JOURNAL OF SYNCHROTRON RADIATION
ISSN 1600-5775

Data management and processing workflow for the Materials Physics and Engineering group beamlines at the Advanced Photon Source

Jun-Sang Park,[a]* Connor Horn,[a,b] Prithvi Ramanathan,[a,c] Peter Kenesei[a] and Siniša Veseli[a]*

[a]Advanced Photon Source, Argonne National Laboratory, Lemont, IL 60439, USA, [b]Applied and Engineering Physics, Cornell University, Ithaca, NY 14853, USA, and [c]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. *Correspondence e-mail: parkjs@anl.gov, sveseli@anl.gov

Received 28 September 2018
Accepted 10 January 2019

Edited by V. Favre-Nicolin, CEA and Université Joseph Fourier, France

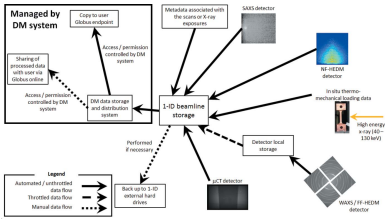
Keywords: data management; data processing workflow.

The ability to store, organize, process and distribute experimental data effectively, efficiently and securely is particularly important for large user facilities like the Advanced Photon Source. In this article, the deployment of the APS Data Management System (DM) at the 1-ID and 6-BM beamlines of the APS is described. These two beamlines support a wide range of experimental techniques and generate data at relatively high rates, making them ideal candidates to illustrate the deployment and customization of the DM system and its tools. Using several usage examples at these beamlines, various capabilities of the DM system are described.

1. Introduction

The ability to store, organize, process and distribute experimental data effectively, efficiently and securely is particularly important as user facilities, like the Advanced Photon Source (APS), host a growing number of users, develop more complex suites of experimental techniques, and generate larger volumes of experimental data. In the realm of synchrotron sources, the advances in detector technologies driving significant increases in data acquisition rates combined with increased photon flux and brilliance and advances in scanning techniques that can fully utilize these new detectors mean that an efficient and secure data management and processing system is critical for the success and sustainable operation of these facilities.

While a comprehensive review of the current state of data storage and archiving systems and data processing practices deployed at various synchrotron user facilities is beyond the scope of this work, it is worthwhile to note several developments at other user facilities. For data management and archiving solutions, the ICAT system (https://icatproject.org/) has been deployed at Diamond Light Source to manage experimental data (https://www.scd.stfc.ac.uk/Pages/Diamond-Light-Source.aspx). The Swiss Light Source at Paul Scherrer Institute in collaboration with the Swiss National Supercomputing Centre have implemented a large data storage system (https://www.cscs.ch/publications/press-releases/2018/589/). More recently, the National Synchrotron Light Source-II (NSLS-II) have deployed a comprehensive data acquisition and management system (http://nsls-ii.github.io/index.html) (Arkilic *et al.*, 2017). These solutions are often augmented by data processing or analysis workflow management systems. For instance, software packages like *scikit-beam* (https://www.bnl.gov/compsci/projects/structural-dynamics-software.



OPEN ACCESS

php) or the *Data Analysis Workbench* (Basham *et al.*, 2015; Filik *et al.*, 2017) framework can be integrated with data acquisition and management systems to allow real-time data analysis and provide feedback to the experimenters. These solutions have similar goals but are customized and developed to address the unique needs of each facility. There are also activities to join these types of efforts across institutions. The neutron scattering community, for instance, have developed a common analysis and visualization framework (http://www.mantidproject.org/Main_Page) (Arnold *et al.*, 2014) that is used by many institutions worldwide. More recently, major European user facilities proposed to build suitable data management and storage infrastructures so that the scientific data generated at those facilities can be accessed and shared more easily and re-used by the scientific community (https://panosc-eu.github.io/).

In this article, we present the APS Data Management System (DM) with an emphasis on the end-user experience. Details about the DM system architecture are presented elsewhere (Veseli *et al.*, 2018); here, we focus on the deployment and customization of the DM system for the 1-ID and 6-BM beamlines of the APS where a wide range of experimental techniques using high-energy X-rays are hosted. The experimental techniques include high-energy diffraction microscopy and high-energy micro-computed tomography (Park *et al.*, 2017), and generate sizable data sets at relatively high data rates, making these beamlines ideal candidates to demonstrate the DM system deployment and the end-user experience.

This article is organized as follows. An overview of the DM system is presented in §2. In §3, we provide an overview of the experimental techniques supported by the APS 1-ID and 6-BM beamlines and the data generation rates associated with these techniques. We highlight the need for a robust data management tool by describing how the experimental data were managed before the DM system deployment. We, then, proceed to provide an example of the deployed DM system at these beamlines, particularly highlighting the consolidation and distribution of large user data sets collected from *in situ* experiments with multiple detectors working simultaneously, if not concurrently (§3.1). We also describe an example of the DM workflow system deployed at the 1-ID beamline that is used to automate data processing (§3.2). In §4, we summarize the impact of the DM system and outline the developments that we are envisioning for the future.

