Experimental Data Curation at Large Instrument Facilities with Open Source Software

Line Catherine Pouchard
Computational Science Initiative
Brookhaven National Laboratory

Kerstin Kleese Van Dam
Computational Science Initiative
Brookhaven National Laboratory

Stuart I. Campbell
National Synchrotron Light Source-II
Brookhaven National Laboratory

Abstract

The National Synchrotron Light Source II operating at Brookhaven National Laboratory since 2014 for the US Department of Energy is one of the newest and brightest storage-ring synchrotron facility in the world. NSLS-II, like other facilities, provides pre-processing of the raw data and some analysis capabilities to its users. We describe the research collaborations and open source infrastructure developed at large instrument facilities, such as NSLS-II, for the purpose of curating high value scientific data along the early stages of the data lifecycle. Data acquisition and curation tasks include storing experiment configuration, detector metadata, and raw data acquisition with infrastructure that converts proprietary instrument formats to industry standards. In addition, we describe a specific effort for discovering sample information at NSLS-II and tracing the provenance of analysis performed on acquired images. We show that curation tasks must be embedded into software along the data life cycle for effectiveness and ease of use, and that loosely defined collaborations evolve around shared open source tools. Finally we discuss best practices for experimental metadata capture in such facilities, data access and the new challenges of scale and complexity posed by AI-based discovery for the synthesis of new materials.

Received 04 January 2019 ~ *Accepted* 24 July 2019

Correspondence should be addressed to Line Pouchard, Brookhaven National Laboratory, PO Box 5000, Upton, New York, 11973-5000. Email: pouchard@bnl.gov

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The National Synchrotron Light Source II (NSLS-II)[1] operating at Brookhaven National Laboratory (BNL) since 2014 for the US Department of Energy (DOE) is one of the newest and brightest storage-ring synchrotron facility in the world. With 25 beamlines in operation, four under construction, and over 30 planned, it has already served more than 1,000 distinct scientists conducting x-ray characterization experiments in 2017, projected to increase up to 1,300 for 2018. Experiments conducted by users from the academic, government and industry sectors and facilitated by the beamline scientific staff acquire data with a range of high throughput detectors. Because today's experiments are conducted as ensembles or groups of related experiments rather than single experiments, users and their samples tend to move between detectors at the same facility and between facilities, taking advantage of the different characterization methods available across the world . Facilities follow a general pattern of acquiring data from samples using high precision instruments, reducing data and reconstructing images, and enabling near real-time analysis. However, the wide variety of detectors, data acquisition methods, storage systems, and data curation philosophies present significant challenges to users and data management support teams at the facilities . These challenges are addressed by using and customizing open source software shared between loosely defined collaborations.

NSLS-II, like other facilities, provides pre-processing of the raw data, and some analysis capabilities to its users. While preparing a sample for an experiment and applying for beamline time takes months, users typically spend between a few hours (beamtime is allocated in eight hours shifts) to a few days on site to perform their experiments. The complex data acquisition process produces large volumes of raw data in the form of images and multi-dimensional data describing the experimental process. Preliminary processing of theses experimental images outputs derived data and is provided at the facility. Preliminary processing aims to remove experimental artefacts and convert from instrumental units (e.g. angle) into physics units (e.g. d-spacing) during the allocated beamline sessions. Increasing computing power at the beamline is required to monitor an experiment in progress and to accommodate the growing frame rate of new detectors. In addition to increasing data volumes it is necessary to provide processing and analytical feedback to a user as quickly as possible to allow adjustment of experimental parameters such as intensity, position of the detectors, or sample environment conditions (e.g. temperature). This presents additional challenges for the infrastructure in terms of storage and computing performance as larger volumes of data needs to be stored and analysed in short periods of time.

This paper describes the research collaborations and open source infrastructure developed at large experimental facilities, such as NSLS-II, for the purpose of curating high value scientific data in the early stages of the data lifecycle . We show the specificity of data management and curation at large facilities, emphasizing aspects of metadata, discovery, and integration with analysis. We demonstrate that, in such facilities, curation activities must be embedded into software along the data life cycle and that effective collaborations evolve around shared open source tools. In addition we discuss best practices for experimental metadata capture, data access and the new challenges of scale and complexity posed by AI-based discovery for the synthesis of new materials. Efforts

1 National Synchrotron Light Source-II: https://www.bnl.gov/ps/

 doi:10.2218/ijdc.v14i1.637

to provide data infrastructure at NSLS-II include the development of a common data acquisition system called Bluesky[2] and a storage system software named Databroker. We increase the impact of these systems by promoting enriched data discovery and provenance tracking across beamlines. Bluesky is already being adopted at other light sources in the US, namely the Linac Coherent Light Source (LCLS), and the Advanced Photon Source (APS).

