Tools for building research objects in Photon and Neutron Facilities

PaN-data ODI

Deliverable D6.3

Grant Agreement Number	RI-283556
Project Title	PaN-data Open Data Infrastructure
Title of Deliverable	Tools for building research objects in Photon and Neutron Facilities
Deliverable Number	D6.3
Lead Beneficiary	STFC
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	30 September 2013 (Month 24)
Actual Delivery Date	8 November 2013 (Month 26)

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

Abstract

We consider the notion of data publication in the context of large-scale scientific facilities. Dataset publication allows access to and citation of data, but in general does not provide sufficient context. We propose instead to publish an investigation, a more complete record of the experiment, including details of the context and parameters of the experiment. We relate this investigation to the emerging concept of a **research object**, and consider how **investigation research objects** can be constructed to support the more complete publication of facilities science. This provides a mechanism for publishing and sharing research information about a facilities experiment in context, which can be annotated, extended with additional related artifacts recording provenance and made available for reuse. This mechanism also enables facilities data to be made available within the web of Open Linked Data.

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 xxxxxx

.

Revision history

Issue	Author(s)	Date	Description
0.1	Brian Matthews, Vasily Bunakov, Catherine Jones and Shirley Crompton, and Antony Wilson (STFC)	30 Sept. 2013	First Draft Structure
0.2	Brian Matthews, Vasily Bunakov, Catherine Jones and Shirley Crompton, and Antony Wilson (STFC)	6th Nov 2013	Complete draft
0.3	Brian Matthews	8th November 2013	Final draft
1.0	Brian Matthews	14th November 2013	Final version

Table of contents

Page

1 Introduction

In this report, we consider how to represent and publish provenance and contextual information as part of a Data Publication and dissemination process.

Data publication is becoming an increasingly accepted part of the future data ecosystem to support research. This involves enabling public access to data by other researchers, with appropriate guarantees of integrity in the management and persistence of the data, and encouraging researchers to cite the use of the data within publications. The intentions behind data publication include: assigning credit and recognition to the collectors of data; encouraging the inspection of data by peers to assess the quality of the data, and validating the assertions of scientific insights claimed in published articles arising from the analysis of the data; enabling the reuse of the data by other researchers to re-analyse to discover new insights and reportable results, thus furthering the value of the research which arises from the data collection. As a consequence, a number of different approaches and infrastructures have been advocated for data publication (for example [1, 2]).

This is also becoming recognised in the field of “facilities science”. We define facilities science as that science which is undertaken at large-scale scientific facilities, in particular in our case neutron and synchrotron x-ray sources, as represented in the PaN-Data consortium, although similar characteristics can also apply for example to large telescopes, particle physics experiments, environmental monitoring centres and satellite observation platforms. In this type of science, a centrally managed set of specialised and high value scientific instruments is made accessible to a community of users to run experiments which require the particular characteristics of those instruments. The facilities have their own dedicated staff and funding to supply a scientific service.

In this report, we concentrate on neutron and x-ray sources. These types of facilities differ from other “big iron” [2] science projects in that whilst the facility itself has the characteristics of “big science”, including large long term investments, specialised support teams, large quantities of data, high-performance computing analysis requirements, the science itself is more characteristic of “small science” (or bench science), with many small experiments undertaken by small research teams taking readings of many samples, with diverse funding sources and intellectual objectives. This mixture of characteristics has influenced how facilities are approaching data publication.

In particular, the institutional nature of the facilities, with the provision of support infrastructure and staff, has allowed the facilities to support their user communities by systematically providing data acquisition, management, cataloguing and access, thus providing some of the advantages of “big science” to a small science community. This has been successful to date; however, as the expectation of facilities users and funders develop, this approach has its limitations in the support of validation and reuse, and thus we propose to evolve the focus of the support provided.

We propose that instead of focussing on traditional artefacts such as data or publications as the unit of dissemination, we elevate the notion of “investigation” as an aggregation of the artefacts and supporting metadata surrounding a particular experiment on a facility to a first class object of discourse, which can be managed, published and cited in its own right. By providing this aggregate “research object”, we can provide information at the right level to support validation and reuse. In this process, we provide the data in the context in which has it has been collected, and thus we are using the data provenance, in the broad sense who has undertaken the experiment and why,

and how the data has subsequently been processed to provide derived data products and presentations.

In this report, we briefly discuss the current facilities approach to managing and publishing data, We then discuss the limitations of this approach, and introduce the concept of an Investigation as a research object, as the unit of publication and access for facilities data, which provides "data in context", including its provenance and derived data products. We discuss how this may be represented as Linked Data, comparing it with other similar approaches to research object in the literature. We then further consider how this Investigation may be used, and the tools support which would be required to collate, maintain and preserve such a research artefact.

This work has also been reported as it has progressed in publications [5] and [18].

