

The Research Data Alliance Photon and Neutron Science Interest Group

Amber Boehnlein,¹ Brian Matthews,² Thomas Proffen,³ and Frank Schluenzen⁴

1 SLAC National Accelerator Laboratory, Menlo Park, California, USA 2 STCF, Daresbury, Cheshire, UK 3 Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA 4 DESY, Hamburg, Germany

Introduction

Scientific research data provides unique challenges that are distinct from classic “Big Data” sources. One common element in research data is that the experiment, observations, or simulation were designed, and data were specifically acquired, to shed light on an open scientific question. The data and methods are usually “owned” by the researcher(s) and the data itself might not be viewed to have long-term scientific significance after the results have been published. Often, the data volume was relatively low, with data sometimes easier to reproduce than to catalog and store. Some data and meta-data were not collected in a digital form, or were stored on antiquated or obsolete media. Generally speaking, policies, tools, and management of digital research data have reflected an ad hoc approach that varies domain by domain and research group by research group. This model, which treats research data as disposable, is proving to be a serious limitation as the volume and complexity of research data explodes. Changes are required at every level of scientific research: within the individual groups, and across scientific domains and interdisciplinary collaborations. Enabling researchers to learn about available tools, processes, and procedures should encourage a spirit of cooperation and collaboration, allowing researchers to come together for the common good. These community-oriented efforts provide the potential for targeted projects with high impact.

A key player in this community-oriented approach is the Research Data Alliance (RDA) [1, 2]. RDA is an international organization designed to promote communities of practice in digital data management across scientific domains, and is supported by research funders in the US, Europe, and Australia. RDA is one mechanism for illuminating the special cases of research data. RDA, governed by a small executive team, encourages a wide-net approach by soliciting individual to form self-organized groups.

RDA supports two types of groups. “Working Groups” (WG) are intended to work towards concrete outcomes, and are finite in duration. “Interest Groups” (IG) are expected to have a longer time span with the idea of providing the opportunity for those with common interests to communicate, or perhaps to even discover that there is a global community who shares the same interests and concerns from a data-centric perspective. After all, many research groups are self-contained from a scientific perspective. IGs are also ideally suited to initiate and guide WGs and foster communication between interest and working groups.

The RDA hosts plenary meetings every six months. Each meeting has a theme and the scientific program is constructed to have joint plenary meetings in the morning, with subsets of the IG and WG meetings

running in parallel in the afternoon. This format gives the chance for the attendees to “sample” what is happening in the other groups, schedule overlapping sessions for closely aligned groups, and set aside time for more private discussions.

The photon and neutron research facilities are all facing data management issues at varying levels, and the objectives of the RDA are congruent with the needs of the facilities and their users. The common practice of leaving data management and curation largely to the data-generating scientists does not conform very well with scientific best practices, essentially requiring that the scientific data and the entire process from data taking to publication is well-documented and preserved. In the ideal world, the original scientific data, associated meta-data, and the provenance information can be traced back from the publication, preferably in a machine-readable way to permit easy data discovery and validation.

The data management issues—in particular, the need for solutions supporting large-scale collaborations—are emphasized by changing scientific approaches. The experiments are increasingly done by international collaborations utilizing several instruments across facilities. The number of publications acknowledging the combined use of synchrotrons, neutron facilities, or free-electron lasers (FELs) or the combined use of different experimental techniques is rapidly increasing. Consequently, the need for data management solutions which work across facilities and continents is growing as well. It’s not only the ability to access and transfer data which matters, but also the ability to use the same tools for the analysis of data, regardless if data have been measured at a spallation source in the US or a synchrotron in Asia. That requires certain commonly agreed conventions; for example, on the coordinate systems describing the location of the sample and the detectors with respect to a reference frame. A standard scientific file format supporting such conventions, as well as interoperability between different conventions, would certainly be beneficial and avoids the recurring development of small scripts converting between different formats and conventions.

