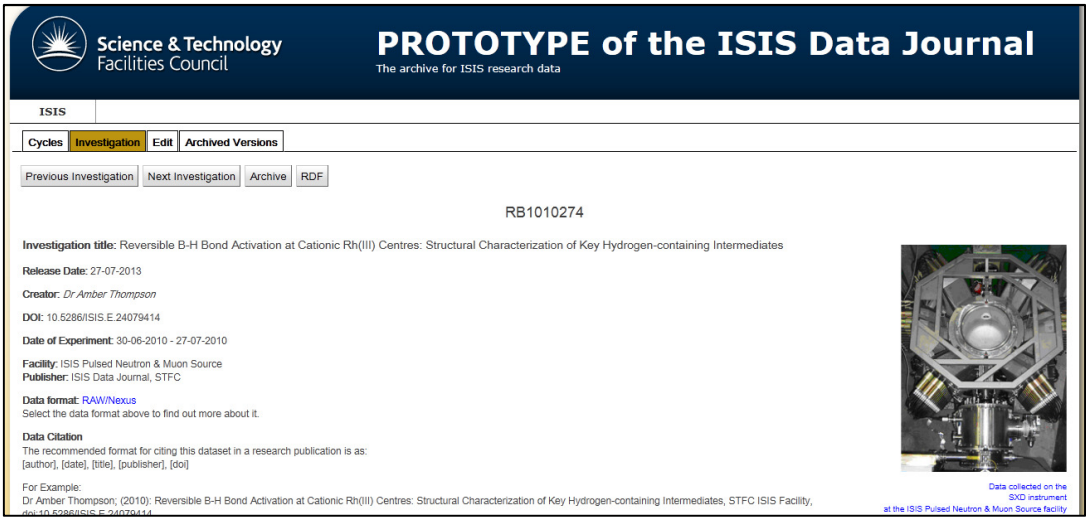


Figure 2 IRO display

4.1.2 Adding Context

The manual addition of content is achieved through the use of the Edit button within the Investigation View. There are currently two methods for adding context manually:

- Adding a subject term from the PaNKOS ontology which describes the experimental technique used to generate the experimental data. The appropriate subject terms are automatically suggested using the information on which instrument was used held within the IRO. There is a many to one relationship between techniques and instruments so it cannot be automatically assigned. In the live IRO, the link to the live PaNKOS term is what is stored.

- Creating a link to an external resource which has a URI. This uses the cito[10] ontology to describe the relationship between the IRO and the external resource and the

[10] http://purl.org/spar/cito/

fabio[11] ontology to describe the external resource. It is possible at the live IRO stage to add and remove these relationships to enable mistakes to be corrected. Examples of external resources include analysis software such as MANTID framework, which has versions identified via a DOI; publications and grants.

Figure 3 illustrates the edit page for adding this additional contextual information and relationships, while Figure 4 illustrates how it is presented to the user as part of the investigation page.

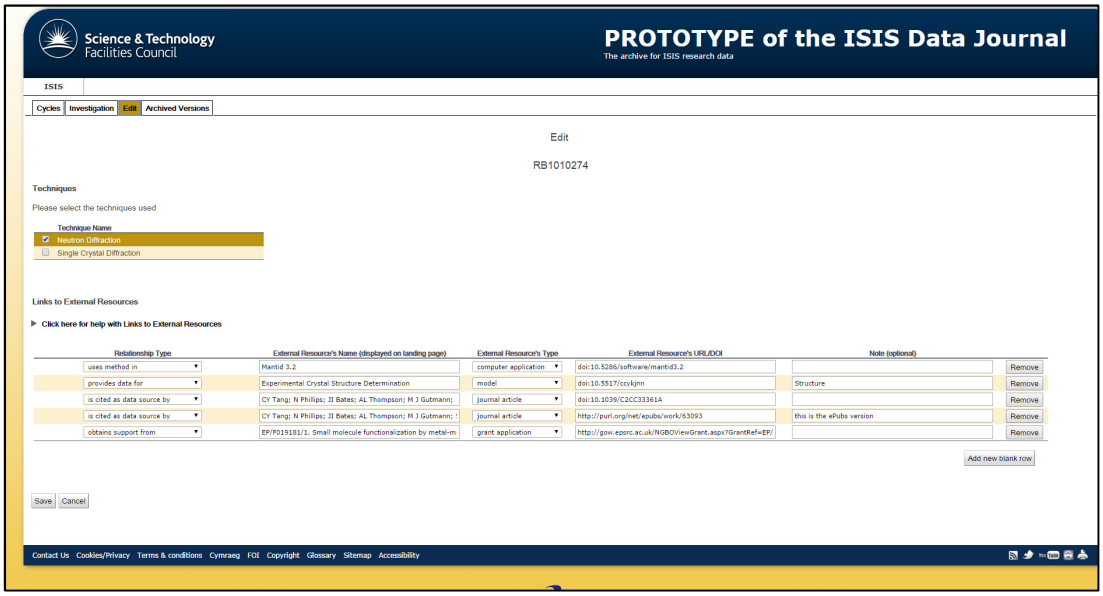


Figure 3 Adding Context to an IRO

In addition we have implemented an automatic look-up for related content in ePubs, the STFC Institutional Repository. ePubs has the capability for relationships to other external resources to be added to a publication's metadata, this includes data DOIs and links to the internal STFC reference to a specific experiment at ISIS, called an RB number. The Data Journal uses the API for ePubs to check for publications referring to the IRO in question.

We don't anticipate allowing the removal of information from a preserved IRO, although parts may be deprecated and so there is no functionality to remove links once the item is preserved.

The current approach assumes that someone related to the content, either the creator or a collaborator would be adding information; this is not a sustainable approach. However fully-automated approaches run the risks of incorrect links and which source is authoritative.

[11] http://purl.org/spar/fabio/

Figure 4 Additional information displayed on the IRO page

4.1.3 Architecture

The major components within the IRO infrastructure are illustrated in Figure 5. Some of these are existing repositories, whiles others have been developed for the prototype. The main components that have been developed are:

- The IRO Builder gathers data from ICAT and writes them out as triples to a triple store
- The Data Journal Service is the central service for viewing and manipulating IROs
- The Data Journal Client is a browser based front end to the Data Journal Service

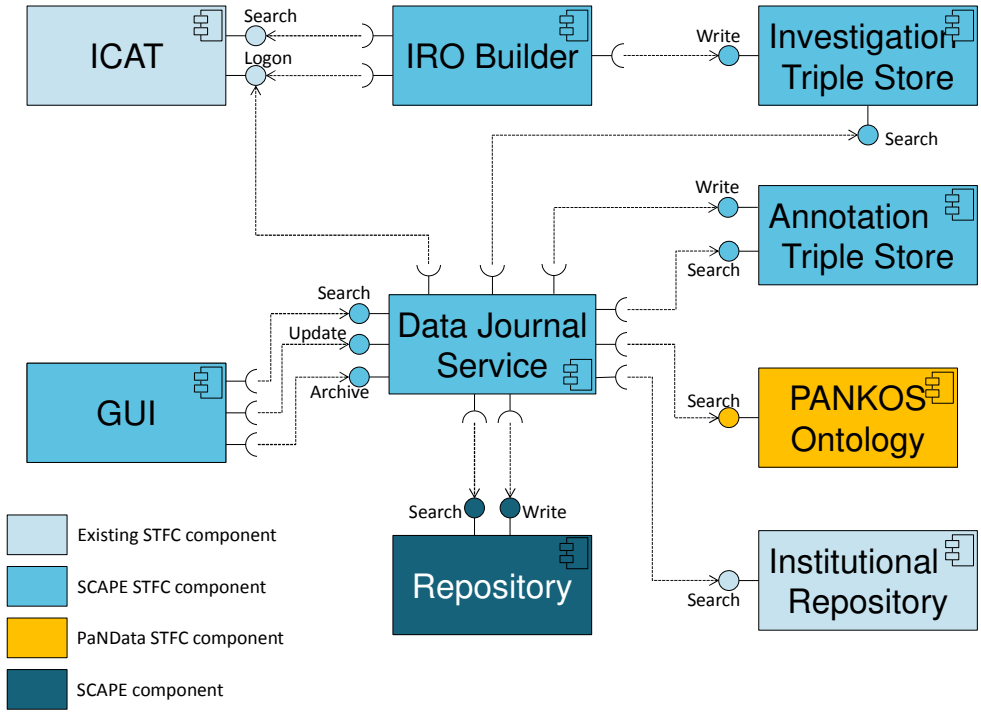


Figure 5: Components providing the IRO functionality

The concept of linked data is at the heart of research objects; therefore it is important that the data be made available in the form of triples.