2 Supporting data management and publication

The neutron and synchrotron radiation facilities support a wide range of different experimental techniques (e.g. crystallography, tomography, spectroscopy, small-angle scattering), and experiments are undertaken within a wide range of different disciplines, including chemistry, biochemistry, materials science, earth science, biology, metallurgy, engineering and archaeology. However, from a data management perspective, they all follow similar processes. User scientists apply for an allocation of time on an instrument supported by a science case, which, if accepted, is followed by one or more visits to the facility's site where a number of samples, prepared by the user in advance, are placed in the target area, and then exposed to the beam of particles for a desired period of time. During the exposure the beam particles are then blocked or deflected by the sample and then detected by banks of sensors arranged around the target area. These sensors then generate data on such parameters as particle counts, angle of deflection, time-of-flight of the particle, energy, or frequency. This raw data is then streamed off via data acquisition and data management systems which collect, aggregate and move the data to short or long term storage to await further analysis. We discussed this process in detail in PaN-Data ODI Deliverable 6.1 [8].

Traditionally, this process has been carried out using standard file systems and tools; however, it has been recognised for some time that with the ever increasing data rates and volumes, and increase throughput of experiments, this approach was becoming increasingly hard to manage by hand with the accompanying risk of data loss or corruption. Consequently, we have systematised the process of data management by developing a data catalogue system, ICAT [3], which is being deployed as a reference catalogue across the PaN-Data consortium in PaN-Data ODI, WP4. This cataloguing component, which is based on an information model capturing a view of a facilities experiment or "investigation" (the Core Scientific MetaData (CSMD) model [4]), within a relational database, provides a common point of gathering information about the experiment. This captures information on the experimental team and intent from the proposal system, and when the experimental visit takes place, will register the data sets, their locations in storage, and experimental parameters. This information is then exposed via an API, either for users to use for browsing and data download on or off site via a web interface (the "TopCat" tool), or else integrating with analysis tools and frameworks so that they can search for and access the data directly. This approach has been successful, and ICAT is being both augmented with additional components.

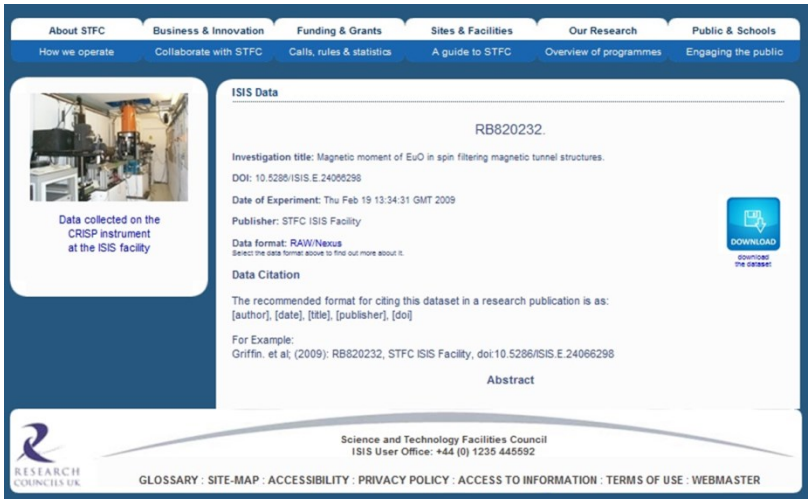


Figure 1: DOI Landing Page for ISIS

The PaN-Data consortium members have recognised the value of releasing data publicly, as reflected in data policies as much is practical [16], data is released for general use after an embargo period of exclusive use to the user. This can be support via the TopCat interface. However, to encourage citation of data and thus attribution and credit for data collection, it is proposed that Digital Object Identifiers (DOIs) for data issued via for example the DataCite consortium[1]; this is discussed in Pan-Data ODI Deliverable 7.1 [17]. Thus for each investigation within for example ISIS or ILL, a DOI is issued, an amount of discovery metadata deposited with the DataCite search engine, and a suitable landing page produced as the "front page" of the data; Figure 1 gives an example of a landing page for the ISIS neutron facility. From this landing page, given suitable permissions for embargoed data, the data can be accessed. ICAT can provide a stable and quality source of metadata, and a route to archival storage. Thus this provides a suitable data publication channel for a facility's data.

2.1 The changing landscape of facilities science

This established process has been successful for data management and the data publication method via DOIs and landing page, whilst still evolving, should provide a mechanism to support basic data discovery, and support citation of data via DOI and a suitable recommended citation format, thus allowing credit to be attributed to experimenters in traditional publications, and following this, allow the facility to via citation tracking to monitor the value of the use of data generated. However, the landscape of facilities science is changing. We summarise some factors [5].

- Instrumentation and data analysis have become more user friendly than in early days of facilities science. This has led to a lesser significance of the instrumentation "gurus" with a current trend of not including them as the authors of papers; the estimate for biology papers is that about half of them do not now include any facility staff members as co-authors [6] so

[1] www.datacite.org

that new methods and forms may be required for the fair and inclusive attribution of research output.

- The advances of instrumentation and Internet have also led to services allowing users to send their samples for remote investigation according to one of the service plans. The sample exposure on a large facility may be just one of the experimental techniques included in the service plan. The service provider then collects the experimental data and supplies them to the user in pre-agreed formats. This implies considering service providers the legitimate agents of facilities science with their inclusion in data management policy.

- Facilities use more than one service to collect data. The user monitoring exercise performed by PaN-data initiative showed that about 7000 (22% of the total) of visitor researchers across Europe have used more than one neutron or synchrotron radiation facility for their investigations[2]. This makes actual the development of common user authentication and user authorization services, as well as experimenting with "virtual laboratories" for the collaborative data analysis.