PaNSIG

Recognizing the commonalities with the RDA and the requirements from the user communities has led to the initiation of a community centric interest group focusing on the issues of the photon and neutron science facilities and user communities. This Photon and Neutron Science Interest group (PaNSig) has been endorsed by the Research Data Alliance at the beginning of 2014 under the name “Research Data Needs of the Photon and Neutron Science Community.” The focus of the interest group is to exploit commonalities and synergies between the photon and neutron research infrastructures aiming at intensified cooperation, mostly but not exclusively concerning the management and analysis of scientific data. Large-scale photon and neutron facilities aim to provide not only the scientific instruments, but also to support the entire scientific process, a goal which is severely hampered by the lack of standards and common, globally applicable solutions. The Photon and Neutron Science Interest Group (PaNSIG) is not going to solve this dilemma, but it’s an excellent opportunity and forum to exchange ideas and concepts in the larger framework of the RDA, and drive developments and evolution of standards and solutions towards a more globally useful infrastructure. After endorsement, the first PaNSIG meeting took place at the RDA third plenary in Dublin, accompanied by co-located meetings on the ICAT data management system and the Umbrella AAI federation [3]. The meeting was kind of a kick-off meeting

with participation from US and European research facilities, representatives of user communities, and service providers like DataCite. The initial PaNSIG activities were indeed focusing on small tasks like extended user surveys, but got considerably more technical, taking up scientific file formats and scientific webinars at the subsequent PaNSIG meeting.

Global Collaborations

Complex experiments at photon or neutron sources are increasingly becoming large-scale endeavors. Some publications on X-ray diffraction experiments at FELs count more than 50 authors from more than 20 different research institutions (e.g., [4]). That is certainly not a particularly typical use case, but illustrates the need for suitable platforms supporting collaborative research. To investigate the issue in a bit more detail, the PaNdata project has issued a survey across the European facilities identifying the common users of the photon and neutron sources [5], which revealed that more than 20% of the 33,000 distinct users have been using more than one facility. One of the first activities of PaNSIG was to extend the survey to include facilities from non-European continents. The first to join the survey was the Spallation Neutron Source (SNS) of the Oak Ridge National Laboratory. As it turned out, SNS has common users with all European neutron sources and—more surprising—also all European synchrotrons. A remarkable 15% of the SNS users also utilize neutron sources in Europe, emphasizing the global and volatile nature of the photon and neutron user communities. The survey was further extended based on publication records of the facilities' publication databases [6]. Each facility is encouraging their users to acknowledge the use of instruments and register publication through user portals. Tracking the publications from acknowledgements is essentially impossible, but the registered publications can be compared to find publications using several facilities across continents. The simplest way of finding matching publications would be based on DOIs providing unique identifiers. Unfortunately, most facilities record all kind of details but not the DOIs. CrossRefs' API [7] allows retrieving DOIs from publication records, but this has its limitations. In particular, false positives in DOI retrieval can obscure the results and the required registration of a publication in several independent databases leads researchers to notify only the most important resources, which can reduce the number of matching publications across facilities. Despite these limitations, matching DOIs registered in the different publication databases show a tight network of links between almost all facilities, indicating that a significant number of experiments actually do use several of the instruments in combination to achieve scientific goals (see Figure 1).

This kind of information is recognized and used by policy makers and funding agencies, as well as providing the motivation to standardize because it is in the interest of the research community. Involvement in the PaNSIG activities and participation in these activities can make the case of the photon and neutron sources considerably stronger on a global scale.

Scientific Webinars: CLNSF

In the US, the Collaboration of the Light- and Neutron-Source Facilities (CLNSF) has initiated a series of webinars with the focus on scientific applications using the instruments provided by the facilities, including all aspects of data management, computation, and data analysis [8]. The goals of the CLNSF

webinars appeared to be well aligned with the goals of the RDA and—of course—the PaNSIG activities. At the fourth RDA plenary in Amsterdam during the second PaNSIG meeting, the IG sought support of the CLNSF webinars. The basis of organizers and contributors could be considerably enlarged, aiming to render the CLNSF webinars a sustainable and scalable effort.

The CLNSF webinars have been successfully re-launched on a new conferencing platform right after the PaNSIG meeting with contributions from various scientific and computational fields (see Figure 2 for two examples) and a lively audience of up to 35 participants for a single webinar. All slides and recordings are available online [9], also providing a nice educational resource. The webinars are aimed at a global audience and are completely open, so anyone interested is invited to join or contribute.

