Some notes on the MDIIIG and PaNSIG joint session:

HDF

The HDFgroup presentation by Lindsay Powers on a new product (HDF product designer) was accompanied by fruitful and lively discussions. It is recognized that HDF provides one of the very few standards for binary data, and sustainability of this core software is crucial for the sustainability and reproducibility of research.

- HDFgroup plans to hold a HDF user conference in 2017.
  - Expressed our support of such an activity
- The product designer essentially serves as a data type registry. It might be a valuable tool to register (and possibly consolidate) NeXus standards. This will be investigated until the next plenary.
- H5serv, h5pyd might allow distributed data access particularly useful for deployment of Data Analysis Services on cloud (and other) platforms. We will investigate if this could be a valuable use case for example for the EOSC project.

Slides: https://goo.gl/IhGHH4

Frictionless Data

Dan Fowler from OKF presented the data packages and their particular power for exchange of tabular data. Though tabular data are not the typical use cases for Photon and Neutron Science applications, there are some experiments which produce light-weight data in legacy formats. Data packages might provide a very convenient way of exchanging these data without the need to actually reformat the original data. We will test the applicability.

Slides: https://goo.gl/lS46RE

Reusing Neutron Data

Devan Donaldson presented insights into sharing and reuse of neutron data he gained at ORNL. The recommendations derived from the small study were not too surprising, but it was very interesting to see how very simple measures can significantly lower the barrier for data sharing (and that there are actual use-cases). Devan would be highly interested to perform similar studies at other facilities and/or investigate cross-facility data sharing.

Slides: https://goo.gl/19qGk3

ActivePaper

Had to be dropped due to time constraints.

Towards working groups

During the various sessions of Materials, Chemistry and PaN interest groups, the notion of “controlled vocabularies”, ontologies and identifier for instruments repeatedly came up. The PaNData project had developed an ontology for facilities, instruments and experimental techniques, which however did not get into production.

Closely related, the partially incoherent standardization of meta-data in NeXus and other HDF5 based meta-formats (netdcf, openpmd etc), is affecting interoperability despite the common format for binary data. Data interoperability and interchange is hence another area, where a well-define work plan could lead to major improvements within a short timeframe.

Brian Matthews has hence proposed to launch one or two working groups in the area of

- Data interoperability and interchange
- Common vocabulary

The work would presumably focus on the needs of material science, chemistry and experiments at Photon and Neutron facilities, but based on the principles of RDA developments on DTR, Vocabulary services etc.

We intend to organize a workshop on data interoperability in conjunction with the next RDA plenary in Barcelona. Case statements will be developed prior to the plenary.

Slides: https://goo.gl/TWsnr5

Some notes on the PaNSIG session:

After a short introduction of the Photon & Neutron Science IG (https://goo.gl/Yi2Yg7), Jaroslaw Nabrzyski (Uni. Notre Dame) gave a short, insightful presentation about Container Strategies for Data & Software Preservation that Promote Open Science. There are no slides available for the talk; however https://osf.io/y9mpx/ provides a wealth of documents and details about the activities. In particular noteworthy (of immediate use):

- The Open Science Framework itself provides a nice and rather holistic environment to organize discussions and documents into groups and projects. Might actually be a good alternative to the way documents are organized within RDA.
- Smart container is an attempt to augment containers with ontology and provenance information, which appear as essential ingredients to achieve long-term sustainability and reproducibility (https://goo.gl/H3mmgO).
- It would certainly be beneficial (at least for us) to co-operate on specific topics (e.g. ontology, provenance, security), exchange knowledge (OSF seems a good platform) and possibly organize a small workshop to take up the developments and ideas.

Raymond Osborn (APS/ANL) gave a great talk about a data processing pipeline at APS, covering various aspects from real-time visualization, standards to distributed data analysis (elucidating the potential use of cloud services for neutron and x-ray data analysis). It was immediately apparent, that adoption of the pipeline to cloud or HPC platforms would require a more suitable data access model. The recent HDF developments around h5serv, h5pyd would smoothly fit into the analysis code (largely based on python and NeXus) facilitating deployment on a broader range of compute platforms. (Slides: https://goo.gl/hVbMZr).

• Frank Schlünzen (DESY) gave a brief account on the current developments and considerations around container utilization at Photon and Neutron facilities (https://goo.gl/1YT9qt). The common theme seems that most of the facilities are at an early stage using container for scientific workflows. The only container implementations chosen so far are Docker and Shifter (which is based on Docker and allows rather simple conversion of Docker into Shifter container). Most facilities expressed security concerns; in particular the handling of credentials for authenticated data access appears problematic. All facilities would like to share not only experiences, but also the actual efforts in container factorization and necessary developments. However, sharing images requires certain trusts (and presumably more than just a simple repository) which would need to be established across facilities. To get a step towards reproducibility, meta-data need to be embedded in a transparent way. Smart container might be suitable way to achieve this. Finally, a controlled vocabulary enabling the matching between scientific experiment and container capabilities could greatly improve users' experience: searching for container in a scientific domain or just by name appears largely inefficient and error prone.

The Reproducibility IG had a great deal of very interesting activities at the RDA 8th plenary. It certainly would provide a good platform to intensify knowledge exchange, developments or co-operations.