- New experimental techniques like neutron tomography, or using robots for manipulating multiple samples, or studies of dynamics of materials. The new techniques produce larger volumes of data; they also raise potential opportunities for researchers to perform comparative and multi-aspect studies for the same samples using different experimental techniques, or using the same experimental technique for much wider variety of different samples. This scales up all three V's of Big Data: Volume, Velocity, and Variety, and makes their analysis more demanding from modelling and from computational points of view.

- Publishers and scholarly institutions such as the International Union of Crystallography are increasingly requiring traceability of published results through final result dataset to the raw data collected at the facility instrument, so that peers can test the validity of the claimed result.

Thus there is an increasing need to reuse and combine results from different sources; to provide sufficient detail to reviewers so that they can reconstruct the experiment to validate results; and to provide mechanisms to allow credit for various participants in the experimental process, suitable for their role, as in for example [7]. Much of this needs to be mediated via automated tools, so the record of the experiment needs to be available in a machine readable format. The current data publication mechanism based on DOIs and landing pages does not support this well as the context of the data collection, the relationships between various research artefacts, and the different roles of individuals in the process is not captured adequately, so we need to rethink what data publication means in this context.

3 Investigations as Research Objects

Our starting point is to consider the research lifecycle in facilities science, given in schematic form in Figure 2 and given in more detail in [8]. From the point of view of the *Facility* (the user scientist may have a different view of their scientific process) investigations tend to go through the same

[2] http://wiki.pan-data.eu/CountingUsers

stages of proposals, preparation, experimental visit, data management, data analysis and visualisation, and publication.

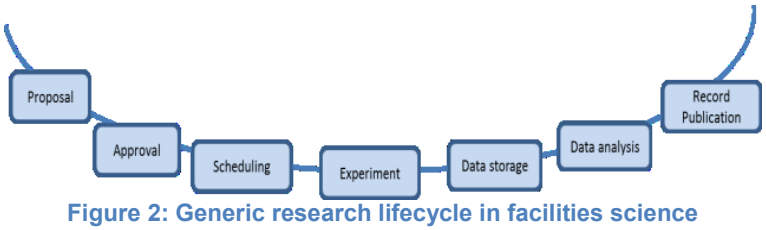


Figure 2: Generic research lifecycle in facilities science

The different stages of research lifecycle produce data artefacts (research proposals, user records, datasets, publications etc.) that are similar across research facilities. Different actors are also involved at the various stages. We also need to record the details of the experiment; which sample was analysed under which experimental conditions, to collect data representing which parameters. Thus by following through the lifecycle of a successful beam time application, we can collect all the artefacts and objects related to it, with their appropriate relationships. As this is strongly related to allocation of the resources of the facility, this is a highly appropriate unit of discourse for the facility; the facility want to record and evaluate the scientific results arising from the allocation is its scarce resources. Thus we propose that the appropriate unit of publication for facilities science is the Investigation.

At one level this is what we already do when we present a landing page for an investigation. Much of the information which is required can be recorded within the ICAT system. It can support describing which sample was used on which instrument to generate which data set under which experimental conditions to measure which parameters. However, the DataCite metadata does not include these, and while some of this information can be found on the landing page (e.g. instrument) and much more can be found by exploring the detailed metadata in TopCat itself, this is human accessible only, not straightforward to find or navigate, and is not distributed in a machine readable form. Further, related artefacts (derived data, publications, provenance information) is not systematically collected or presented, although now ICAT has the capability to collect this information [9, 10]. What we propose to do is publish the investigation as a single aggregated unit which can be identified and delivered to the user in a machine readable format and contain sufficient contextual information to support discovery of all the components of the investigation and their relationships, so they are available for validation and reuse; that is publish the investigation as a *Research Object.*

The notion of Research Objects has been explored in a number of projects in recent years (e.g. [11, 12, 13]), and Research Objects have been defined as:

> *… semantically rich aggregations of resources that bring together data, methods and people in scientific investigations. Their goal is to create a class of artefacts that can encapsulate our digital knowledge and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge*[3]

Research Objects (ROs) as implemented can be seen to have the following characteristics.

[3] http://www.researchobject.org

- Information about research artefacts and their attributes and relationships are represented as Linked Data; thus RDF is used as the underlying model and representation, with URI used to uniquely identify artefacts. As ROs are linked data objects, they can link into to the existing Linked Data cloud to provide additional context information and be managed by the standard tools of Linked Data and the Semantic Web.

- Standard vocabularies are used to represent relationships describing the research process, such as workflow (workflow4ever[4]), provenance (e.g. Prov-O[5]), and citation (e.g. cito [14]). Use of standard vocabularies encourages shared understanding, enables reuse and allows the use of tools which are tailored for their specialised semantics.

- A bound is provided on the object as an aggregation, so we can determine membership of the research object; typically, OAI-ORE[6] is used for this purpose.

- The whole research object can be identified via a URI, so its own history and attributes can be related as a first class research artefact in its own right.

The notion of the boundary of a RO is particularly important. A research artefact can be linked to a number of research artefacts. An investigator or instrument can participate in a number of investigations; a publication may use the output of several investigations to support its results. If this is represented as a simple web of linked data, then it would be difficult to distinguish which artefacts and relationships are members of which research object. We need a notion of defining a boundary to determine membership of the RO; OAI-ORE, with its notions of Aggregation and Resource Map provides such a boundary. Research Objects are thus highly suitable as a mechanism to represent and publish Investigations.