Scientific File Formats

One of the most recent PaNSIG activities was the attempt to establish an interest group on scientific file formats and a working group on HDF5 extensions. Scientific file formats are an essential element achieving interoperability between applications or between scientific experiments using different instruments. Scientific file formats are also crucial for archival and long-term preservation of data. Data at light sources are still pre-dominantly stored as tiff images. This might be very convenient from the application point of view, but the lack of meta-data and the huge overhead in writing and reading millions of small files impair the performance of data infrastructures. Photon and neutron sources have a long history of collaborating on scientific file formats. In structural biology, for example, CBF has been in use for decades, with a well-defined meta-data structure accompanying the binary data. On the other hand, NeXus is a well-established standard at neutron facilities worldwide and increasingly is becoming a standard at light sources, as well. For example, the detector manufacturer DECTRIS supports NeXus natively on the most recent detectors [12], since the underlying HDF5-libraries become essential for high-throughput data recording.

Recognizing the importance of the two independent standards has led to intensive work on the interoperability between CBF and NeXus. The standards were originally hardly compatible, but with slight modifications on both sides, full interoperability has been achieved through collaborative efforts of the IUCr, the NeXus developers, the NeXus International Advisory Committee (NIAC), and the PaNdata initiative [13]. These kinds of efforts could well be promoted and supported by PaNSIG and the RDA. HDF5 is one of the very few scientific file formats fulfilling the needs of photon and neutron sources and, in particular, the needs of tFELs, which are operating at an extremely high repetition rate and facing particularly challenging concurrent read and write requirements.

The most important theme of the PaNSIG meeting and the Scientific File Formats BoF sessions at the RDA fourth plenary was consequently the standardization, interoperability, and evolution of file formats like HDF5 and netCDF. With prominent participation of the HDF group and UNIDATA, the discussion on evolution of HDF5 and implementation of new features like SWMR (single writer, multiple reader) and Virtual Datasets has stimulated investigations of future implementations, funding schemes, and support for new developments. Several photon facilities are currently driving and funding HDF5 developments.

Defining and prioritizing future developments (see Figure 3), as well as collaborative funding to realize the implementations, will remain important issues for PaNSIG.

The discussions and developments are continuing (see also [15]), and will be a topic at the next PaNSIG meeting as well as a co-located workshop.

Conclusions

Membership of PaNSig is open to all members of RDA, which in turn is free to all individual researchers and practitioners. PaNSIG can act as a mediator between the RDA as an organization and the vast spectrum of activities on the one hand, and the facilities and user communities on the other hand. Several of the colleagues involved in PaNSIG are also taking an active part in the RDA/CODATA Materials Data, Infrastructure & Interoperability IG [16], Structural Biology IG [17], Federated Identity Management IG [18], and several other RDA groups, which might provide a convenient way to stay informed and launch projects of mutual interest. The Materials IG has also proposed to hold a joint sessions at the fifth RDA plenary in San Diego [19], which will be a fruitful addition to the PaNSIG sessions.

PaNSIG will continue the activities around the user and publication recordings. To arrive at common user or publication databases, or just a federation of existing instances, appears very unlikely. However, the consequent use of persistent identifiers and means for automated discovery of publications based on experiments at photon and neutron sources will remain on the agenda. It could provide facilities with a much easier tool to populate publication databases and save users from registering publications in an ever-increasing number of user portals. Participating in the RDA or PaNSIG activities can have its benefits.

The RDA covers a very broad range of aspects around scientific data without aiming for specific implementations, avoiding community specific solutions, which work just for a single scientific case, as well as the ultimate one, which works for all use cases but not too well for any of them. The RDA offers, however, an excellent overview on policies, recommendations, best practices, and existing implementations and services.

PaNSIG, on the other hand, is more community-centric and can offer a view on actual implementations and identify and initiate opportunities for cooperative developments. Great open-source tools for data management or data analysis, such as ICAT, SPOT, or Mantid and DAWN, are good examples of collaborative efforts leading to powerful and usable products for photon and neutron science cases. Furthermore, PaNSIG offers opportunities to shape the future direction and collaborative funding of implementations of essential standards like HDF5.

