RDA-PanSig Workshop on Interoperability

**Alba Synchrotron,** **Barcelona, Spain,** 3-4 April 2017

It is well-known that the users of Photon and Neutron sources move between facilities and instruments, using and combining different techniques, together with simulation data, to further their science (see for example:<http://pan-data.eu/Users2012e-Results> ). They thus collect data in different formats, with different sets of experimental parameters and with different data set conventions, depending on the local practices of the facility and instrument. If users wish to search, access, combine and compare, and apply analysis software to data collected from different sources, they will need to take these differences into account, which could be a significant barrier to the effective use of the data. Further, software developers have to deal with different formats and parameters, and future re-users of data would need to search and access data in different systems.

A key issue is that the size of data sets already exceeds the capacity of facility users to handle them, so it is becoming essential that the facilities provide resources that allow such data to be rapidly stored and managed. However, it is not sufficient just to provide a repository for large data sets, since it may be impractical for users to download them. Rather, it is necessary that such data servers can return specific subsets of the data, such as data slabs or associated metadata, as requested by the user. Such granular access to the data presents a number of technical challenges that this workshop is designed to address.

In this workshop, we will explore how to support and enable the interoperability of data within and across user facilities. This would include considering the following aspects:

1. **Data Discovery**

How do we integrate existing online resources, such as ICAT, the Materials Data Facility, and Globus Online, so that the users are able to locate data to which they have access, no matter where it is stored? This would consider the use of common data and metadata standards and protocols for discovery, access and interpretation of data sets.

1. **Authentication**

How do we provide a unified authentication scheme that determines whether the users have access to the remote data? This would consider common schemes to identify users and how to accommodate different local authentication schemes.

1. **Server Queries**

What queries are required so that users can inspect the remote data and selectively download subsets of the data and metadata?

1. **Transfer Protocols**

How do we handle network requests and how do we serialize the data that is returned?

1. **APIs**What kind of APIs are needed/provided? How would applications need to be modified?
2. **Reproducibility**Do we need to capture and store queries to make a) processes reproducible and/or b) facilitate actions repeatedly accessing the same data slab?
3. **Controlled Vocabulary**

Do we need controlled vocabularies for data and/or resource discovery?

The results will be disseminated to facilities around the world for comment, with a view to providing international standards for accessing data through cloud servers. An open-source project would be established to implement such servers, although facilities would be free to implement their own to accommodate local variations in server and data architecture provided they respond to the agreed network queries and transfer protocols.

The aim of the workshop would be to identify and refine likely areas of work which could form a programme of work for a future RDA Working Group, within the scope of the RDA Interest Group on the Research Data Needs of Photon and Neutron Science. We would expect that the outputs of the workshop include:

* A workshop report, outlining a draft specification and a development roadmap to support a interoperable infrastructure across facilities, for consultation with facilities.
* Input into a draft RDA Working Group proposal describing a specific set of time limited activities within the roadmap, such as specifying a reference architecture and core interchange protocols
* Identifying potential implementations to provide a proof of concept of the working group recommendations.

**Possible contributors:**

* Facilities tools developers: e.g. ICAT, MANTID, DAWN, NeXpy
* Data repositories used by facilities : e.g. MDF
* Formats used by facilities: e.g. NeXuS
* Projects exploring aspects of interoperability: e.g. EOSC-Pilot, OpenAire
* User communities and publishers: e.g. IUCr, NIST
* Members of other RDA Groups.

**Agenda**

**Monday 03 April, Alba Synchrotron**

13:00 - 13:45 Welcome: And Registration (Lunch possible at the Canteen)

14:00 - 14:15 Introduction and purpose of meeting Dr. Brian Matthews

14:15 - 15:30

Data Discovery: Status reports. Interoperability requirements at different institutes.

***Would participating institutes like to give a view on on requirements for interoperability?***

15:30 - 15:50 Tea/Coffee break

15:50 - 16:35 Data Discovery: ICAT - Steve Fisher ??

16:35 - 17:05 RDA International Materials Resource Registries WG - Laura Bartolo ??

17:05 - 17:30 Discussion

19:30-22:00 Dinner

**Tuesday 04 April, Alba Synchrotron**

09:00 - 09:45 NeXpy, MDF and HDF5 - Ray Osborn

09;45 - 10:30 Using Nexus at Diamond - Alun

10:30 - 10:50 Coffee/Tea break

10:50 - 10:35 Using Nexus at Soleil - Majid

10:40 - 13:00 Discussion

13:00 - 14:00 Lunch

14:00 - 15:30 EU PROJECTS:

14:00 - 14:30 CalipsoPLUS JRA2. Remote data analyisis

14:30 - 15:00 EU PROJECTS: LEAPS. Brainstorming for the future European Photon Sources Strategies (on IT and data management)

15:00 - 15:30 European Open Science Cloud and Photon and Neutron science Pilot demonstrator : Brian Matthews (?)

15:30 - 15:50 Coffee / Tea Break

15:50 - 17:00

* EU PROJECTS: PaNData community and PaNDaaS submission
* Summary and next steps.