The IRO Builder extracts the metadata from ICAT and produces CSMD conformant RDF, which is then used as input to a triple store. The IRO Builder takes two timestamps as parameters, which are used to retrieve data modified between these two timestamps. This makes it straightforward to keep the triple store up to date with the selected values in ICAT. These metadata about investigations then form the core of the IROs. A Fuseki[12] triple store is used to store the triples generated by the IRO Builder.

The relationships with other resources are made using Open Annotations[13], a vocabulary under development within the World-Wide Web Consortium for the representation of shared annotations of web accessible resources. In this case, a annotation has a body and a target, together with some associated metadata documenting the annotation. The Figure 6 shows an annotation from the PaNKOS ontology. Here the target is the investigation and the body is information about the target, i.e. the PaNKOS term; that this is providing a term in the ontology to categorise the investigation is recorded by the use of the value *tagging* for motivation of the annotation.

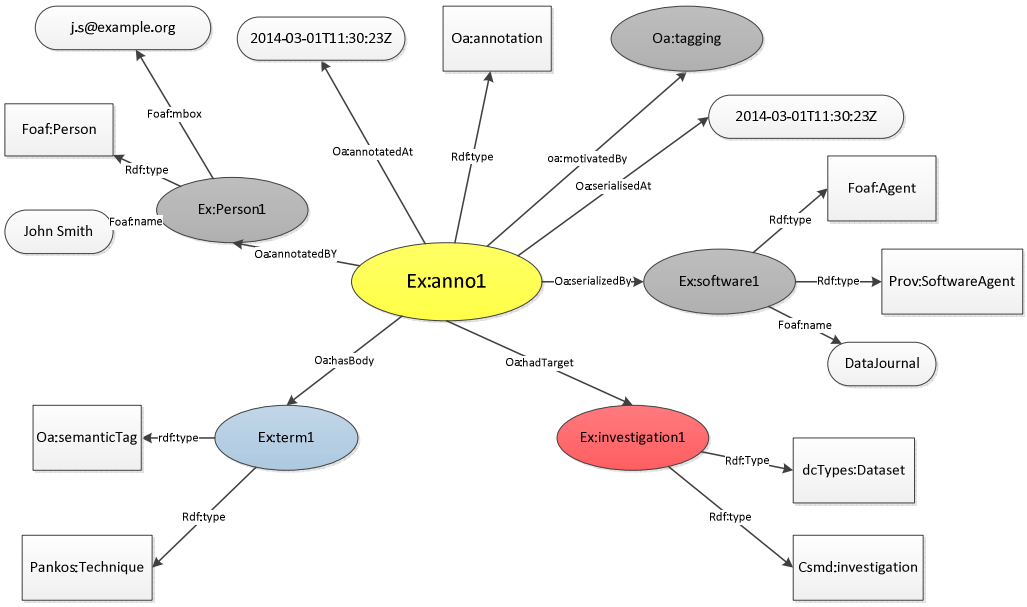


Figure 6 Example of Open Annotation used in the IROs

A similar approach is used to record a relationship to an external resource such a related publication, but in this case the reason for the link, and hence the value of oa:motivatedBy is *commenting*. In addition to the type of oa:Annotation it is also of type cito:CitationAct to enable the description of the type of relationship to be captured; this is illustrated in Figure 7.

[12] http://jena.apache.org/documentation/serving_data/
[13] http://www.w3.org/community/openannotation/

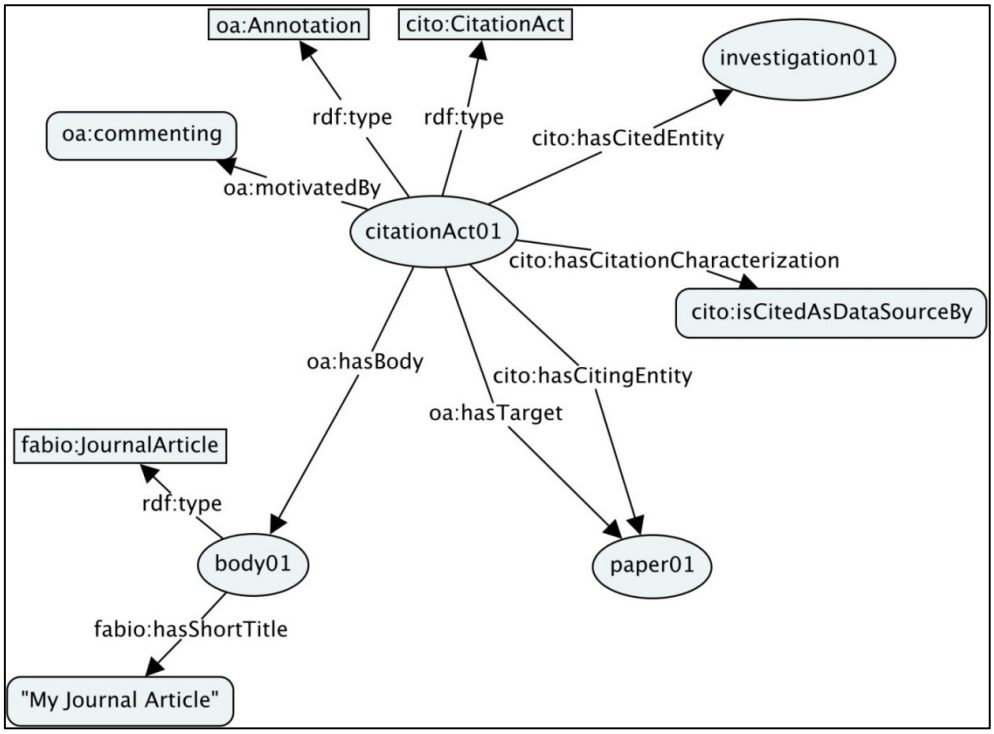


Figure 7: Example of Open Annotation for linking external resources

An IRO will evolve over time as further research outputs are created, such as publications, this leads to an expectation that the minimum criteria for completeness may change over time. The SCAPE IRO validator has been designed to validate against different levels. The IRO Validator assesses an IRO against a given validation level. Three levels are have considered, however currently only level 1 is implemented. The levels are cumulative and so passing level 2 implies that level 1 has been achieved.

- Level 1: check the release date is before now and it has been assigned a DOI.
- Level 2: check that all the links are valid.
- Level 3: check that there are links to one or more publications.

The checks are meant to reflect the lifetime of an IRO. Calling the validator with a level of 2 will first check that the IRO passes the level 1 validation.