Related Efforts

An overview of data curation and preservation challenges at other large scale, experimental facilities is provided in (Matthews, Crompton, Jones, & Lambert, 2015), focused on the programs launched by the UK Science and Technology Facilities Council (STFC). Preserving scientific data by itself is not sufficient to secure future understanding of physical phenomena, this requires curating and preserving many other research objects associated with experimental context and set-up, instruments and their calibrations, raw data transformations, and computational analyses. Curation and preservation of research objects and artefacts (RO), rather than data, has been well advocated; ROs contain organized, semantically rich resources aggregated into units of knowledge (Bechhofer, De Roure, Gamble, Goble and Buchan, 2010). As noted in a recent review of the Materials Genome Initiative (MGI) high throughput experimental methodologies (Green et al., 2017), materials science lags behind in data curation, data access, standardized data formats, and optimized coordination. The result is underdeveloped interaction between research efforts, the kind that would fully exploit existing infrastructure elements and promote opportunistic discoveries on a large scale in energy production, conversion, and storage, catalysts, microelectronics, and other areas.

Previous efforts related to metadata include the Core Scientific MetaData (CSMD) (Matthews et al., 2010) and the NFFA-Europe efforts. CSMD, developed within STFC, describes data from experimental facilities with core concepts (including datafile, dataset, investigation, investigator, software version, parameter), facility descriptions (shift, instrument, facility-cycle, facility scientist, etc.) and others. "Sample" is included as Auxiliary Information alongside format and type, and cannot be described in details with this format. While instantiations of CSMD focus on raw data, the extended version of CSMD describes derived data, the analysis process, and attributes of analysis software (Yang, Matthews and Wilson, 2013). The Nanostructures Foundries and Fine Analysis (NAFFA-Europe) is an effort that brings together European research labs providing access to characterization methods and computation at the nanoscale. This effort aims at federating data management and curation for facilities in the European Union, the scope of its metadata includes description across the entire data life cycle. As such, the high-level metadata vocabulary element "sample" needs to be further specified, as our own effort does.

2 Bluesky: https://nsls-ii.github.io/bluesky

Experimental Data Infrastructure at NSLS-II

NSLS-II Data Acquisition System

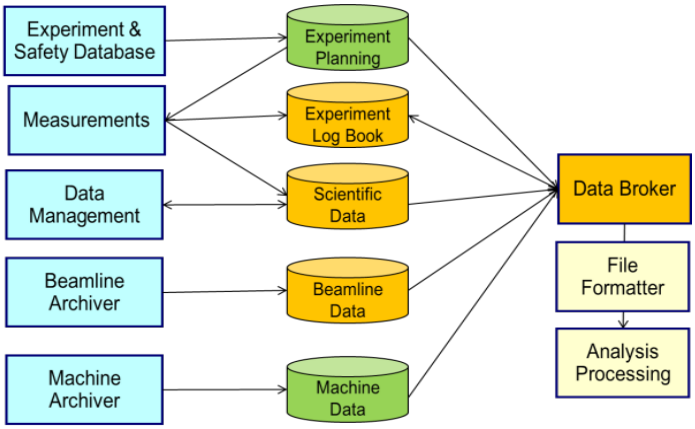

Figure 1. Schematic representation of experimental data workflow at NSLS-II beamlines. This workflow independently runs at each of the beamlines. Once acquired, metadata is stored in a Databroker database and data in a file system partition for each beamline.

At the core of the NSLS-II data acquisition and data management system is an essential set of flexibly structured data stores (see Figure 1) . These stores are accessed using the Databroker Application Programmer's Interface (API). This set of data stores are tailored for specific types of experiments, and store not only the raw data from a detector, but also additional metadata from the beamline, accelerator, logbooks, the proposal system, sample management and scheduling systems. The Databroker is a key piece of data handling interface software[3]. It enables the development of data analysis procedures for visualization and analysis of the real-time data to allow decision making, and for post-processing of all data after the actual experiment, without the scientist having to worry about the specific on-disk data format or representation. The main aim is to be able to separate the data I/O code from the scientific code by providing a common data API to all of the experimental data, irrespective of where it comes from, or what format it is stored in.

