Requirements for data catalogues within facilities

Milan Prica[1], George Kourousias[1], Alistair Mills[2], Brian Matthews[2]

[1] *Sincrotrone Trieste S.C.p.A, Trieste, Italy*
[2]*Scientific Computing Department, STFC, Didcot, UK*

It has become increasingly common for Photon and Neutron facilities to support a data catalogue. This provides a systematic record of the data generated within facilities including information on experiments undertaken on instruments, the context in which data was collected, and information on how the data itself is stored. The contextual information would typically contain information on the experiment or other process (e.g. simulation, data analysis) which was undertaken to generate the data, such as information on the experimental team, the instrument, the sample and the parameters set. The data storage information would contain information on where the data set are located and organizes, together with information and controls on how to access the data. Examples of data cataloguing systems which are in use in facilities include ICAT [1] and Tardis [2].

Data catalogues are seen as a key part of the information infrastructure allowing the user community context based access to their data, both within the facility and at their home institution, providing the facility with a record of the experiments undertaken for audit and impact analysis, and providing the basis for publishing and sharing data with the wider community, encouraging review and reuse. As such data catalogues become federated, allowing cross-searching across facilities, they will support scientists to perform cross-facility, cross-discipline interaction with experimental and derived data. This will also deliver a common data management experience for scientists using the participating infrastructures particularly fostering the multi-disciplinary exploitation of the complementary experiments provided by neutron and photon sources.

The PaN-Data Open Data Infrastructure project [3] is investigating the deployment and use of data catalogues to provide a common approach to federated data publication, search and access within its participating facilities across Europe. As the initial stage of this activity, it has developed a set of criteria for evaluating data cataloguing systems for potential reuse and redeployment across the facilities. These criteria range from the metadata stored in the catalogues, the extent to which they can be integrated into the facilities processes, their scalability, to their support and total cost of ownership. In this presentation, we shall discuss the role of data catalogues in facilities infrastructure, which we then use to motivate and justify these criteria, and consider how they would apply to existing data cataloguing systems.

References

1. The ICAT Information Catalogue http://www.icatproject.org/
2. The Tardis Project http://tardis.edu.au/
3. The PaN-Data consortium http://pan-data.eu/

Email corresponding author: brian.matthews@stfc.ac.uk
Preference: Oral Presentation
Topic: Data Management