4.1.4 Further development

This prototype data journal represents a first step into using the notion of research object to capture and publish contextual and provenance information about facilities experiments. Further opportunities lie in linking to external reference data sources, such the Cambridge Crystallographic Database Centre (CCDC)[14], the Protein Data Bank[15], or the Material Genomes Initiative[16] which provide a deposit of reference data for the analysed structures in different domains; when such data can be associated with facilities experiments, the mecha-

[14] http://www.ccdc.cam.ac.uk
[15] http://www.wwpdb.org/
[16] www.nist.gov/mgi/

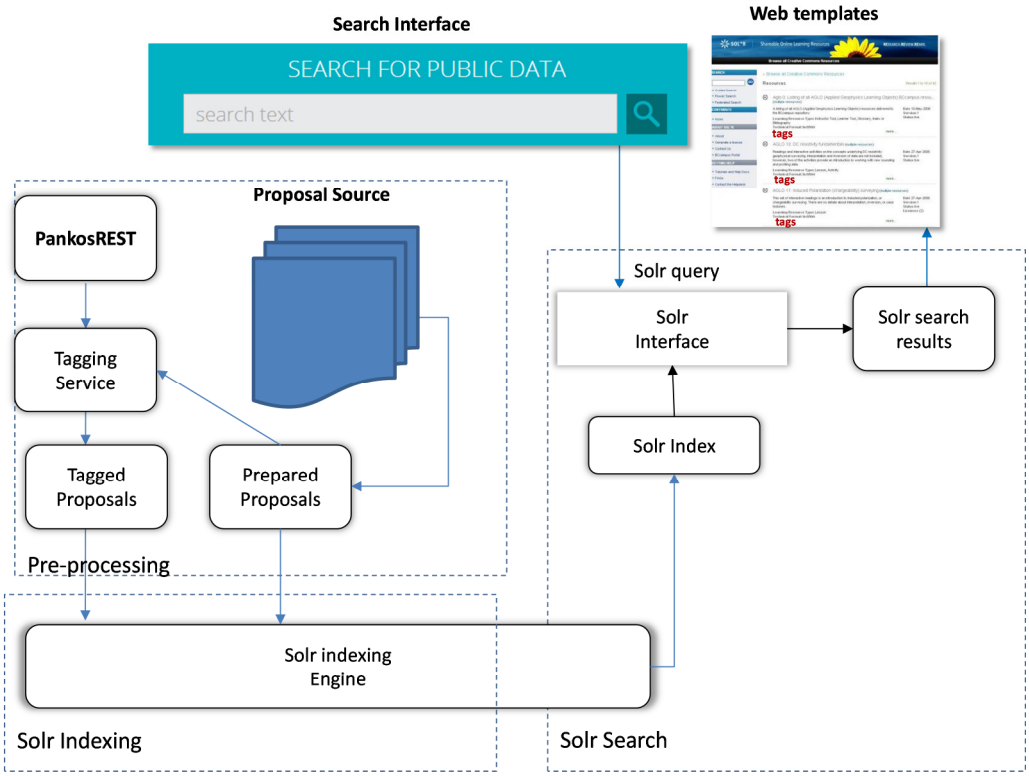
nisms of research objects described here can be used to capture and present the relationship.

4.2 Tagging proposals with PaNKOS[17]

Most facilities organise their proposals and data or by instruments. Some go further by also using uniquely assigned experiment numbers, run numbers, and investigator names. The information about specific experimental techniques employed by an instrument is not captured alongside with the proposals or the data. This is perhaps because this knowledge is typically local to the responsible instrument scientists and it is hard to capture by requesting the scientists (although most facilities provide such information on their website).

This section presents the design and implementation of a small demo system that shows how to use PaNKOS to annotate the proposals of a facility so that the facility and users can search and explore the proposals using experimental techniques, offering a new way to the facilities and the users to examining the research conducted at a facility.

Figure 8 illustrates the system architecture of incorporating PaNKOS into the proposal indexing process of a facility. We have deployed and tested this architecture and the suite of technologies, primarily based on Apache Solr[18], with ILL and ISIS data over the summer of 2014. As a result, two demonstrations have been developed.



[17] This section describes the joint work between STFC and ILL over the summer of 2014.
[18] http://lucene.apache.org/solr/

Figure 8: System Architecture for the Indexing of Facility Proposals

One is a conventional search engine like interface, as shown in Figure 9[19]. The proposal abstract shown here is publically available via its corresponding ISIS data landing page. No extra information is revealed by showing this page in this public deliverable. However, the difference is that experimental techniques are now associated with this particular experiment. In addition, because of the hierarchal organisation of techniques in PaNKOS, this experiment is associated with not only small angle neutron scattering technique but also scattering technique, where the former is a **subTechnique** of the latter.

Figure 9: Annotating the Proposals with PaNKOS - a Solr based search interface

Another demonstration is a more visually attractive presentation based on the same information stored in solr. But, this round, the techniques are indexed as facets within solr. Figure 10 illustrates a visualisation of some generated (not real) proposal data using PaNKOS techniques. So, one step advanced than the other demonstration, this demo leverages the hierarchy of techniques to provide a perhaps more intuitive interface for scientists to explore the collection of proposals and data.

[19] Note that proposal information is obscured as this is from a live proposal system and information may as not as yet be public.

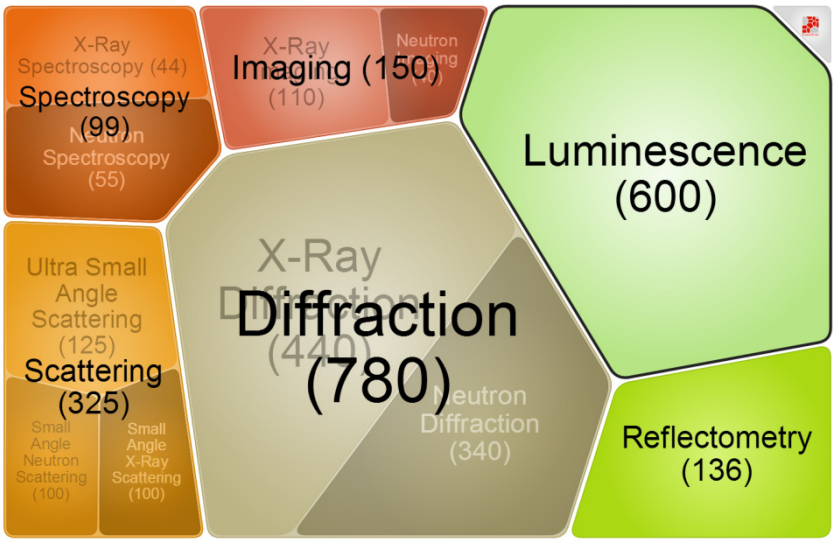


Figure 10: Exploring the Proposals by PaNKOS techniques

5 Publishing PaNData records using OAI-PMH

A further aspect of data sharing is the publishing of core metadata which can be made available to general purpose metadata harvesters, aggregators and search engines which provide search tools for researchers across disciplines and countries. Such tools increase the opportunity that data collected and made publically available by facilities will become accessible for inspection and reuse. One such search engine is provided by DataCite as discussed in Deliverable 7.1. However, this is only available for data which has been published and has had a DOI associated with it. Other metadata aggregators use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[20], a standard for publishing and accessing metadata from archives. Thus, to better support the aggregation and sharing of metadata about facilities experiments, an interface supporting this protocol should be provided.

In order to provide a OAI-PMH service, PaNData collaborated with the EU funded EUDAT project[21], that delivers four key services:

- B2SHARE - service for individual researchers or teams to share their "long tail data",
- B2SAFE - safe and highly available replication service,
- B2FIND – data discovery service allowing users to find data stored in B2SAFE and B2SHARE, also collecting a large number of dataset descriptions from various disciplines, with links to the actual datasets stored elsewhere beyond EUDAT,
- B2STAGE – service for shipping large amounts of data to high-performance computing systems.

