Sharing research facilities data in common data infrastructures

© Vasily Bunakov © Alistair Mills

Science and Technology Facilities Council,
Harwell, United Kingdom
vasily.bunakov@stfc.ac.uk , alistair.mills@btinternet.com

© Piotr Oramus

AGH University of Science and Technology,
Kraków, Poland
oramus@student.agh.edu.pl

Abstract

The work describes the collaboration between a large experimental research facility and emerging national and cross-national data infrastructures, with the purpose of sharing experimental data and making it findable in common multi-disciplinary data catalogues.

1 Introduction

Many of the major centres of scientific research provide both the instruments for the research, and the infrastructure for storing and processing data. This is typical for large research facilities like synchrotrons, neutron sources, powerful lasers that grant timeslots to visitor scientists for their specific investigations and provide infrastructure for data collection and preservation. Generally, scientists work on the science and facility IT engineers work with the data; this leads to a requirement that these two groups collaborate. Another requirement for collaboration comes from the emerging e-infrastructures that transcend institutional and national borders and research disciplines.

Although research facilities make the data available, they do not provide a large range of access methods. The purpose of our work was to provide an industry standard protocol for accessing the data so that a large number of researchers can find the records about datasets produced by research facilities and access them easily.

New routes to existing data and metadata are important as in the last decade the number of data sources in Europe has increased enormously. It is no longer viable for most researchers to track all of the data which are relevant to their investigations, so data discovery services provided by a cross-discipline infrastructure are essential. Our work is an example of a productive collaboration between a discipline-specific data centre – ISIS neutron and muon facility [3] that is a part of a wider landscape of similar neutron and photon facilities in Europe [9] – and EUDAT e-infrastructure [1] using popular metadata standards and protocols.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Use case description

EUDAT has developed several services, namely:

- B2SHARE – a data publishing service;
- B2SAFE – a secure and reliable replication service;
- B2FIND – a data discovery service (data catalogue);
- B2STAGE – a data delivery service for the rapid delivery of large volumes of data towards high-performance computing;
- B2ACCESS – user authentication service used by some of the above services.

EUDAT services are deployed centrally by project participation organizations with free registration and access for researchers, or the services can be deployed by interested parties in their own environment as all the software in support of these services is open source. We have focused on using the centrally deployed instance of EUDAT B2FIND [8] which consumes records delivered by data providers using OAI-PMH [2], maps them to its own metadata schema, and publishes them in a common data catalogue. The OAI-PMH specification is straightforward and allows the use of different metadata schemas; however, within a single metadata schema, quite different interpretations of metadata elements are possible; EUDAT always negotiates the meanings of metadata elements with the data provider.

The data provider in our case is the ISIS neutron and muon source [3] that collects data during scientific investigations, and that catalogues the data using the ICAT software platform [4]. ISIS has a data management policy [7] that provides public access to most of its publicly funded data at the end of an embargo period of three years. The ISIS policy requires that users of the data register with ISIS, and ISIS records their activity. Registration is free, but the management of ISIS wants to be aware of the use of its data when assessing the impact of the facility.

The work of providing ISIS data in EUDAT involved the following steps:

- evaluation of the available technology;

- building the metadata harvester;
- mapping the domain-specific metadata to a more popular schema;
- mapping the data provided by the service end point to the requirements of B2FIND;
- provision of a service end point for publishing metadata;
- liaison with EUDAT B2FIND for testing the end point and harvesting the data records.

There were two main challenges to address during implementation. The first challenge was the mapping of the metadata: from ISIS to OAI, then from OAI to B2FIND. The second challenge was to avoid compromising the data policy set by ISIS.

The first challenge was technical and required careful programming as well as discussions with specialists knowledgeable of the metadata models for both the data provider and the data consumer.

The second challenge required access to the data records so that the harvester could collect them. In order to get this access, ISIS provides suitable credentials, and it was decided to restrict harvesting to the data records with persistent identifiers in DataCite [10], as this implies that the records are not withheld by ISIS under its data embargo policy.

3 Technology stack and metadata mapping

We chose the Qualified Dublin Core (QDC) metadata schema [6] to represent the data from ISIS. This schema is well known, has a large user base and is one of the schemas recognized by the EUDAT B2FIND metadata mapping interface. The data from ISIS is well structured but it is in a schema that is not supported by the EUDAT B2FIND. The main purpose of B2FIND is data discovery rather than the harmonization of metadata schemas. Table 1 presents the mapping from ICAT metadata schema to QDC and to EUDAT B2FIND schema. This mapping is essential for the semantics of the ISIS data records once they are harvested by EUDAT.