4 Building an Investigation Research Object

We outline the major steps of building a research object to represent facility's investigations.

4.1 Representing CSMD in RDF

We can represent the CSMD as an OWL ontology, as discussed in PaN-Data ODI Deliverable 6.2 [19]. This will allow us to represent metadata as RDF triples within triple stores (or provide a triple based front end onto metadata databases such as ICAT via for example a SPARQL endpoint) and allows us to publish data about investigations into Linked Open Data. Figure 3 gives a sample of the OWL representation; the full model can be found on the ICAT Google Code site[7]. The OWL representation has a base URI: `http://www.purl.org/net/CSMD/4.0#`

```
<owl:Class rdf:about="csmd:Investigation">
    <rdfs:label>Investigation</rdfs:label>
    <rdfs:comment>An investigation or experiment</rdfs:comment>
</owl:Class>
```

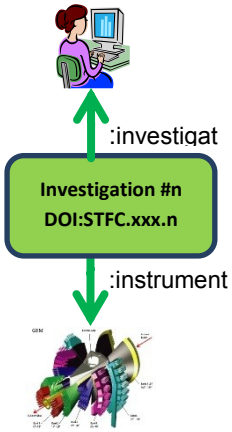
[4] http://www.wf4ever-project.org/
[5] http://www.w3.org/TR/prov-o/
[6] http://www.openarchives.org/ore/
[7] https://code.google.com/p/icatproject/

```
    <owl:Class rdf:about="csmd:Facility">
        <rdfs:label>Facility</rdfs:label>
        <rdfs:comment>An experimental facilty</rdfs:comment>
    </owl:Class>

    <owl:Class rdf:about="csmd:Dataset">
        <rdfs:label>Dataset</rdfs:label>
        <rdfs:comment>A collection of data files and part of an investiga-
tion</rdfs:comment>
    </owl:Class>

    <owl:Class rdf:about="csmd:Datafile">
        <rdfs:label>Datafile</rdfs:label>
        <rdfs:comment>A data file</rdfs:comment>
    </owl:Class>
```

Figure 3: A fragment of the CSMD Ontology

4.2 Constructing an investigation research object

As the facilities lifecycle is enacted within an experiment, we can then construct the research object. Thus, immediately after an investigation has been approved, we can initialise the research object, assigning a DOI at this early stage, and providing some basic information from the proposal, such as instrument used and investigator, as in Figure 4, which also includes a prototypical fragment in RDF-Turtle of the investigation object at this stage.




```
:n a  csmd:Investigation ;
   csmd:investigation_doi  doi:stfc.xxx.n

csmd:investigation_investigationUser :iu1;
  csmd:investigation_instrument :inst1 .

:iu1  a csmd:investigationUser ;
    csmd:investigationUser_user  :u1 .

:u1 a csmd:User .
:inst1 a csmd:Instrument  .
```

Figure 4: Initialising the Investigation Object

As the experiment in undertaken, we can add further information to the investigation object, to build a more complete picture of the collection of raw data on a sample, again as in a simplified view in the figure below. This step captures the information presented on the current DOI landing page.

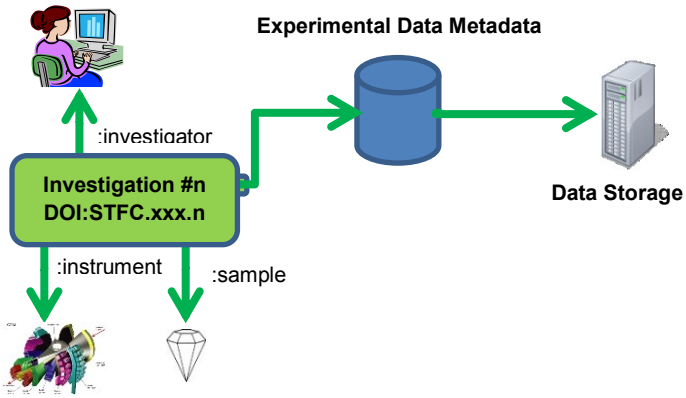


Figure 5: Investigation Object after the Experiment

As the experimental lifecycle goes on, as for example analysis of the data through software packages, and publications and other auxiliary content is added to the investigation, together with the parameters and configurations used, and provenance information collected, we can continue to add to the Investigation object, building an eventual object which may contain references to objects in different repositories, ownerships and locations, brought together in a single linked structure as in Figure 6. This describes a linked data structure which include references to derived data products, software packages used to generate derived products, and publications. These use the CSMD vocabulary, in combination with the CITO ontology for representing citations [14].

Thus this provides a complete picture of the full investigation. This is a dynamic object; further entities could be added it, further derived datasets, publications, or annotations for example as further reuse is undertaken of the research object.