[20] http://www.openarchives.org/pmh/
[21] www.eudat.eu

Thus PaNData collaborated with EUDAT in order to:

- make an evaluation of the technology available and to install an OAI-PMH endpoint that serves as a machine interface between PaNData and EUDAT,
- create a metadata model that maps PaNData records stored in ICAT to XML records published through OAI-PMH,
- create a metadata harvester, converter and publisher component that requests data from ICAT and populates the OAI-PMH endpoint with the actual PaNData records,
- agree on the mapping of the PaNData records published through OAI-PMH with the EUDAT metadata published through B2FIND,
- create metadata converters and publish transformed PaNData records in B2FIND,
- verify PaNData records in B2FIND with the original PaNData record owners.

A few OAI-PMH platforms for publishing PaNData records were considered, with the list in Appendix E. The choice for the implementation was jOAI, a Java implementation of OAI-PMH; it has good documentation and it is mature.

The metadata model for publishing PaNData records in OAI-PMH is based on Qualified Dublin Core which is a well-known generic metadata format, understandable by a variety of data practitioners across the globe. Appendix F contains the mapping of PaNData Core Science Metadata model supporting within the ICAT catalogues to the OAI-PMH schema and corresponding fields in the EUDAT metadata schema.

The harvester, converter and publisher from ICAT to Qualified Dublin Core in OAI-PMH endpoint was implemented in Python with some details available in Appendix G.

When testing the PaNData to EUDAT interface, about 700 publically released STFC ISIS records were chosen, processed and published through B2FIND service[22]. Verification of the representation of PaNData ISIS records in EUDAT B2FIND by the original record owners has been requested, and the communication channel with EUDAT is open for discussion. The implementation of PaNData OAI-PMH service for ISIS is currently published at:

https://icat-10.PaNData.stfc.ac.uk/oai/provider?verb=Identify

Other PaNData communities can configure their data as a “set” in the existing OAI-PMH endpoint, or they can set up their own OAI-PMH endpoint using the same technology.

[22] http://b2find.eudat.eu/group/PaNData

6 A Final word on Provenance

Within the work on provenance PaNData we have covered a range of topics within area of the supporting the wider scientific lifecycle beyond the data collection. We have demonstrated the need for support for recording analysis steps within data management tools, which is now supported within ICAT and linked to within tool frameworks such as Mantid. We have shown the utility of providing a common contextual framework of vocabulary to record and share information such as techniques and other properties. Investigation research objects offer the opportunity to systematically record provenance and other contextual links and publish as a single intellectual entity, the record of the science undertaken.

However, work within this topic is not as yet complete, and we provide a view to the future, building on the work of this workpackage and other work within PaNData. Currently, many instrument scientists within facilities spend a portion of their time helping users to install data analysis software. In order to ensure a productive yield from the experiments, scientists also invest vast efforts in guiding and/or directly involving in the data analysis process. Most scientists rely on conventional means, emails, sometimes, site visits, to help users analysing data, who has typically returned to their home institution, which may have limited access to the right computing resources. This can be a time consuming and costly process, to users, scientists, and ultimately, the facility, affecting the facility's capability and capacity to support, enable, and deliver an excellent science programme. We believe that these are significant factors towards the 2-year time lag between an experiment and papers, and the fact that, many experiments to not ever have their results published. However, facilities tend to have the right expertise, the right software and access to high-performance computing capability. This situation can be drastically improved, if data management technologies supported by such technologies as Clouds are employed in the right way so that experimental scientists can access compute resources at facilities after the experiment.

If virtual machines all the data analysis software installed were to be offered to experimental teams to support post experiment data analysis pipelines. This would provide access to the data, be it from an experiment, data analysis, and/or simulation process. Compute resources would then be provided for data processing, regardless when, where, and how data are produced (with a suitable resourcing model). Storage and compute resources will be dynamically managed, supporting the needs of different instruments, and, different types of processing. The data would be systematically, flexibly and securely managed, meaning that:

- Data would be securely guarded so that only the authorised personnel can access the virtual machine;
- Data would be backed up so that scientists do not need to worry about losing their work;
- Data provenance would be systematically catalogued throughout the whole analysis pipeline so that scientists can come back to the data any point, and search through them easily;
- Data would be archived in a long term archival storage so that data can be made available for long term use, and

- Finally, data and the virtual machines would be made accessible to the collaboration team remotely without users worrying about software installation, configurations, and updates.

From a compute point of view, users would not need to worry about moving data or sending data between members of the team. Data would be made available on any VM, appearing as a native part of the file system. All these would significantly lower the entry bar of analysing science data, leading to profound positive impact on the excellent science that facilities do.

Appendix A. A survey Questionnaire about metadata capture, storage, usage, and standardisation

This survey aims to provide a snapshot of the role of metadata in the European photon and neutron facilities.

A) About your facility

1. Is your facility a neutron facility or a photon facility? Neutron/Photon/Both/Both and more

B) Data storage

2. Store experiment data 6 months after they are captured? Yes/No/Planned
3. Store analysed data 6 months after they are captured? Yes/No/Planned
4. Store supplementary data for papers? Yes/No/Planned

C) Metadata capture

5. Capture metadata about users and proposals? Yes/No/Planned
6. Capture metadata about experiment data? Yes/No/Planned

D) Metadata storage/catalogue mechanism

7. Use databases to store information about experiment data? Yes/No/Planned
8. Store experiment metadata inside data files (file format is _irrelevant_in this context)? Yes/No/Planned
9. Use databases to store information about analysed data? Yes/No/Planned
10. Use databases to store information about supplementary data about papers? Yes/No/Planned

E) Metadata standardisation

11. Use a set of mandatory metadata to describe the data gathered at your facility, regardless of which beamline/instrument they are collected? Yes/No/Planned
12. Use metadata standard to describe data gathered at your facility? Yes/No/Planned

F) Metadata usage in services

13. Use DOIs in your facility? Yes/No/Planned

G) Others (Optional)

14. In general, synchrotron facilities collect more diverse range of metadata than neutron facilities. Is it true? Yes/No

Notes:

* If you can be specific, you are very welcome to write additional comments.
* If your organisation runs multiple facilities (e.g. neutron and synchrotron), you are welcome to produce a reply for each facility.

======

Metadata is data about data. Metadata can describe anything about data, for example,
- Data content: what is inside a data file or a set of data files from an experiment, this can be words, or snapshots of images, or a significant part of spectra related to experiment data
- Data context: who, when, where, what about an experiment, and how data is collected, e.g. parameters, experiment techniques (e.g. tomography, diffraction), name of facility, name of investigators, affiliations etc., software that produce the data
- Data location: where data is stored, e.g. the file path on disk, tape, or a remote server
- Data identifier: how the data should be referred to. Digital Object Identifier (DOI) is an example of such.
- The format of data, e.g. Nexus, RAW, TIFF