NSLS-II beamline data acquisition system is built to provide an essential infrastructure that can grow and evolve over the life of the facility. NSLS-II has developed a suite of open source software packages that not only support new detectors and new techniques as they continue to evolve, but also allows the analysis of the quality of the data as it is being collected and the performance of the beamline. The user facing package is called Bluesky and is for experiment control and the collection of scientific data and metadata that is stored in Databroker databases. Bluesky interacts with the beamline hardware through the standard programming language, Python using the EPICS interfacing software.

At the start of this project, centralized computing resources for data management and analysis at NSLS-II are in the planning stages. The Databroker high level features include search on metadata with searchable fields customized at each beamline, automated end-of-run data export to industry-standard formats, a sample management

3 Databroker: https://nsls-ii.github.io/databroker

 doi:10.2218/ijdc.v14i1.637

inventory database, and a service architecture to allow remote access to Databroker functionality. Our effort extends the Databroker solution and promotes searchability across beamlines by designing search mechanisms across multiple databases and a sample metadata schema for use in the NSLS-II facility.

Scientific Metadata for Sample Description

This section presents our efforts and results in designing a metadata schema that will describe user samples. A metadata schema describing samples was not available for our users at NSLS-II. Standard file formats exist for crystallography data, a branch of materials science leading the way with standard descriptions; a crystal is a material characterized by a lattice pattern of microscopic structures. The Crystallography Information File and Framework (CIF) is a flat file format (text) representing the structure of a material (Hall and McMahon, 2005). The CIF, designed in 1990, has been very successful at combining text and scientific data in a format both human- and machine-readable and has been adopted by numerous journals, databases, and users to archive and distribute crystallographic information. CIFs include unique identifiers as a CIF ID, citation details including a DOI, a version number, and scientific data describing for instance the position and angles of atoms in a crystal, space group, and more. The CIF standards has enabled the development of over a hundred open source software modules in use by the community since its inception[4]. The inclusion of both scientific data describing a particular compound and the citation details of the original citing paper in a single text file ensures keeping track of a provenance trace for a crystal structure deposited in a CIF database. Referencing a CIF is one way to embed curation in software development and data life cycles for an experiment. A CIF cross-reference in any schema opens the possibility of exploiting a standard format with its associated tools and referenced sources.

We designed a metadata schema for samples that captures how samples are presented to a beam. In our project, samples are described by several main objects: a constituent object, and a container object itself made of a geometry object, an out-of-beam constituent object and an in-beam constituent object. Constituents describe the chemistry of the sample and its container. Taken together, these describe sample structure and the conditions present at the beam when examining a sample. The container object describes the environment containing a sample presented to a beam. The sample can be characterized outside its operating environment (e.g. a sample in air) and/or where the studied properties arise (e.g. a compound in a gas at a certain temperature and pressure, a catalyst in an acid solution). The schema includes attributes such as sample maker, collaborators, Principle Investigator group, preparation start-date, tags, and common name. In addition, our schema harnesses the extensive data and metadata descriptions in CIFs by including a CIF ID and its corresponding CIF database in the constituent object. CIF IDs are unique identifiers only in the context of the CIF database they belong to; as quality and degree of curation in these databases vary, it is important to reference the source of a CIF. Sample common names are a convenient way for referring to sample representing combinations of chemical elements and reactions, for instance titanium dioxide. But materials with the same name could have different properties due to variations in their atomic structure. Mention of a CIF and its original database source in sample metadata ensures that the measured properties in the experiment can be compared to a known structure. CIF curation is harnessed into our metadata at an early stage of the experiment. Further curation of

4 IUCr Crystallographic Information Framework: https://www.iucr.org/resources/cif/software

metadata in our system is ensured by the use of google forms linked to validation scripts for data entry testing for values in required fields.