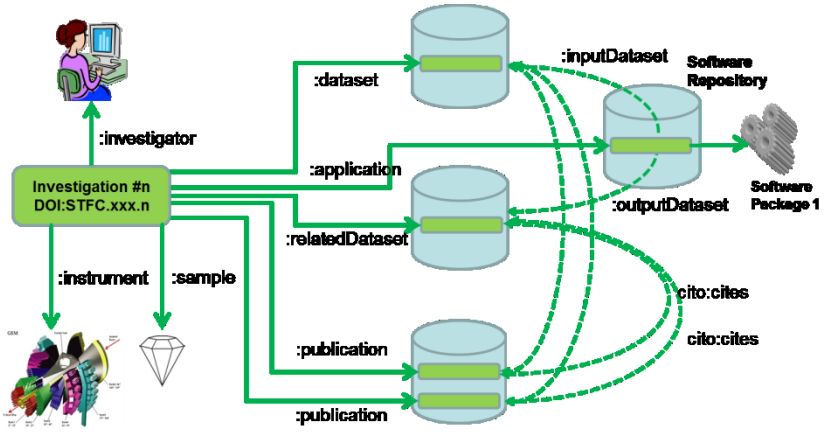


Figure 6: Investigation Object after a complete lifecycle

Thus an investigation research object can be constructed in to aggregation. However, the research resources within the linked data graph can also be connected to other objects. For example, a publication could use data from several investigations. The publication should be included in each investigation object, but any particular investigation should not include fully the other investigations. Thus we need to provide a boundary. As mentioned above, other approaches have used

OAI-ORE to provide a boundary of what is included within the research object, and we propose to follow a similar approach.

4.3 OAI-ORE Basics:

Dissemination of the linked-data instances of the provenance records is done using the OAI-ORE technology. The OAI-ORE defines standards for the description and exchange of aggregations of Web-based resources in a linked-data compliant way. The key OAI-ORE concepts are: Aggregation (A) - a set of Web-based Resources; Aggregated Resource (AR) - a Resource that constitutes (together with other resources) an Aggregation; and Resource Map (ReM) - a brief description of an Aggregation.

So, as illustrated in Figure 7, the provenance record within an IRO would be encapsulated within an OAI-ORE Aggregation as an Aggregated Resource. In order to publish the record, we assign a DOI to the corresponding OAI-ORE Aggregation (identified by an OAI-ORE Aggregation URI). So, when the DOI is de-referenced, the client is redirected (using HTTP 303 re-direct as recommended by the linked-data principles) from the Aggregation URI to the URI of the Resource Map that describes the Aggregation.

The Resource Map serves as a landing or splash page providing a description of the Aggregation (not Aggregated Resource), which includes the URI for the Aggregated Resource (e.g. a provenance record). The client is then able to de-reference the URI for the Aggregated Resource to retrieve it. It is important that the contents and format of the Aggregated Resource remain static for an indefinite period of time in order to adhere to the DOI rules.

The Aggregation description contained within a Resource Map may also include information about other static or non-static resources related to the Aggregated Resource using an appropriate vocabulary. In effect, this enables the provider of a workflow instance to be able to seamlessly link to other related resources that he or she may not have control over – one of the principle advantages of linked-data. In addition, a Resource Map may be provided in multiple formats (e.g. HTML, RDF, XML) based on the client's request.