Appendix B. Sections of PaNKOS describing ISIS and ILL, the instruments, and the techniques

```
#################################################################
#
#    Individuals – Facilities – ILL and ISIS
#
#################################################################

PaNKOS:ISIS rdf:type PaNKOS:NeutronSource ,
                :NamedIndividual ;
        PaNKOS:preferredName "ISIS"^^rdfs:Literal ;
        PaNKOS:hasInstrument PaNKOS:ALF  ,
                        PaNKOS:CHIPIR  ,
                        PaNKOS:DEVA  ,
                        PaNKOS:EXEED  ,
                        PaNKOS:HET  ,
                        PaNKOS:IMAT  ,
                        PaNKOS:LMX  ,
                        PaNKOS:PEARL  ,
                        PaNKOS:ZOOM  ,
                        PaNKOS:ARGUS ,
                        PaNKOS:CRISP ,
                        PaNKOS:EMU ,
                        PaNKOS:ENGIN-X ,
                        PaNKOS:GEM ,
                        PaNKOS:HIFI ,
                        PaNKOS:HRPD ,
                        PaNKOS:INES ,
                        PaNKOS:INTER ,
                        PaNKOS:IRIS ,
                        PaNKOS:LET ,
                        PaNKOS:LOQ ,
                        PaNKOS:MAPS ,
                        PaNKOS:MARI ,
                        PaNKOS:MERLIN ,
                        PaNKOS:MUSR ,
                        PaNKOS:NIMROD ,
```

```
                        PaNKOS:OFFSPEC ,
                        PaNKOS:OSIRIS ,
                        PaNKOS:POLARIS ,
                        PaNKOS:POLREF ,
                        PaNKOS:ROTAX ,
                        PaNKOS:SANDALS ,
                        PaNKOS:SANS2D ,
                        PaNKOS:SURF ,
                        PaNKOS:SXD ,
                        PaNKOS:TOSCA ,
                        PaNKOS:VESUVIO ,
                        PaNKOS:WISH .


PaNKOS:ILL rdf:type PaNKOS:NeutronSource ,
                :NamedIndividual ;
        PaNKOS:preferredName "ILL"^^xsd:string ;
        rdfs:comment "Institut Laue-Langevin"^^xsd:string ;
        PaNKOS:hasInstrument PaNKOS:BRISP ,
                        PaNKOS:CYCLOPS ,
                        PaNKOS:CryoDEM ,
                        PaNKOS:D10 ,
                        PaNKOS:D11 ,
                        PaNKOS:D16 ,
                        PaNKOS:D17 ,
                        PaNKOS:D18 ,
                        PaNKOS:D19 ,
                        PaNKOS:D1B ,
                        PaNKOS:D20 ,
                        PaNKOS:D22 ,
                        PaNKOS:D23 ,
                        PaNKOS:D2B ,
                        PaNKOS:D3 ,
                        PaNKOS:D33 ,
                        PaNKOS:D4 ,
                        PaNKOS:D7 ,
                        PaNKOS:D9 ,
                        PaNKOS:FIGARO ,
                        PaNKOS:GRANIT ,
                        PaNKOS:IN1 ,
                        PaNKOS:IN10 ,
                        PaNKOS:IN11 ,
                        PaNKOS:IN13 ,
                        PaNKOS:IN14 ,
                        PaNKOS:IN15 ,
                        PaNKOS:IN16B ,
                        PaNKOS:IN20 ,
                        PaNKOS:IN22 ,
                        PaNKOS:IN3 ,
                        PaNKOS:IN4C ,
                        PaNKOS:IN5 ,
                        PaNKOS:IN6 ,
                        PaNKOS:IN8 ,
                        PaNKOS:LADI-III ,
                        PaNKOS:OrientExpress ,
```

```
                    PaNKOS:PF1B ,
                    PaNKOS:PF2 ,
                    PaNKOS:PN1 ,
                    PaNKOS:PN3 ,
                    PaNKOS:S18 ,
                    PaNKOS:SALSA ,
                    PaNKOS:SuperADAM ,
                    PaNKOS:VIVALDI .


#################################################################
#
#    Individuals - ISIS Instruments and techniques
#
#################################################################

PaNKOS:ALF rdf:type :NamedIndividual ;
        PaNKOS:preferredName "ALF"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:CHIPIR rdf:type :NamedIndividual ;
        PaNKOS:preferredName "CHIPIR"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:CosmicNeutronRadiation .

PaNKOS:EXEED rdf:type :NamedIndividual ;
        PaNKOS:preferredName "EXEED"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:Time-of-FlightDiffraction .

PaNKOS:HET  rdf:type :NamedIndividual ;
        PaNKOS:preferredName "HET"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:Spectroscopy .

PaNKOS:IMAT  rdf:type :NamedIndividual ;
        PaNKOS:preferredName "IMAT"^^rdfs:Literal ;
        PaNKOS:supportsTechnique                    PaNKOS:NeutronTomography,
PaNKOS:NeutronImaging, PaNKOS:NeutronDiffraction .

PaNKOS:LMX  rdf:type :NamedIndividual ;
        PaNKOS:preferredName "LMX"^^rdfs:Literal ;
   PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:DEVA rdf:type :NamedIndividual ;
        PaNKOS:preferredName "DEVA"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:RF-mSR .

PaNKOS:PEARL  rdf:type :NamedIndividual ;
        PaNKOS:preferredName "PEARL"^^rdfs:Literal ;
  PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .

PaNKOS:ZOOM  rdf:type :NamedIndividual ;
         PaNKOS:preferredName "ZOOM"^^rdfs:Literal ;
         PaNKOS:supportsTechnique PaNKOS:VSANS  .

PaNKOS:ARGUS rdf:type PaNKOS:MuonSpectrometer ,
                :NamedIndividual ;
```

```
        PaNKOS:preferredName "ARGUS"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:MuonSpectroscopy .

PaNKOS:CRISP rdf:type PaNKOS:Reflectometer ,
                :NamedIndividual ;
        PaNKOS:preferredName "CRISP"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:Reflectometry .

PaNKOS:EMU rdf:type PaNKOS:MuonSpectrometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "EMU"^^rdfs:Literal ;
       PaNKOS:supportsTechnique PaNKOS:MuonSpectroscopy .

PaNKOS:ENGIN-X rdf:type PaNKOS:NeutronDiffractometer ,
                  :NamedIndividual ;
          PaNKOS:preferredName "ENGIN-X"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:GEM rdf:type PaNKOS:NeutronDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "GEM"^^rdfs:Literal ;
       PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .

PaNKOS:HIFI rdf:type PaNKOS:MuonSpectrometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "HIFI"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:MuonSpectroscopy .

PaNKOS:HRPD rdf:type :NamedIndividual ;
        PaNKOS:preferredName "HRPD"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .

PaNKOS:INES rdf:type PaNKOS:NeutronDiffractometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "INES"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .

PaNKOS:INTER rdf:type PaNKOS:Reflectometer ,
                :NamedIndividual ;
        PaNKOS:preferredName "INTER"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:Reflectometry .

PaNKOS:IRIS rdf:type PaNKOS:NeutronSpectrometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "IRIS"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:NeutronSpectroscopy .

PaNKOS:LET rdf:type PaNKOS:NeutronSpectrometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "LET"^^rdfs:Literal ;
       PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:LOQ rdf:type PaNKOS:SANSInstrument ,
              :NamedIndividual ;