We then developed software that harvests the data records from the ISIS data catalogue, maps them to the QDC schema and passes them to the OAI-PMH server that implements a popular standard for automatic data harvesting [2] required by EUDAT B2FIND ingest mechanism. We considered several implementations of OAI-PMH, and chose a Java implementation called jOAI [5] as it is mature, well documented and widely used. The data records acquisition component is a Python wrapper to ISIS ICAT API.

The resultant technology stack is presented by Figure 1. The bottom layer is a domain-specific data catalogue supported by the research facility (ISIS); the top layer is a multidisciplinary data catalogue supported by a common data infrastructure (EUDAT); the middle layers are components that enable a transformation from a domain-specific implementation to a common data discovery service.

We have stored the software which was developed in this project in a public repository, so that others can examine it for the details [12]. The software is modest in size, and can be easily deployed on a small computer. The computer has to execute a script once per hour to find new data, and it has to run a jOAI server continuously.

Table 1 Mapping from ICAT metadata to Dublin Core and EUDAT B2FIND

ICAT field	QDC term	B2FIND field
Investigation ->doi	dc:identifier	-
Investigation ->title	dc:title	title
Investigation ->summary	dc:description	notes
Instrument ->fullName Investigation ->name InvestigationParameter->name (multiple)	dc:relation	tags
"dx.doi.org/" + Investigation->doi	dcterms:references	URL
User->fullName	dc:creator	author
-	-	spatial
Name of the organization (as a literal)	dc:contributor	maintainer
Description of a facility (as a literal)	dc:subject	discipline
-	-	PublicationYear
Investigation->releaseDate	dcterms:issued	PublicationTimestamp
en	dc:language	Language
Facility->name Facility ->fullName Facility->url	dc:publisher	Origin
DatafileFormat ->name DatafileFormat ->type DatafileFormat ->version DatafileFormat ->description	dc:format	Format
Facility title (as a literal)	dc:relation	GeographicDescription
Web link (URL) to ISIS Data Management Policy	dc:rights	Rights
-	dc:relation	Project
Country code (as a literal)	dc:relation xsi:type= "dcterms:ISO3166"	Country
-	-	GeographicCoverage
Investigation ->startDate	dcterms:temporal	TemporalCoverage: BeginDate
Investigation ->endDate		TemporalCoverage: EndDate

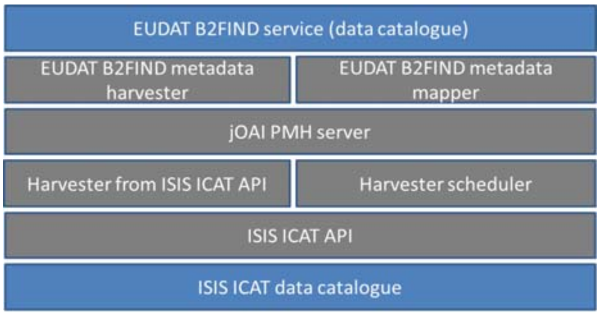

Figure 1 Technology stack for the facility-specific data discovery service

For the published information to be visible, it is necessary to register the jOAI server with a discovery service, such as B2FIND. The operation of the discovery service is the responsibility of a third party such as EUDAT.

The essential flow of work of the software is the following:

- Once per hour, the software connects to the ICAT and requests details of any new records to publish. A suitable record has a Digital Object Identifier and a Release Date since the last time the software was run;
- For each record identified, the software serializes the record as a QDC object and passes it to the jOAI publisher;
- Once per hour, the jOAI publisher checks for new objects and publishes them.

In this way, new records created by the data owner, are generally available within two hours, with no manual processing. No changes, other than configuration, are required to the ICAT server, the jOAI server or the discovery service. For the owner of the data, the additional processing required to provide this service is negligible. For the owner of the discovery service, the additional processing is negligible.

4 Data discovery use case

The services that we have developed in course of this work support the following data discovery use case. In order to find data, the researcher uses a Google-style free string search in the B2FIND data catalogue [8], and locates candidate datasets of interest. This is similar to using any search engine, except that B2FIND is likely to be more relevant as it has a harvesting policy which ensures that it searches a known set of sources; many of the sources known to B2FIND are of little general interest, and are not harvested by general purpose search engines.

Having received search results, the user selects one of the candidates located by B2FIND. B2FIND presents more information about the chosen candidate. In the case of an ISIS record, this information includes the DOI assigned to the dataset by the DataCite service [10]. The DOI link references a web landing page supplied by the ISIS facility; the landing page contains an actionable link that allows the user to get the data collected during the experiment, with the user access to the actual experimental data regulated by a facility data management policy – which in the case of ISIS is a liberal policy which encourages research data reuse [7].