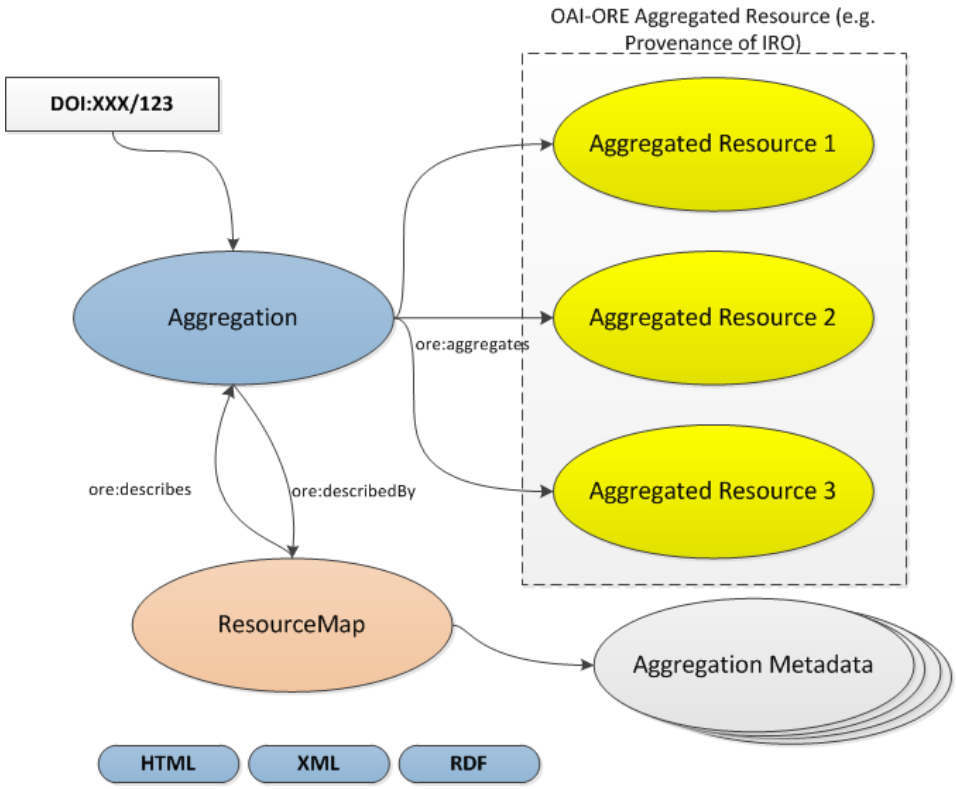


Figure 7: An OAI-ORE representation of linked provenance records within an IRO

4.4 Using OAI-ORE as an aggregation constructor

OAI-ORE provides some core constructs for capturing aggregations. The class `ore:Aggregation` provides an abstract concept for aggregating resources (`ore:AggregatedResources` in OAI-ORE), with an object property `ore:aggregates` as the combining mechanism. `ore:ResourceMap` describes the aggregation, the resources and the relationships between them. Thus to represent an Investigation Research Object, which is an aggregation, we declare that the Investigation class is a subclass of `ore:Aggregation`:

```
csmd:Investigation  rdfs:subClassOf ore:Aggregation .
```

This follows the approach of the Core Research Object Model[8], and thus we can also declare:

```
csmd:Investigation  rdfs:subClassOf ro:ResearchObject .
```

Further, we can declare the core relationships between Investigations and other resources in the CSMD using sub-properties:

```
csmd:investigation_dataset rdfs:subPropertyOf ore:aggregates .
```

We can thus use OAI-ORE to construct the investigation research object with minimal changes to our information model.

[8] http://www.researchobject.org/ontologies/

5 General Architecture and Management for IROs

The concept of adding additional context and linking to related items into an Investigation Research Objects (IROs) is not limited to this specific domain, and can be generalised to any activity where an investigation or experiment is carried out within a large scale facility context.

5.1 Requirements

- Should be able to use existing facility infrastructure for the capture of the investigation and associated experimental data
- Should support the publication and discovery of IRO. The metadata should be able to be discovered in both human readable and machine readable formats.
- Should allow the user to input additional supplementary material, such as publications and analysed data and accurately combine it with the current Investigation Research Object (IRO).
- Should be checkable to ensure both the persistence of links and the persistence and integrity of the additional context.
- Should be able to search for additional links to augment the IRO with additional supplementary information
- Should allow for electronic navigation between different linked outputs for this to be successful over time, then the objects being connected need to have persistent identifiers so that there is some assurance that the links will also persist.

5.2 Management of IROs

As with any object, there are a series of processes to create, manage and preserve the research object, these of course will be guided by the policy in place within the organisation. As it is a complex object, there are additional considerations both from a management and policy viewpoint.

5.2.1 Creation and population of IROs

A research object which describes an investigation can have many links to other objects, and is intended to be built over time – the publication will come many months if not years after the experiment was run – we do not want to wait for this to occur before building the object. This means the valid minimum set of elements will either be a low barrier, or needs to change over time. In the architecture, it is anticipated that the core of the research object will be built from a facilities existing infrastructure that manages research data.

The type of information captured in a facilities science research object is subject to change. There is the issue of new, replacement versions of the objects being referenced through improvements in analytical techniques, where only the latest version needs to be shown, compared against the state where the version of the object as it was referenced is the important link. Depending on others to provide the persistent identifier and object and their choice of how this persistent mechanism is implemented will affect the make-up of the research object. An obvious change to be managed is if the data is migrated to another format – what is the effect on the investigation research object?

5.2.2 Validation and Verification

If the research object is to be used, re-used and preserved then there needs to be notions of what a well-formed research object is and what the significant properties of that object which need to be identified and preserved are.

Taking the concept of well-formed as a starting point, we have already noted that the research object will increase in richness of content over time and therefore the object at creation may have a small set of required links. The presence of these links at creation would be relatively straightforward to ascertain. It is a more difficult proposition to identify at what point additional resources should be considered to be missing, rather than not yet created.

As the research object is an OAI-ORE aggregation, the organisation publishing the aggregation may not own the resources linked to, and so if they become unusable, there is the issue of what are the mandatory links which make it a feasible intellectual entity and are there any which if they disappeared would mean that the research object was no longer valid.

One of the issues in preservation is rights – does the content holding institution have the rights to modify the object to enable successful preservation. In the case of a research object, one of the links may be to another content holding institution and what may happen if the other content holding institution changes and chooses to no longer hold the item that is of importance to the research object is an open question.

In addition to these issues, there is an additional issue to ensure that the model and object generated by that model, themselves conform to the OAI-ORE specification and that the ontologies used are known and retrievable.

5.2.3 Use and access

There needs to be the ability to find and identify research objects of use to the designated community and for the information to be displayed effectively.

5.3 Architecture and components

Figure 8 below shows a generic architecture for investigation based research objects (IRO). The items in grey would form part of local infrastructure which provides the investigation object and associated context which is available at the initial creation phase. The items in green are proposed tools that would be required; these would be as generic as possible. The items in blue provide tools to enable the metadata to be used by others, to display the IROs and to enable others to add supplementary data to the IRO; these would be dependent on the local infrastructure and requirements.

Relating these tools to the description of management processes above; the IRO Builder would create the research object and the Supplementary Data, RO annotator and Link searcher would provide functionality to enable additional context and links to be added. The RO validator would ensure that the IRO was not only correctly formed, but would also look for changes over time and finally the user interface, through the data journal concept, would enable end users to find and use these aggregations of content.