References

- [1] Treloar, *Learned Publishing* **27**(5), 9–13 (2014).
- [2] F. Berman, R. Wilkinson, and J. Wood, Building global infrastructure for data sharing and exchange through the research data alliance. Available at: http://www.dlib.org/dlib/january14/01guest_editorial.html
- [3] Details and slides available at: <https://indico.desy.de/event/1stow>
- [4] J. Küpper, *Phys. Rev. Lett.* **112**, 083002 (2014).
- [5] Available at: <http://pan-data.eu/Users2012-Results>
- [6] Available at: <http://pan-data.eu/Users2014-Results>
- [7] Available at: <http://search.crossref.org/help/api>
- [8] Available at: <https://confluence.slac.stanford.edu/display/CLNSF/Home>
- [9] Available at: <https://indico.desy.de/categoryDisplay.py?categId=333>
- [10] D. Gursoy, CLNSF Webinar on TomoPy. Available at: <https://indico.desy.de/categoryDisplay.py?categId=333>
- [11] Y. Nashed, CLNSF Webinar on Ptychographic reconstruction. In J. Deng et al., *Simultaneous Cryo X-ray Ptychographic and Fluorescence Microscopy of Green Algae*, PNAS 2015 (in press).
- [12] DECTRIS success stories, available at: https://www.dectris.com/success_stories.html?page=2.
- [13] H. Bernstein, et al. Coping with BIG DATA image formats: Integration of CBF, NeXus and HDF5: A progress report. Available at: <http://www.eposters.net/pdfs/coping-with-big-data-image-formats-integration-of-cbf-nexus-and-hdf5-a-progress-report.pdf>
- [14] Available at: <http://www.hdfgroup.org/>
- [15] Available at: <http://confluence.diamond.ac.uk/display/HDF5DEV/HDF5+Developments>
- [16] Available at: <https://rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html>
- [17] Available at: <https://rd-alliance.org/groups/structural-biology-ig.html>
- [18] Available at: <https://rd-alliance.org/groups/federated-identity-management.html>

Figures

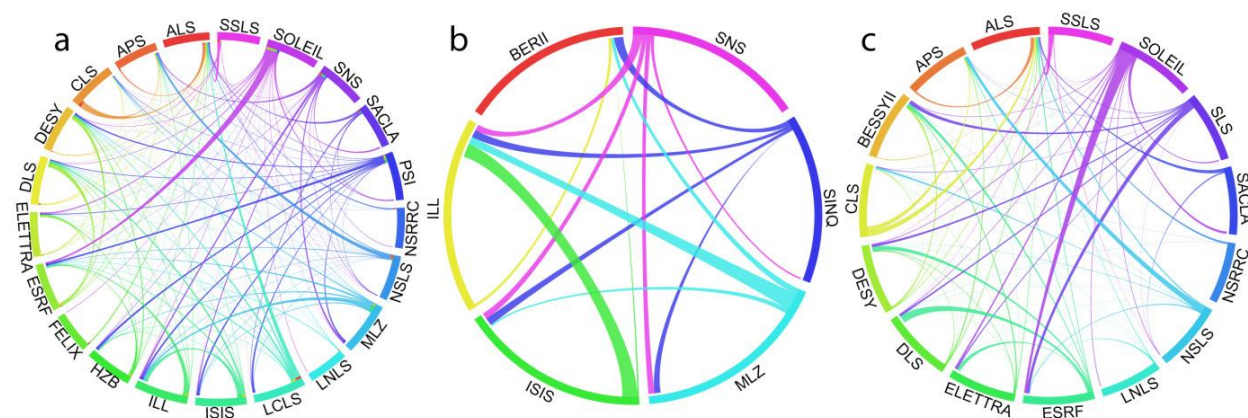


Figure 1: (a) Common publications (DOI) across all facilities; (b) for neutron sources; (c) for synchrotrons.