Discovery Across Multiple Beamlines

The design of new materials and compounds with predicted properties that achieve specific physical and chemical requirements under given experimental conditions are a key scientific advancement in materials science and a major goal of characterizing the physical structure of new material samples at the atomic level using particle accelerators. Traditionally new materials are synthesized following domain experts' intuition informed by deep knowledge and guiding the process. The choice of materials may be dictated by computational structures and the scientific literature. New samples, for instance crystals grown in labs, are then examined in a beam for solving their structure, and computational analysis is performed to determine physical properties. In order to retain a competitive innovation advantage, for a research organization, a company, and nationwide, it is necessary to accelerate this process of discovery and automate the search for new combinations of elemental structures that may result in materials with the desired physical properties. This is the goal of the Material Genome Initiative, with its named analogy to genomics and reference to the large scale automated discovery enabled by sequencing genomes. In order to fulfil the promise of automated discovery in materials, infrastructure must be built that provides connections between multiple data sources, databases, experimental notebooks, and the scientific literature. Numerous databases exist that focus on experimental or computational structures and properties, but the connections between these building blocks are missing. Only when many data sources can be assembled by providing connecting infrastructure can we hope to scale up discovery and take advantage of the "big data" opportunities offered by high throughput detectors.

In our effort at NSLS-II, we are building one such connection by enabling search across the metadata stores of multiple beamlines using the open source, highly scalable and extensible, commodity software Elasticsearch[5] (ES). Our goal is to enable federated search across large, heterogeneous data sources, such as databases with different schemas, no schema, and other sources such as files, providing one of the many connections needed between existing infrastructure components. Using ES, scientist users visiting the facility are able to quickly look up the parameters and calibration of a prior experiment they have performed at our facility during their current experiment. Samples brought to NSLS-II for an experiment may have been examined with other detectors. With ES, users can search across experiments performed at several beamlines and other sources in one interface. We built an ES index from fields in Databroker metadata databases. The ES index accepts JSON documents as input, with multiple benefits for our goal: first, no complex mapping is required to implement search across multiple database fields; and second, database content, and flat files can co-exist in the same index. This last feature enables bringing to users additional data sources not part of Databroker. Our system supports searching reference structures in CIF format from the Crystallography Open Database (COD)[6], an open access collection of files for organic, inorganic, and metal-organics compounds and minerals. As our sample metadata includes a CIF ID field, a user can quickly look up CIF references from COD and experiment parameters from Databroker in the same search interface. The user will find additional details such as the original citing paper, space group, atomic coordinates,

5 Open Source Search and Analytics: https://www.elastic.co/
6 Crystallography Open Database: http://www.crystallography.net/cod/

 doi:10.2218/ijdc.v14i1.637

etc. that are included with CIF for their sample during their experiment. ES also supports powerful search functions that include numerical ranges and fuzzy search, which takes care of the erratic use of hyphens and typos in chemical compound searches (Ni-Ca vs. Ni Ca, "perofskit" instead of perovskite).

Provenance Labels for Streaming Image Analysis

Advanced analysis capabilities provided at NSLS-II are specific to each beamline as techniques vary greatly according to the experimental and analysis mode. In our streaming analysis pipeline now deployed at two beamlines the computational tasks required to perform analysis are organized in a workflow as a Directed Acyclic Graph (DAG) (Pouchard, et al., 2019). A DAG is a network of connected nodes representing processes on the data, each with input and output, and an order of execution. An example of simple analysis is the calculation of statistics for a window of several images streaming from the acquisition system. Our system provides a provenance label and a timestamp for each node in the analysis workflow graph and its data. The labels associate an experimental scan plan and data IDs from Databroker. The DAG workflow execution analyzes images while data acquisition from the experiment is still in progress (streaming). In addition, using the provenance labels, a user can review the computational steps for each data transformation from the retrieved images to the final results. This review facilitates the validation of results while an experiment is in progress. After an experiment has completed, users can run the computational workflow again, assured that they are using the exact same data and conditions to verify their results. The provenance labels provide access to every node of the computational graph so that users can also re-run the workflow varying data input, and analysis parameters as desired. The attribution of provenance labels in streaming analysis supports quality control of results, replication of analysis, and brings us one step closer to reproducibility of an experiment.