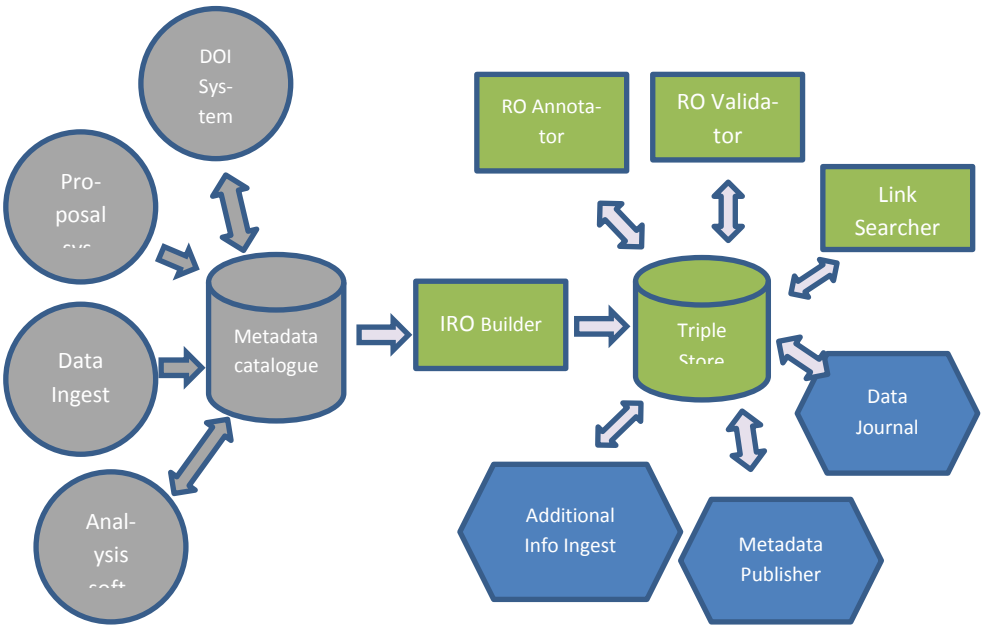


Figure 8: General architecture for investigation research object

STFC are undertaking an initial implementation of Investigation Research Objects. This involves extracting the reference investigation data from the ISIS ICAT data catalogue and building a representation of the IROs in a suitable triple store. It is anticipated that this could form the basis of the data publication and annotation system. Thus derived data products could be added to IROs for example, and used to generate landing pages. One concept which is being prototyped in a notion of a data journal for a facility, as a formal record of the experimental output of a facility.

6 Using Investigation Research Objects

The use of IROs can enable a number of potential uses and benefits.

6.1 Provide management for research artefacts beyond the experiment

Using IROs allows the facility e-infrastructure to control the creation of the aggregation, and thus the resource map describing the Aggregation will be an authoritative collection of related research artefacts. . We identify the aggregated resources using its own URI in to the Aggregation (locking in the context). When other Resources reference the individual Aggregated Resources or the Aggregation itself, it uses the IROs URIs. This helps to direct web traffic to the facility hosted services, as the Resource Map facilities discovery and enable (point and click) access to the Aggregated Resources on or via our server.

Further, exposing an Investigation DOI (Aggregation) as a compound research object allows us to reference all the related research artefacts using the same DOI. There is no need to register separate DOIs for its Dataset or Datafile as these objects can be referenced as an Aggregated Resource w/n the Investigation DOI (Aggregation).

Consistent presentation of the research artefacts can be support, as the facility can use content negotiation to match what to return for a request, thus supporting multiple views and formats

6.2 Facilitate eScholarship:

The use of IROs provides a context to bind the distributed research artefacts together, with all Aggregated Resources are linked within the context of an Investigation, and unambiguously define the relationship between the resources within the Investigation. This can aggregate heterogeneous resources on the web independent of repository containment, including provenance info on the Aggregated Resources. This allows the traceability of research results to be followed, in both a human and machine interpretable manner, providing a basis for validation of results.

Further, by publishing the IROs on the web, improve discovery as can be enabled as it can consumed by web crawlers, data mining applications, and be made available as part of the Linked Open Data network. The rich semantics and metadata should facilitate interpretation and re-use, while the use of ORE means that the Investigation and its related research artefacts is packaged under One URI, and thus the compound object can be referenced and annotated directly, to facilitate exchange. Usage and citation statistics can be gathered on an Investigation Object level, so we can gather more accurate statistics of use in context.

6.3 Supporting multiple viewpoints

Regardless of discipline there is an acknowledged “life cycle” of research, which is realised in many ways depending on the audience and purposes; a researcher, a funder, a research organisation, a publisher or a preservation institution will focus on different aspects of this life cycle and bring additional contextual links relating to their business process and requirements. For these different stakeholders the central object to which context is added will be different as their world viewpoint is different, for example a publisher will want to establish links from the publication; a funder may wish to do the same for grants. We have described building the links to the investigation from our viewpoint as a facility which is responsible for the creation, discovery and curation of the investigation undertaken at the facility.

The use of research objects supports well this notion of different points of view. Publishing data within a linked open data context in particular makes notions of what constitutes a coherent viewpoint of relevant resources and relationships hard to capture. By providing boundaries and criteria for membership, research objects can support multiple points of view within one data infrastructure. Thus different stakeholders can construct, use and reuse the context relevant to them, and also be credited to the portion of the object which is appropriate to their contribution.