       PaNKOS:preferredName "LOQ"^^rdfs:Literal ;
```

```
        PaNKOS:supportsTechnique PaNKOS:SmallAngleNeutronScattering .

PaNKOS:MAPS rdf:type PaNKOS:NeutronSpectrometer ,
                :NamedIndividual ;
          PaNKOS:preferredName "MAPS"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:MARI rdf:type PaNKOS:NeutronSpectrometer ,
                :NamedIndividual ;
          PaNKOS:preferredName "MARI"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:NeutronSpectroscopy .

PaNKOS:MERLIN rdf:type PaNKOS:NeutronSpectrometer ,
                  :NamedIndividual ;
            PaNKOS:preferredName "MERLIN"^^rdfs:Literal ;
            PaNKOS:supportsTechnique PaNKOS:NeutronSpectroscopy .

PaNKOS:MUSR rdf:type PaNKOS:MuonSpectrometer ,
                :NamedIndividual ;
          PaNKOS:preferredName "MUSR"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:MuonSpectroscopy .

PaNKOS:NIMROD rdf:type PaNKOS:NeutronDiffractometer ,
                  PaNKOS:SANSInstrument ,
                  :NamedIndividual ;
            PaNKOS:preferredName "NIMROD"^^rdfs:Literal ;
            PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:OFFSPEC rdf:type PaNKOS:Reflectometer ,
                   :NamedIndividual ;
             PaNKOS:preferredName "OFFSPEC"^^rdfs:Literal ;
             PaNKOS:supportsTechnique PaNKOS:Reflectometry .

PaNKOS:OSIRIS rdf:type PaNKOS:NeutronSpectrometer ,
                  :NamedIndividual ;
            PaNKOS:preferredName "OSIRIS"^^rdfs:Literal ;
            PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:POLARIS rdf:type PaNKOS:NeutronDiffractometer ,
                   :NamedIndividual ;
             PaNKOS:preferredName "POLARIS"^^rdfs:Literal ;
             PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .

PaNKOS:POLREF rdf:type PaNKOS:Reflectometer ,
                  :NamedIndividual ;
            PaNKOS:preferredName "POLREF"^^rdfs:Literal ;
            PaNKOS:supportsTechnique PaNKOS:NeutronDepolarisation ,
                                     PaNKOS:PolarisedNeutronReflectometry ,
                                     PaNKOS:Reflectometry .

PaNKOS:ROTAX rdf:type PaNKOS:NeutronDiffractometer ,
                 :NamedIndividual ;
           PaNKOS:preferredName "ROTAX"^^rdfs:Literal ;
           PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .
```

```
PaNKOS:SANDALS rdf:type PaNKOS:NeutronDiffractometer ,
                 PaNKOS:SANSInstrument ,
                 :NamedIndividual ;

          PaNKOS:preferredName "SANDALS"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction .

PaNKOS:SANS2D rdf:type PaNKOS:SANSInstrument ,
                :NamedIndividual ;
         PaNKOS:preferredName "SANS2D"^^rdfs:Literal ;
         PaNKOS:supportsTechnique PaNKOS:SmallAngleNeutronScattering .

PaNKOS:SURF rdf:type PaNKOS:Reflectometer ,
               :NamedIndividual ;
        PaNKOS:preferredName "SURF"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:Reflectometry .

PaNKOS:SXD rdf:type PaNKOS:NeutronDiffractometer ,
              PaNKOS:SingleCrystalDiffractometer ,
              :NamedIndividual ;

       PaNKOS:preferredName "SXD"^^rdfs:Literal ;
       PaNKOS:supportsTechnique PaNKOS:NeutronDiffraction ,
                       PaNKOS:SingleCrystalDiffraction .

PaNKOS:TOSCA rdf:type PaNKOS:NeutronSpectrometer ,
               :NamedIndividual ;

         PaNKOS:preferredName "TOSCA"^^rdfs:Literal ;
         PaNKOS:supportsTechnique PaNKOS:NeutronSpectroscopy .

PaNKOS:VESUVIO rdf:type PaNKOS:NeutronSpectrometer ,
                 :NamedIndividual ;
          PaNKOS:preferredName "VESUVIO"^^rdfs:Literal ;
          PaNKOS:supportsTechnique PaNKOS:NeutronSpectroscopy .

PaNKOS:WISH rdf:type PaNKOS:NeutronDiffractometer ,
               :NamedIndividual ;
        PaNKOS:preferredName "WISH"^^rdfs:Literal ;
        PaNKOS:supportsTechnique PaNKOS:NeutronPowderDiffraction .


###################################################################
#
#    Individuals - ILL Instruments and techniques
#
###################################################################

PaNKOS:BRISP rdf:type PaNKOS:Time-of-flightSpectrometer ,
                :NamedIndividual ;
         PaNKOS:preferredName "BRISP"^^xsd:string ;
         rdfs:comment "BRISP - TOF Spectrometer for Small Angle Inelastic Scatter-
ing"^^xsd:string ;
         PaNKOS:supportsTechnique PaNKOS:SmallAngleInelasticScattering .
```

```
PaNKOS:CYCLOPS rdf:type PaNKOS:LaueSingleDiffractometer ,
                PaNKOS:SingleCrystalDiffractometer ,
                :NamedIndividual ;
          PaNKOS:preferredName "CYCLOPS"^^xsd:string ;
          rdfs:comment "CYlindrical Ccd Laue Octagonal Photo Scintillator"^^xsd:string ;
          PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .


PaNKOS:CryoDEM rdf:type PaNKOS:NuclearParticlePhysics ,
                :NamedIndividual ;
          PaNKOS:preferredName "CryoDem"^^xsd:string .

PaNKOS:D10 rdf:type PaNKOS:SingleCrystalDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D10"^^xsd:string ;
       rdfs:comment "Single-crystal four-circle diffractometer with three-axis energy analysis"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:D11 rdf:type PaNKOS:LargeScaleDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D11"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:SmallAngleNeutronScattering .

PaNKOS:D16 rdf:type PaNKOS:LargeScaleDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D16"^^xsd:string ;
       rdfs:comment "Small momentum transfer diffractometer with variable vertical focusing"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:Diffraction .

PaNKOS:D17 rdf:type PaNKOS:Reflectometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D17"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .

PaNKOS:D18 rdf:type PaNKOS:PowerDiffractometer ,
              :NamedIndividual .

PaNKOS:D19 rdf:type PaNKOS:SingleCrystalDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D19"^^xsd:string ;
       rdfs:comment "Thermal neutron diffractometer for single-crystal and fibre diffraction D19"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:D1B rdf:type PaNKOS:PowerDiffractometer ,
              :NamedIndividual ;
       PaNKOS:supportsTechnique PaNKOS:PowderDiffraction .

PaNKOS:D20 rdf:type PaNKOS:PowerDiffractometer ,
              :NamedIndividual ;
       PaNKOS:preferredName "D20"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:PowderDiffraction .
```

```
PaNKOS:D22 rdf:type PaNKOS:LargeScaleDiffractometer ,
                :NamedIndividual ;
        PaNKOS:preferredName "D22"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:SmallAngleNeutronScattering .

PaNKOS:D23 rdf:type PaNKOS:SingleCrystalDiffractometer ,
                :NamedIndividual ;
        PaNKOS:preferredName "D23"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .


PaNKOS:D2B rdf:type PaNKOS:PowerDiffractometer ,
                :NamedIndividual ;

        PaNKOS:preferredName "D2B"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:PowderDiffraction .

PaNKOS:D3 rdf:type PaNKOS:SingleCrystalDiffractometer ,
         :NamedIndividual ;
         PaNKOS:preferredName "D33"^^xsd:string .

PaNKOS:D33 rdf:type PaNKOS:LargeScaleDiffractometer ,
                :NamedIndividual ;