Discussion and Lessons Learned

Sample Metadata

Sample metadata has been a crucial and somewhat neglected piece in the materials characterization process. In addition to the elements noted above, details about the manufacturing of the sample would be useful. However, users may be uninterested in providing even minimum details about their sample as a result of habit or due to lack of time. They may also be reluctant to disclose details even when strict permissions are in place to safeguard intellectual property. Another obstacle to providing detailed metadata is that users may know only basic information, such as chemical composition, for their samples: they come to the beamline to learn more. High throughput detectors enabling the characterization of sample collections are relatively new, and users have not felt the need historically to record metadata using methods other than spreadsheets and lab notebooks. Metadata recorded with these methods are not easily searchable, shareable, and interoperable. A part of the solution is to require users to fill out some metadata relevant to scientific research when they request time allocation to a beamline. While facilities usually request basic safety information about a sample for transport and

removal (e.g. has it been irradiated?), few request that users add scientific metadata as well. For instance, the Spallation Neutron Source (SNS) at Oak Ridge National Lab includes a Description metadata element (which can include a chemical formula) and a state element (e.g. single crystal, powder). The Diamond Light Source facility in the UK requires more extensive sample metadata such as function of the sample, sample type (state) and container descriptions to be filled out from a controlled vocabulary.

At some beamlines operating at NSLS-II, users are required to capture some metadata before the beam is activated and are encouraged to fill out additional metadata fields. Referencing published structures for chemical compounds and their atomic-level characterizations in the constituent object of our schema using a CIF value is an essential part of embedded curation that provides a link to previous results. As a best practice, all beamlines should implement similar requirements, thus helping to move towards embedded curation in systems they already use, such as the proposal system and safety information forms. This practice would ensure that at least some scientific metadata is captured, but it is insufficient by itself as many characteristics of the sample are discovered after experiment and analysis. Thus easy-to-use mechanisms for capturing scientific metadata for samples further down the experimental process are needed in addition.

Data Access and Sharing

The discovery system provides incentives to users for adding more metadata about the sample to their experiments. Based on self-assessment of users at one beamline, it is clear that they immediately perceived the added value of our discovery system as they can search in the COD available through our discovery system for reference structures during their experiment. Returning users are able to compare their current experiment parameters and resulting images with those performed months before or with a different detector. Even a minimal amount of metadata such as atomic count in chemical composition goes a long way towards enhancing the impact of the discovery system.

One major obstacle to automated data discovery on a large scale in materials is the lack of data access and the ubiquitous presence of paywalls. The culture of open access to data and publications in materials science is in its infancy when compared to other disciplines such as biology. For instance, during a demonstration of our discovery system, domain experts noted that the Inorganic Crystal Structure Database[7] would be a better source of CIF files than COD because its content is constantly updated and curated and its reference structures are of better quality. We could bring this content into our system by negotiating access policies and costs. BNL libraries have negotiated access for manual search and download of individual structures, however bulk access via API suitable for computational processing remains elusive. While COD is open source, free, and sufficient for prototyping purposes, a production system for NSLS-II would require access to higher quality structures.

From a broader perspective, users in materials science are less reluctant to share their data with colleagues they know than providing access to the general community. Data portals are seen as enablers for collaboration within the scope of a project. Sharing data at the early stages of a research life cycle allows scientists to get better help from data experts with setting up analysis. Describing samples and experiments with a common format available in a data portal promotes modelling studies both pre- and post-experiment (pre-experiment modelling, as takes place in experimental design, fosters a principled, systematic approach to data collection). A reluctance to sharing raw

7 Inorganic Crystal Structure Database: http://www2.fiz-karlsruhe.de/icsd_home.html

 doi:10.2218/ijdc.v14i1.637

data, even within the scope of a project, sometimes comes from the fear of being misunderstood, as more advanced but proprietary analysis software not available to all members may lead to different conclusions. The role of funding agencies in requiring that materials data be made more accessible, for instance in following up with promises made in data management plans, cannot be underestimated.

Data Policies

In contrast to the European facilities, such as those under the UK STFC, many US facilities have not all adopted data archival policies. For instance the APS does not provide any long term data archiving and management in its policies. LCLS Data Retention Policy distinguishes between types of data (raw, derived, results) with different tiers of access (front or deep storage) and back-up for the length of retention (from four months to two years and more). NSLS-II has kept all data and images since its first flight in 2014, is now committed to a minimum data retention and access period of one year, and is implementing a multi-tier storage policy. A User Agreement between BNL and the user home institution is put in place that covers liability, intellectual property, and financial fees (when the experiment is performed by a user affiliated with a for-profit institution). Users transfer data to their institutions via networks and leave for home with backup copies on portable devices. Data volume is one of the obstacles to preservation often cited by facilities. As data grow, transferring these data over networks becomes precarious, even with the reliance upon robust transfer protocols. Increasingly, analysis capabilities that provide data reduction onsite are provided at NSLS-II and other user facilities, resulting in lower data volumes for transfer. While facilities in the US should develop strong data policies, institutional motivation to do so is growing. Making a strong business case for preservation, such as described in (Matthews, Crompton, Jones and Lambert, 2015) is needed to overcome the cultural and financial obstacles inherent in developing and implementing such policies.