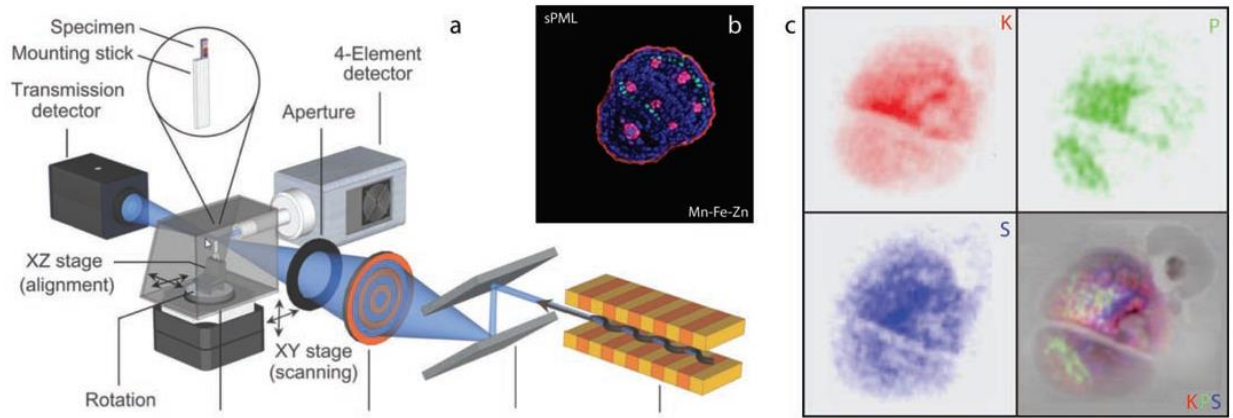


Figure 2: Examples from CLSNF webinars. (a) Set-up for X-ray fluorescence tomography; (b) distribution of chemical elements in arabidopsis seeds [10]; (c) combined cryo-ptychography and fluorescence imaging of marine algae (*Ostreococcus* sp.) [11].

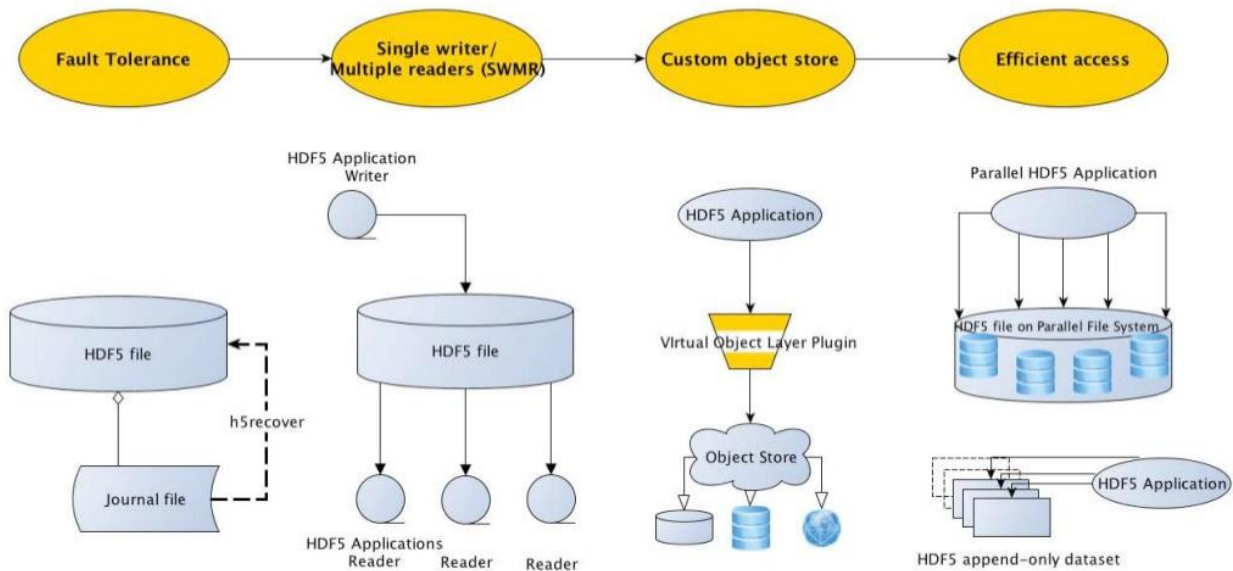


Figure 3: Road map for HDF5 1.10 [14].