        PaNKOS:preferredName "D33"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:SmallAngleDiffraction ,
                        PaNKOS:SmallAngleNeutronScattering .


PaNKOS:D4 rdf:type PaNKOS:PowerDiffractometer ,
               :NamedIndividual ;
       PaNKOS:preferredName "D4"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:PowderDiffraction .

PaNKOS:D7 rdf:type PaNKOS:Time-of-flightSpectrometer ,
               :NamedIndividual ;
       dc:description  "D7"^^xsd:string ;
       PaNKOS:preferredName "D7"^^xsd:string ;
       rdfs:comment "Diffuse Scattering Spectrometer"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .

PaNKOS:D9 rdf:type PaNKOS:SingleCrystalDiffractometer ,
               :NamedIndividual ;
       rdfs:comment " Hot neutron four-circle diffractometer"^^xsd:string ;
       PaNKOS:preferredName "D9"^^xsd:string ;
       PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:FIGARO rdf:type PaNKOS:Reflectometer ,
                  PaNKOS:Time-of-flightSpectrometer ,
                  :NamedIndividual ;
          PaNKOS:preferredName "FIGARO"^^xsd:string ;
          rdfs:comment "Fluid Interfaces Grazing Angles ReflectOmeter"^^xsd:string ;
          PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .


PaNKOS:GRANIT rdf:type PaNKOS:NuclearParticlePhysics ,
```

```
                :NamedIndividual ;
          rdfs:comment "A 2nd-generation gravitational neutron spectrometer"^^xsd:string ;
          PaNKOS:preferredName "GRANIT"^^xsd:string .

PaNKOS:IN1 rdf:type PaNKOS:ThreeAxisSpectrometer ,
              :NamedIndividual .

PaNKOS:IN10 rdf:type PaNKOS:BackScatteringSpectrometer ,
              :NamedIndividual ;
        rdfs:comment "Cold neutron backscattering spectrometer IN10"^^xsd:string ;
        PaNKOS:preferredName "IN10"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:ElasticNeutronScatteringSpectroscopy ,

        PaNKOS:InelasticNeutronScatteringSpectroscopy .

PaNKOS:IN11 rdf:type PaNKOS:Spin-echoSpectrometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "IN11"^^xsd:string ;
        rdfs:comment "Spin-Echo Spectrometer IN11"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:SpinEchoSANS .

PaNKOS:IN12 rdf:type PaNKOS:ThreeAxisSpectrometer ,
              :NamedIndividual .

PaNKOS:IN13 rdf:type PaNKOS:NeutronSpectrometer ,
              :NamedIndividual ;
        rdfs:comment "CRG - thermal neutron backscattering spectrometer IN13"^^xsd:string
;
        PaNKOS:preferredName "IN13"^^xsd:string ;
        PaNKOS:supportsTechnique                PaNKOS:BackScatteringSpectroscopy,
PaNKOS:ToFSpectroscopy .

PaNKOS:IN14 rdf:type PaNKOS:ThreeAxisSpectrometer ,
        :NamedIndividual ;
   PaNKOS:preferredName "IN14"^^xsd:string .

PaNKOS:IN15 rdf:type PaNKOS:Spin-echoSpectrometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "IN15"^^xsd:string ;
        rdfs:comment "Spin-echo spectrometer with time-of-flight option and focussing op-
tion"^^xsd:string ;
        PaNKOS:supportsTechnique PaNKOS:SpinEchoSmallAngleNeutronScattering .

PaNKOS:IN16B rdf:type PaNKOS:NeutronSpectrometer ,
               :NamedIndividual ;
          PaNKOS:preferredName "IN16B"^^xsd:string ;
          PaNKOS:supportsTechnique PaNKOS:BackScatteringSpectroscopy  .

PaNKOS:IN20 rdf:type PaNKOS:ThreeAxisSpectrometer ,
              :NamedIndividual ;
        PaNKOS:preferredName "IN20"^^xsd:string ;
        rdfs:comment "Thermal neutron three-axis spectrometer with polarisation analy-
sis"^^xsd:string .

PaNKOS:IN22 rdf:type PaNKOS:ThreeAxisSpectrometer ,
```

```
        :NamedIndividual ;
   PaNKOS:preferredName "IN20"^^xsd:string .

PaNKOS:IN3 rdf:type PaNKOS:TestInstrument ,
                :NamedIndividual ;
   PaNKOS:preferredName "IN3"^^xsd:string ,
                                "Test Instrument"^^xsd:string .

PaNKOS:IN4C rdf:type PaNKOS:Time-of-flightSpectrometer ,
                 :NamedIndividual ;
         PaNKOS:preferredName "IN4C"^^xsd:string ;
         PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .

PaNKOS:IN5 rdf:type PaNKOS:Time-of-flightSpectrometer ,
                :NamedIndividual ;
   PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .

PaNKOS:IN6 rdf:type PaNKOS:Time-of-flightSpectrometer ,
                :NamedIndividual ;
   PaNKOS:preferredName "Cold neutron time-focussing time-of-flight spectrometer
"^^xsd:string ;
   PaNKOS:supportsTechnique PaNKOS:Time-of-FlightSANS .

PaNKOS:IN8 rdf:type PaNKOS:ThreeAxisSpectrometer ,
                :NamedIndividual .

PaNKOS:LADI-III rdf:type PaNKOS:LargeScaleDiffractometer ,
                    PaNKOS:Quasi-LaueDiffractometer ,
                    :NamedIndividual ;
         PaNKOS:preferredName "LADI-III"^^xsd:string ;
         rdfs:comment "Quasi-Laue diffractometer LADI-III"^^xsd:string ;
         PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction .

PaNKOS:OrientExpress rdf:type PaNKOS:SingleCrystalDiffractometer ,
                       :NamedIndividual ;
          PaNKOS:preferredName "Orient Express"^^xsd:string ;
          PaNKOS:supportsTechnique PaNKOS:NeutronLaueDiffraction ,

PaNKOS:SingleCrystalDiffraction .


PaNKOS:PF1B rdf:type PaNKOS:NuclearParticlePhysics ,
                 :NamedIndividual ;
         PaNKOS:preferredName "PF1B"^^xsd:string .

PaNKOS:PF2 rdf:type PaNKOS:NuclearParticlePhysics ,
                :NamedIndividual ;
         PaNKOS:preferredName "PF2"^^xsd:string ;
         rdfs:comment "Ultracold neutron facility"^^xsd:string .

PaNKOS:PN1 rdf:type PaNKOS:NuclearParticlePhysics ,
                PaNKOS:Spectrometer ,
                :NamedIndividual ;
   rdfs:comment "PN1 (Lohengrin) an exotic isotope spectrometer"^^xsd:string ;
   PaNKOS:preferredName "PN1"^^xsd:string .
```

```
PaNKOS:PN3 rdf:type PaNKOS:GammaRaySpectrometer ,
                PaNKOS:NuclearParticlePhysics ,
                :NamedIndividual ;
          rdfs:comment "Gamma-Ray Spectrometers GAMS4 and GAMS5"^^xsd:string ;
          PaNKOS:preferredName "PN3"^^xsd:string ;
            PaNKOS:supportsTechnique PaNKOS:SingleCrystalDiffraction ,
                                                        PaNKOS:Spectroscopy .

PaNKOS:S18 rdf:type PaNKOS:NeutronInterferometer ,
                :NamedIndividual ;
            rdfs:comment "Neutron interferometer"^^xsd:string ;
            PaNKOS:preferredName "S18"^^xsd:string ;
            PaNKOS:supportsTechnique PaNKOS:NeutronInterferometry ,
                                                        PaNKOS:USANS .

PaNKOS:SALSA rdf:type PaNKOS:PowerDiffractometer ,
                  :NamedIndividual ;
                    PaNKOS:preferredName "SALSA"^^xsd:string ;
                    rdfs:comment "Strain imager for engineering applications SALSA"^^xsd:string
;
                    PaNKOS:supportsTechnique PaNKOS:PowderDiffraction .

PaNKOS:SuperADAM rdf:type PaNKOS:Reflectometer ,
                     :NamedIndividual ;
                    PaNKOS:preferredName "SuperADAM"^^xsd:string ;
                    rdfs:comment "The reflectometer SuperADAM is an angle dispersive fixed
wavelength machine which combines high flux due to a focussing monochromator with a
high Q resolution."^^xsd:string ;
                    PaNKOS:supportsTechnique PaNKOS:USANS .

PaNKOS:VIVALDI rdf:type PaNKOS:SingleCrystalDiffractometer ,
                    :NamedIndividual ;
            PaNKOS:preferredName "VIVALDI"^^xsd:string .
```