Machine Learning (ML) for Materials Discovery

The relatively new advent of ML-based discovery for synthesizing new materials presents additional challenges to curation. The materials science disciplines are represented in numerous databases that could in principle form the basis for building training models and drawing statistical inferences. Performing large parameter sweeps across millions of structures and properties in databases, relating them in a meaningful way, and aggregating feature sets are required tasks to produce promising designs and simulate future experiments. Training models require large amounts of representative data in order to perform accurate predictions. Materials science databases contain either large numbers of similar types of samples (for instance in ICSD) or small numbers of diverse ones but not both. Databases can be specialized for types of materials (for instance inorganic thin film materials, organic polymers) and methods of data acquisition (Zakutayev et al., 2018). Despite the appearance of abundant data, the specialization in small data sources and lack of diversity in large ones can impede the creation of accurate models. Examples of curated, materials structures that have been successfully used for ML applications can be found in the Materials Project, however these are computational structures obtained from ab initio calculations, and not experimental structures . Both types of structures are needed to achieve the goal of initiatives like the MGI.

Another challenge when applying ML to materials design is the precision and completeness of metadata found in existing databases. Properties of existing materials and samples relevant to properties algorithms are trying to predict must be present in training samples in order to draw inferences with a reasonable degree of accuracy. Multiple relevant properties must also be present to enable aggregation in feature sets for ML algorithms to perform. If metadata is incomplete, this can result in exclusion of records from data used to train models, further reducing the potential size of training sets. Another consequence of incomplete metadata is the poor accuracy of predictions. The quality of datasets should be quantified for metadata completeness, as such weights could be used in model tuning. The curation tasks performed at NSLS-II include sample data annotation with metadata elements and storing experimental configurations, as described in the previous sections. To improve curation, NSLS-II could require its users to fill out most metadata fields instead of the few currently required. Synthesis parameters, such as manufacturing temperatures and pressures for new samples, are not currently included in our sample metadata but could be required as well as they are instrumental in determining properties.

Collaborations Through Open Source Software

Collaborations are key to the effective deployment and customization of infrastructure at a large experimental facility. They are loosely defined as teams form and dissolve around a specific project, a software installation, a demo or a shared short-term purpose. The quality of the open source software exemplified in the extent of documentation and availability of support influences effectiveness. While standardization from national or international institutions is not required, using software that other institutions also use builds familiarity and facilitates adoption of tools by users, especially for the growing number of users who move between facilities. Diversity of skills and experience in our teams and an ability and willingness to communicate across disciplines, in this case physicists, chemists, material scientists, computer scientists, and curation experts is crucial to the success of our effort, with computer and data scientists often providing the glue through the tools they develop.

Our effort aligns well with MGI through high-level collaborations. In particular, we will deposit our schema into the NIST Materials Data Curation System (MDCS), one project of the High-Throughput Experimental Materials Collaboratory in MGI, once MDCS becomes operational. MDCS will allow describing, depositing and sharing materials metadata in its repositories thanks to user-defined schema templates. The purpose is to create a nationwide federated network of materials data and experiments focused on all aspects from materials synthesis to characterization by sharing software tools, federating repositories, and interoperable metadata. However, providing open access to metadata schemas, as with MDCS, is only a first step towards interoperability. Adoption of software developed by other facilities appears sometimes easier than re-using metadata developed elsewhere.

No facility will have the expertise and resources to develop all the software necessary to cover the wide variety of experiments done at the NSLS-II. To facilitate community participation, NSLS-II has organized the first of a series of hackathons with other participants from other facilities and hosted the New Opportunities for Better User Group Software (NOBUGS) conference in 2018, a meeting aimed at promoting interactions between engineers and scientists working on software for X-ray, neutron and muon sources around the world. Collaborations through the Research Data Needs of the Photon and Neutron Science Community Interest Group (PaNSIG) at the Research Data Alliance is

a more formally organized effort aiming to support better data management and curation practices in large facilities. Focused on common scientific file formats, metadata, and persistent identifiers (including for instruments and samples), this effort points to the future for more robust data curation practices at experimental facilities, such as NSLS-II.