6.4 Data publication

We would propose to use investigation research objects as the unit of publication for our facilities data. Thus we would identify investigation and their related resource maps by persistent identifiers, and use them to generate a landing page. This would be extensible to provide access to the research object in its entirety and include related entities to provide more information in context which could be accessed by other automated agents. Metadata associated with the DOI would need to be changed. Currently, the Datacite metadata field ResourceType supports Dataset and

Collection (amongst others), neither of which is correct[9] in this context. We would propose that the list of allowed values for this field is extended to include the notion of experiment, study or investigation.

Using the notion of research object as a more open ended bounded object raises the notion of what exactly is being published persistently in this case. If we add additional information are we maintaining stability? Research Objects are well suited to notions of versioning, where we can relate objects together as they change, thus keeping the old boundary stable. Further, we would propose to have different levels of assurance in our case. The core information on the experiment (sample, instrument, parameters, raw dataset) would remain constant, with other information being secondary and subject to possible extension; this would made clear in the presentation.

6.5 Data preservation

Shifting the focus from the data to the investigation makes the data preservation activity a more complex one, as it moves from activities relating to the preservation of a well-defined digital object to include not only the digital object but also activities to ensure that the complex linked data, OAI-ORE resource map maintains it integrity and meaning, and links still point to resolvable objects. For preservation purposes it is important that these links are permanent to ensure the integrity of the object.

7 Summary

The work presented in this report represents a work in progress. Further discussions are required to agree the correct representation of Investigations as research objects, and design and implementation work to provide tools support so that investigation research objects can be constructed, maintained and published as linked data. However, we see that this could form the basis of a data publication route for facilities data via enhanced landing pages.

ACKNOWLEDGEMENTS

This work is joint work between the teams of PaN-Data work-package 6, and the SCAPE project (Scalable Preservation Environments, EC Grant agreement no: 270137) at STFC. In this work, the PaN-Data WP6 concentrated on the aspects of Provenance, and SCAPE considered further aspects related to Preservation (SCAPE).

REFERENCES

1. Lawrence B., Jones C., Matthews B., Pepler S., Callaghan S. Citation and peer review of data: Moving towards formal data publication. Int. Journal of Digital Curation, 6(2) (2011)
2. Parsons, M.A., Fox, P.A. Is Data Publication the Right Metaphor? Data Science Journal Vol. 12 (2013)

[9] http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

3. Flannery, D., et.al. ICAT: Integrating Data Infrastructure for Facilities Based Science. e-Science: Fifth IEEE International Conference on e-Science (2009)
4. Matthews, B., et. al. Using a Core Scientific Metadata Model in Large-Scale Facilities. 5th International Digital Curation Conference, London, UK, (2009)
5. Bunakov, V., Matthews, B. Data curation framework for facilities science. Data 2013 2nd International Conference on Data Management Technologies and Applications (2013)
6. Mesot, J. A need to rethink the business model of user labs? Neutron News, 23 (4), (2012)
7. Marcos, E. et al. Author order: what science can learn from the arts. Communications of the ACM, , 55(9),39-41 (2012)
8. Matthews, B. et al., Model of the data continuum in Photon and Neutron Facilities. PaNdata ODI, Deliverable D6.1. (2012). http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D6.1.pdf
9. Yang, E. Matthews, B., Wilson, M. Enhancing the Core Scientific Metadata Model to Incorporate Derived Data. Future Generation Computer Systems 29 (2) 612-623 (2013)
10. Fisher, S.M., Phipps, K., Rolfe, D. ICAT Job Portal: a generic job submission system built on a scientific data catalogue. 5th International Workshop on Science Gateways, (2013).
11. Bechhofer, S. et al. Why linked data is not enough for scientists. Future Generation Computer Systems, 2013, 29(2),599-611 (2013)
12. Shaon, A., Callaghan, S., Lawrence, B., Matthews, B., Osborn, T., Harpham, C. Opening up Climate Research : a linked data approach to publishing data provenance. 7th International Digital Curation Conference (DCC11), Bristol, England, (2011)
13. Belhajjame K, et. al. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Heraklion, Greece, (2012)
14. Shotton, D. CiTO, the Citation Typing Ontology J Biomed Semantics. 1(Suppl 1): S6. (2010) . doi: 10.1186/2041-1480-1-S1-S6
15. Wilson, M.. Meeting a scientific facility provider's duty to maximise the value of data. In DataCite Summer Meeting, Digital Research Data in Practice (DataCite2012), Copenhagen, Denmark. (2012) http://epubs.stfc.ac.uk/work-details?w=62852
16. Common policy framework on scientific data. PaNdata Europe, Deliverable D2.1, (2011) http://wiki.pan-data.eu/imagesGHD/0/08/PaN-data-D2-1.pdf .
17. Implementation of persistent identifiers for PaNdata datasets. PaN-Data ODI, Deliverable D7.1 (2013) http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D7-1-final.pdf
18. Matthews, B., Bunakov, V., Jones, C. and Crompton, S. Investigations as research objects within facilities science. First Workshop on Linking and Contextualizing Publications and Datasets, Malta, September 26th, 2013 (2013)
19. Common ontology definition and definition of tools to support the use of provenance for Photon and Neutron Facilities. PaN-data ODI Deliverable D6.2 (2013)