Appendix C. Weblinks of PaNKOS

PaNKOS files and more: http://PaNData.org/ontology/
PaNKOS Fuseki web server for interactive access to PaNKOS: http://PaNData.org:8009/

Appendix D. PaNKOSREST APIs

The REST web application is deployed at an internal host at RAL named: http://ahost.rl.ac.uk. It offers 6 types of information from PaNKOS. They are shown as follows. For brevity, we only include a sample json return message for the last API.

a) Get facilities and instruments:
http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q1

SPARQL query:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>

SELECT ?facility ?instruments
WHERE {
 ?facilityType rdfs:subClassOf PaNKOS:Facility .
 ?facility          rdf:type ?facilityType ;
                              PaNKOS:hasInstrument ?instruments .
 }
order by ?facility
```

b) Get things in PaNKOS: http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q2

SPARQL Query:

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>

select ?ThingsInPaNKOS
where
{
  ?ThingsInPaNKOS rdfs:subClassOf owl:Thing.
}
```

c) Get all PaNKOS techniques:
http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q3

SPARQL Query:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>

select ?allPaNKOSTechniques
where
{
  ?x PaNKOS:hasSubTechnique+/PaNKOS:preferredName ?allPaNKOSTechniques .
}
```

d) Get all facilities: http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q4

SPARQL Query:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>

   SELECT ?facility
   WHERE {
    ?facility rdf:type PaNKOS:NeutronSource .
}
```

e) Get facilities and techniques:
http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q5

SPARQL Query:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>
prefix dc: <http://purl.org/dc/elements/1.1/>

SELECT  ?inst ?tech
WHERE {
   PaNKOS:ISIS PaNKOS:hasInstrument ?inst .
   ?inst       PaNKOS:supportsTechnique  ?tech .
}
```

f) Get facilities, instruments, and techniques:
http://ahost.rl.ac.uk:8080/PaNKOSRest/rest/PaNKOS/q6

SPARQL Query:

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix PaNKOS: <http://www.purl.org/PaNKOS#>

SELECT ?facility ?instrument ?technique
WHERE {
   ?facilityType rdfs:subClassOf PaNKOS:Facility .
   ?facility       rdf:type ?facilityType ;
   PaNKOS:hasInstrument ?instrument .
   ?instrument PaNKOS:supportsTechnique ?technique .
}
order by ?facility
```

Sample JSON return message (incomplete):

```
{
  - head: {
      - vars: [
            "facility",
            "instrument",
            "technique"
        ]
    },
  - results: {
      - bindings: [
          - {
              - facility: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#ILL"
                },
              - instrument: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#BRISP"
                },
              - technique: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#SmallAngleInelasticScattering"
                }
            },
          - {
              - facility: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#ILL"
                },
              - instrument: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#CYCLOPS"
                },
              - technique: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#Diffraction"
                }
            },
          - {
              - facility: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#ILL"
                },
              - instrument: {
                    type: "uri",
                    value: "http://www.purl.org/pankos#CYCLOPS"
                },
```


Appendix E. OAI-PMH servers considered for PaNData endpoint.

Table 2: Table of OAI-PMH servers

OAI server name	Home	Provider	Comments
jOAI	http://www.dlese.org/oai/	DLESE (Digital Library for Earth System Education) – geoscience community	Java implementation used by EUDAT.

pyOAI	http://infrae.com/download/OAI/pyoai	Python community	Python implementation; a part of MOAI platform https://pypi.python.org/pypi/\MOAI/2.0.0 that can aggregate content from different sources (static files, relational databases) and expose it via OAI-PMH endpoint. Also can be a “pipe” getting data from other OAI-PMH endpoints.
Nesstar	http://www.nesstar.com/software/\oai_pmh_server.html	Norwegian Social Science Data Services	Java implementation popular in social science data archives.
OAICat	http://oclc.org/research/activities/\oaicat.html	OCLC (Online Computer Library Centre)	Java implementation popular in the libraries.

Appendix F. The mapping of PaNData ICAT fields to the OAI-PMH schema and corresponding fields in EUDAT metadata schema.

Table 3: Table of field mappings

	EUDAT Field	ICAT Field(s)	Comment
dc:identifier	-	Investigation->doi	
dc:title	title	Investigation->title	
dc:descritption	notes	Investigation->summary	
dc:relation	tags	Instrument->fullName Investigation->name InvestigationParameter->name (multiple)	Investigation->name + some string co-create RB number; InvestigationParameter->name is usually multiple; Semicolon separated.
dcterms:references	URL	"dx.doi.org/" + Investigation->doi	
dc:creator	author	User->fullName	Multiple.
-	spatial	-	N/A
dc:contributor	maintainer	Science and Technology Facility Council, ISIS	Tom Griffin proposed term similar to "STFC ISIS"; resultant string is a publisher and preferred name for ISIS taken from PaNKOS ontology.
dc:subject	discipline	"Clean energy and the environment, pharmaceuticals and health care, nanotechnology and materials engineering, catalysis and polymers, fundamental studies of materials"	Taken from PaNKOS ontology.

-	PublicationYear	-	Can be extracted from PublicationTimestamp.
dcterms:issued	PublicationTimestamp	Investigation->releaseDate	
dc:language	Language	en	ISO 639-1.
dc:publisher	Origin	Facility->name Facility->fullName Facility->url	Semicolon separated.
dc:format	Format	DatafileFormat->name DatafileFormat->type DatafileFormat->version DatafileFormat->description	Semicolon separated.
dc:relation	GeographicDescription	"ISIS, Harwell, United Kingdom"	This is an example constant.
dc:rights	Rights	ISIS Data Management Policy http://www.isis.stfc.ac.uk/user-office/data-policy11204.html	To be explicitly agreed.
dc:relation	Project	-	Empty right now, ISIS wants to use this field in the future.
dc:relation xsi:type="dcterms:ISO 3166"	Country	GB	GB is an ISO 3166-1 alpha-2 code alternatively: ISO 3166-2:GB
-	GeographicCoverage	-	N/A
dcterms:temporal	TemporalCoverage:BeginDate	Investigation->startDate Investigation->endDate	Timestamps delimited with semicolon
	TemporalCoverage:EndDate		

Appendix G. The harvester, converter, and publisher from ICAT to Qualified Dublin Core in OAI-PMH endpoint.

In order to deploy OAI-PMH to publish ICAT data, the following steps are required:

- deploy the data provider joai, then configure it;
- publish the Metadata namespace definition; this step is generally not required, as it has already been published;
- configure, then execute the collector script.

Detailed instructions on each of these steps are provided in references [1, 2, 3]. The software is provided with preloaded configuration scripts so that the user has little more to do than added the credentials for the ICAT to be queried.
The software has been written so that once configured, the collector runs once per hour, and finds new records in the ICAT and passes them to the provider.
The metadata namespace definition has been published on http://PaNData.org
The mapping of the the ICAT fields to the Metadata model is implemented in the software in the directory called src.
There is an example of OAI-PMH deployed, populated and operating is at the following location: https://icat-10.PaNData.stfc.ac.uk/oai/provider?verb=Identify
[1]https://code.google.com/p/PaNData/source/browse/trunk/contrib/erasmus2014/poramus/pmh/readme.txt Description of how to configure and deploy the OAI-PMH server.
[2]https://code.google.com/p/PaNData/source/browse/trunk/contrib/erasmus2014/poramus/pmh/namespace/readme.txt Description of the deployment of the namespace definition.
[3]https://code.google.com/p/PaNData/source/browse/trunk/contrib/erasmus2014/poramus/pmh/script/properties_description.txt Description of the properties of the collector.