Embedding curation of raw and derived data at the beamline is a good step towards the preservation of valuable experimental results but not sufficient to ensure the reproducibility of final results. A wide dissemination of software developed for NSLS-II is well established through open versioning and software repositories, such as GitHub and bitbucket, but preservation is not guaranteed for the long term. Local policies ensuring a unified approach to curation at NSLS-II are in progress.

Acknowledgements

This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DESC0012704 with the U.S. Department of Energy. This research used resources of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. This work was partially supported as part of GENESIS: A Next Generation Synthesis Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award Number DE-SC0019212.

References

Arkilic, A., Allan, D., Caswell, T., Li, L., Lauer, K., & Abeykoon, S. (2017). Towards integrated facility-wide data acquisition and analysis at NSLS-II. *Synchrotron Radiation News, 30*(2), 44-45. doi:10.1080/08940886.2017.1289810

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *Nature Preceedings*. doi:10.1038/npre.2010.4626.1

Bicarregui, J., Matthews, B., & Schluenzen, F. (2015). PaNdata: Open data infrastructure for photon and neutron sources. *Synchrotron Radiation News, 28*(2), 30-35. doi:10.1080/08940886.2015.1013418

Boehnlein, A., Matthews, B., Proffen, T., & Schluenzen, F. (2015). The research data alliance photon and neutron science interest group. *Synchrotron Radiation News, 28*(2), 43-47. doi:10.1080/08940886.2015.1013421

Bunakov, V., Griffin, T., Matthews, B., & Cozzini, S. (2016). *Metadata for experiments in nanoscience foundries.* Paper presented at the International Conference on Data Analytics and Management in Data Intensive Domains. Ershovo, Moscow, Russia. doi:10.1007/978-3-319-57135-5_18

Green, M.L., Choi, C., Hattrick-Simpers, J., Joshi, A., Takeuchi, I., Barron, S., . . . Gregoire, J. (2017). Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews, 4*(1), 011105. doi:10.1063/1.4977487

Hall, S.R., & McMahon, B. (2005). *International tables for crystallography, definition and exchange of crystallographic data* (Vol. 8): Springer Science & Business Media.

Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., . . . Ceder, G. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials, 1*(1), 011002. doi:10.1063/1.4812323

Kalidindi, S.R., & De Graef, M. (2015). Materials data science: Current status and future outlook. *Annual Review of Materials Research, 45*, 171-193. doi:10.1146/annurev-matsci-070214-020844

Kleese Van Dam, K., Li, D., Miller, S.D., Cobb, J.W., Green, M.L., & Ruby, C.L. (2011). Challenges in data intensive analysis at scientific experimental user facilities. *Handbook of Data Intensive Computing* (pp. 249-284): Springer. doi:10.1007/978-1-4614-1415-5_10

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., & Kleese, K. (2010). Using a core scientific metadata model in large-scale facilities. *International Journal of Digital Curation, 5*(1), 106-118. doi:10.2218/ijdc.v5i1.146

Matthews, B., Crompton, S., Jones, C., & Lambert, S. (2015). Towards the preservation of the scientific memory. *International Journal of Digital Curation, 10*(1). doi:10.2218/ijdc.v10i1.361

Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation, 10*(2), 176-192. doi:10.2218/ijdc.v10i2.342

Pouchard, L., Juhas, P., Billinge, S., Wright, C.J., Campbell, S., Park, G., Stavitski, E., and Van Dam, H. (2019). Provenance Infrastructure for Multimodal X-Ray Experiments and Reproducible Analysis. In Kerstin Kleese Van Dam, Stuart Campbell, Kevin Yager, and Richard Farnsworth (Eds.), *Handbook on Big Data and Machine Learning in the Physical Sciences. Vol 2: Advanced Analysis Solutions for Leading Experimental Techniques*. World Scientific.

Yang, E., Matthews, B., & Wilson, M. (2013). Enhancing the core scientific metadata model to incorporate derived data. *Future Generation Computer Systems, 29*(2), 612-623. doi:10.1016/j.future.2011.08.003

Zakutayev, A., Wunder, N., Schwarting, M., Perkins, J. D., White, R., Munch, K., . . . Phillips, C. (2018). An open experimental database for exploring inorganic materials. *Scientific data, 5*, 180053. doi:10.1038/sdata.2018.53