Apart from its usage in EUDAT B2FIND, the OAI-PMH endpoint for ISIS ICAT and the appropriate metadata mapping are being tested for the new Research Data Discovery Service (RDDS) which is a national UK initiative similar to EUDAT B2FIND but with a different scope of research data records collected [11]. RDDS is going to become another public channel for the dissemination of experimental data collected by the ISIS facility, along with EUDAT, DataCite and research papers that cite data DOIs. Figure 2 represents the flow of data records and data persistent identifiers between different services of a common data discovery ecosystem.

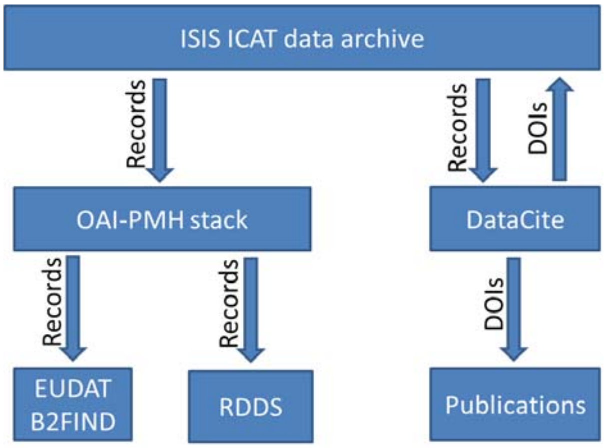

Figure 2 Data records and data DOIs flow

After a period of testing with a few harvesting e-infrastructures, the OAI-PMH stack has the potential to become part of the ICAT software distribution [4] that is used by other neutron and photon facilities in Europe. This should make it easier for other facilities to supply their data records to data discovery portals. It was not possible during the course of the project described in this paper to assess the impact of this work on the various stakeholders. However, the existence of projects such as EUDAT and RDDS and their active collaboration with this project supports our belief in the need for such projects. As we continue to work in this area, we will learn more about the needs of the stakeholders, and change our implementation to support those needs.

Conclusion

We considered the effort to implement the OAI-PMH endpoint and supply data records in e-infrastructures worthwhile for the following reasons:

- large research facilities such as ISIS have an interest in sharing data; it may be a legal or policy requirement that they publish this data, especially data that is collected in a publicly funded investigation; many investigators consider that the provision of data enhances the value of their research and consider that data citation is as valuable as publication citation, hence more routes to citable data are beneficial for researchers;

- sharing data in multi-disciplinary catalogues like B2FIND and RDDS attracts new collaborators, facilitates data reuse within a discipline, and encourages cross-discipline research;
- we are working within a community of European facilities which are adopting common standards for software and infrastructure [9]; the software developed in the course of this work and shared in GitHub [12] provides added value in the technology stack already adopted by similar research centres, which makes our solution organizationally scalable;
- other e-infrastructures can use the ISIS ICAT OAI-PMH endpoint that is now running as beta-service [13], to harvest data records for ISIS investigations with actionable links to publicly available data; metadata cross-walks need to be defined between the OAI-PMH metadata and the e-infrastructure metadata; this is similar to EUDAT, and aims to avoid semantic misinterpretation of metadata elements.

This work provides foundation IT-components and from an organizational point of view, may serve as a model for sharing data collected by large research facilities in common cross-disciplinary data infrastructures. The work is a contribution to the emerging European research data ecosystem comprising traditional research centres, common national and transnational e-infrastructures, research teams located in smaller labs in universities and industry, as well as individual researchers willing to share data. The work aims to increase the efficacy and efficiency of using the public funds allocated for research and development, by providing new routes for data publishing and data reuse.

Acknowledgements

This work is supported in part by Horizon 2020 EUDAT and the UK JISC RDDS projects, although the views expressed are the views of the authors and not necessarily of the projects.

References

[1] EUDAT: the collaborative Pan-European data infrastructure. http://www.eudat.eu

[2] Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh

[3] ISIS neutron and muon research facility. http://www.isis.stfc.ac.uk

[4] ICAT project. http://icatproject.org

[5] jOAI. http://www.dlese.org/oai

[6] DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms

[7] ISIS data policy. http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

[8] EUDAT B2FIND service. http://b2find.eudat.eu

[9] PaNdata initiative. http://pan-data.eu

[10] DataCite service. http://www.datacite.org

[11] UK Research Data Discovery Service. https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

[12] PMH component in ICAT GitHub repository https://github.com/icatproject-contrib/pmh

[13] ISIS ICAT OAI-PHM endpoint (beta-service). http://oai.eudat.stfc.ac.uk/oai/provider?verb=Identify