RESEARCH METADATA MANAGEMENT AT THE AUSTRALIAN SYNCHROTRON

Richard. Farnsworth, Alistair Grant, Andrew Rhyder, Australian Synchrotron, Melbourne Australia
Nick Hauser, Bragg Institute ANSTO, Sydney Australia.

Abstract

This paper details the approach the Australian Synchrotron [1] is using, in collaboration with the Australian Neutron Source, run by the Bragg institute, part of ANSTO [2] (Australian Nuclear Science Technology Organisation) called OPAL (Open Pool Australian Light-water Reactor) for some of the data and metadata management issues. It explores the data and user policies, describes the quantity and quality of data and demonstrates the way forward based on both existing and future directions in e-research, network communications, user proposal and material databases, portal technologies and integration techniques. The role of standards for access and metadata creation is also explored. This work is funded by an educational infrastructure grant administered by Australian National Data Services.

DATA POLICY

In order to progress with publicly funded research facilities data and metadata publishing the data policies must be clear. The answer to the questions who owns the data, when can you make it public, what can you do with it should be clear. At the Australian Synchrotron twenty-four months is allowed to the principle investigator to publish publicly. The period is thirty-six for Bragg institute instruments (ANSTO, OPAL). Either facility may choose to process the raw data in order to make is accessible or publishable. There is a growing trend worldwide towards open technical data. There is also a growing trend towards publishing not only a scientific paper, but also the raw data that was used to produce it.. As both the Australian Synchrotron and the Bragg institute are publicly funded, technical data created at either location should be at some point made available to the public. Currently there are some Australian mechanisms for achieving this. One is called the ANDS portal. [3] ANDS stands for the Australian National Data Service.

DATA QUANTITY

The Australian Synchrotron operates nine beamlines producing around two to three Terabytes of experimental data per day across a wide variety of disciplines from protein crystallography, medical, through to the conservation and restoration of cultural objects and works of art. In 2009 over five hundred groups conducted research at the Australian Synchrotron. More are expected this year and the next. The Australian Synchrotron expects to be producing at least eight terabytes per day when the next round of ten beamlines are installed in the coming two to five years. Even if the Australian Synchrotron just keep operating the existing beamlines, there will be a significant increase in data collection because of both the continual improvement in detectors and the overall efficiency or "duty cycle" of the beamlines. The objective of this project to make that data available publicly. The data will be stored in the curated archives immediately; however authorisation for access will be allowed or otherwise depending on when the data can be made available. Much smaller volumes of data are created at the Bragg institute.

This project is seeking to provide services so that researchers and institutions can manage their data. To give them the power of something like "Google" over their data – that is the ability to search, catalogue and access. This promotes the use and re-use of data and so adds to the efficiency of the data generating ability of each facility.

MECAT

The chosen a name for this project is "MeCAT" [4], as a nod to a similar project/product called ICAT. It was a requirement to name the project, rather than the technology being used. The project is to enhance the technology to enable those things aforementioned efficiencies.

COLLABORATION

It is worthwhile noting the collaboration details. The two facilities have decided that if they collaborate and pooled resources between the two similar facilities in Australia, we could effectively get twice the efficiency of the software development dollar in terms of software resources.

It also a major step towards the creating an Australian culture of same software in similar institutions. This leads to the same experiences for researchers. This is becoming increasingly strategically important to both facilities. It would be ideal if every institute used exactly the same software everywhere such that experimenters trained in the use of software one area or instrument could use the same skills in another. This is probably never going to be completely possible, but this project assists by moving toward that philosophical direction. It also helps with the data management, because the software automatically moves the data for those researchers that come from known institutions to their home institutions.

OBJECTIVE OF PROJECT

The objective of this project is to provide services to researchers to manage their experimental data and to provide data search and access to the broader research

community. These services will enable better use and reuse of the data. The ultimate aim is to combine these services into a common environment to allow project teams to interact with the instruments and even allow for a data collaboration between ANSTO and AS

MECAT FOCUS

The Australian Synchrotron intends to focus this project on three of its nine beamlines, because there is a great deal of disparity between all nine - too much for the project to deal with in the first instance. The Australian Synchrotron will be looking at the Soft x-ray spectroscopy, Infrared Microspectroscopy, and Macromolecular Crystallography beamlines. The Macromolecular Crystallography beamline already is using parts of the MECAT project software. The Australian Synchrotron is intending to take data from the experimental end stations, the proposal database, scheduling database and the EPICS control systems. At ANSTO, their scope is a little different. The Bragg institute is considering all of the neutron beam instruments, these instruments are smaller data volume producers, although of no less importance. Instead of EPICS they use a control system which is a local adaptation of the Swiss Spallation Instrument Control System (SICS) a collaboration from the Paul Scherrer Institute (PSI). The actual implementation at the Bragg Institute is nearly identical.

Both institutes will produce an ARCS compatible data repository, ARCS is the Australian Research Collaborative Services. This will then allow a set of standards to harvest that data publicly using metadata.

ARCS

ARCS lets researchers look for data, transfer data and to share material. It uses a concept called the "Data fabric" which is an overloaded term that has only recently been defined more precisely. It's been used like the "cloud", but aligned for research data purposes and has central; data storage, security etc. for facilities across Australia. As time goes by, there will be more and more experimental facilities using the data fabric provided by ARCS. The following description of ARCS [4] is pertinent:

The Australian Research Collaboration Service provides tools and services that enable researchers to operate at the forefront of their fields. It is intended to allow them to securely store large volumes of data for more collaboration. These tools and services also enable the transfer of data for faster analysis and result, to share material for convenience and control and finally to share data securely only with authorised

TARDIS

The MECAT project has chosen to use a particular technology to help collect the data for the databases,. It is called TARDIS, and stands for The Australian Repository for Diffraction ImageS, Ref [5]

TARDIS is a collaborative venture; coming out of the eResearch community. It is Australian and was started at Monash University, Australia [6]. It puts data into the dataset for various communities. It started off as a development for assisting users of the Protein Crystallography beamline. It has been made open source and is used for managing groups of files for a given experiment.

TARDIS available at the website ref [5] and is used for managing a group of files associated with an experiment – as per the following schema.

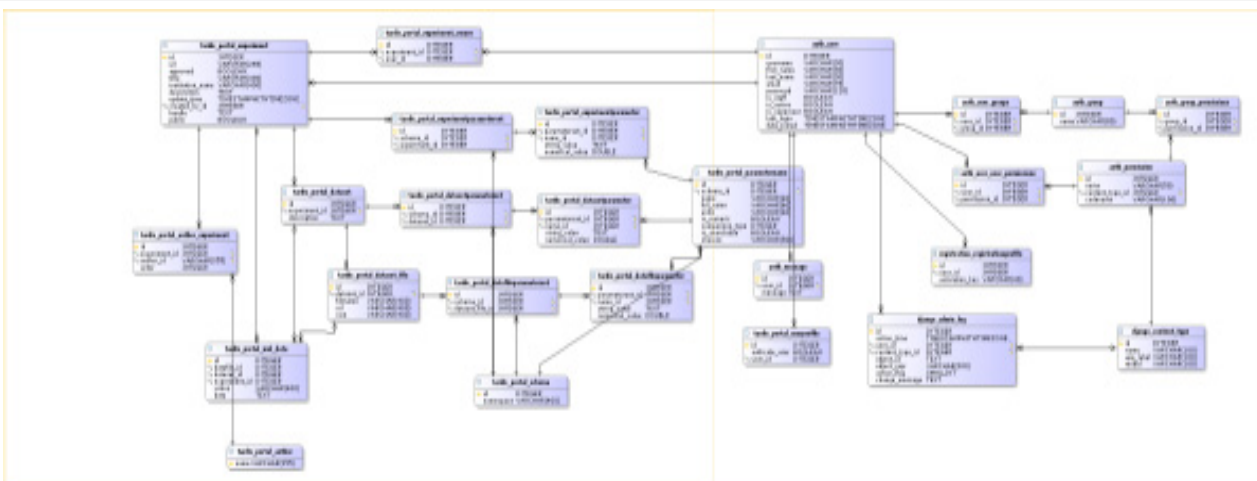

Figure 1: TARDIS Schema showing authorisation and experiment, parameters, database and data files

TARDIS SCHEMA

Figure 1 shows how TARDIS [5] looks like in its full schema. If we remove the authorisation, the complexity reduced significantly. It encompasses an experiment which consists of the "Dublin core" type information – that is title, author date etc. Then there are the soft parameters stored against that. Those parameters will generally be unique to the instrument and science being used for those experiments.

Now examine at the datasets themselves, there may be many datasets associated with a given experiment. This is the way that works for the two facilities in question and similar intuitions. Finally there are the files themselves and where they are located.
In summary, the schema consists of the experiment, the experimental data, datasets and files.

CONCLUSION

MeCAT is a joint project between the Australian Synchrotron and ANSTO to improve Metadata management and publication at the facilities. It is using and extending open source tools called TARDIS and will offer Australian Scientists greater capabilities to share and reuse data.

REFERENCES

[1] http://www.synchrotron.org.au/
[2] http://www.ansto.gov.au/
[3] http://ands.org.au/
[4] http://mecatproj.wordpress.com/
[5] http://tardis.edu.au
[6] http://www.monash.edu.au/