2. Data management system architecture

As a US Department of Energy user facility, the APS operates many end-stations with a suite of experimental techniques to investigate a diverse set of material systems at multiple length scales. For this article, an *experiment* represents an object denoted by an arbitrary name that is associated with a collection of data files generated at a particular beamline or endstation during a block of allocated beam time. For the 1-ID and 6-BM beamlines, the arbitrary name of an experiment is generated using the name of the principal experimenter and the time of the allocated beam time. The terms *APS user* and *user* are used interchangeably with experimenter; an experimenter is associated with an experiment if the experimenter is part of the team that was granted beam time. This distinction is necessary to control which *experimenter* has access to which *experiment* data.

The DM system (Veseli *et al.*, 2018) developed at the APS offers a common framework and set of tools around which the beamlines can build their data acquisition, storage, distribution and processing workflows. Fig. 1 shows a schematic of the system. A more elaborate schematic describing the components of each service is presented in Veseli *et al.* (2018). While the DM system consists of many components, it can be broadly broken down into two main functional parts — DM Software and DM Storage. These are highlighted in green in Fig. 1 and they are built to interact with the existing APS systems like the data acquisition (DAQ) systems that are already in place for experiments (highlighted in yellow in Fig. 1).

The components of the DM Software can be classified into one of the following categories:

(i) APS-wide services. These services include the central DM database (DB) that keeps track of all APS experiments and experimenter identifiers. These identifiers can be managed by a suite of administrative tools. APS-wide services also include the storage management service responsible for managing data files and their access permissions. For example, when a new file is transferred into the DM Storage the storage management service ensures that the file is associated with the

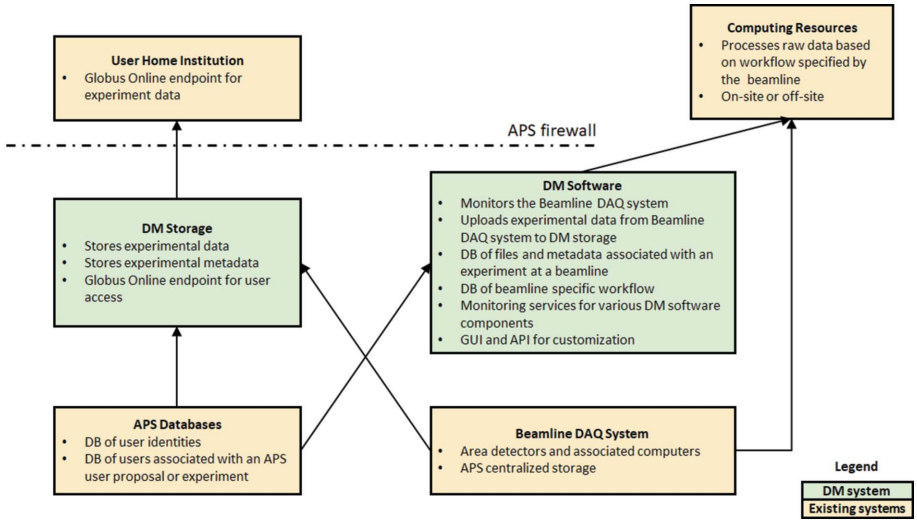


Figure 1
Schematic of the APS DM system. A more detailed version of this diagram has been given by Veseli *et al.* (2018).

correct experiment. It also ensures that proper permissions are applied to that file which enables remote access to the file for authorized users.

(ii) Beamline-specific services. These services consist of the Data Acquisition (DAQ) service, beamline metadata catalogue service, and workflow processing service. The DAQ service is responsible for monitoring a designated file system location and transferring new or modified data files to the DM Storage. The beamline metadata catalogue service keeps track of the metadata associated with a data file such as its file size, MD5 checksum value, and original location in the file system before its transfer to the DM Storage. This metadata catalogue currently does not include scanning or X-ray exposure related information. The workflow processing service manages and executes beamline-defined workflows where a workflow is a set of processing steps that are executed to reduce or analyze experimental data. It is capable of submitting jobs to available computing resources and monitoring their progress.

(iii) Monitoring services. Every DM service has built-in monitoring interfaces that enable external applications to find out more about its state (*e.g.* operational status of a service, number of files in the queue). Those interfaces are used by custom Nagios (https://www.nagios.org/) plug-ins built for the DM system.

(iv) User interfaces. The DM users and administrators can interact with the DM system via a web portal, a desktop graphical user interface (GUI), or full set of command line tools, as well as via Python and Java Application Programming Interfaces (APIs). The command line tools can be used as the building blocks to generate a set of customized tools to match the needs of a particular beamline. Fig. 2 shows an example of DM Python API usage. Here, the Python script is tasked with retrieving the list of experiment files associated with a particular experiment named `smith_apr18`. More information about these interface options are available at https://confluence.aps.anl.gov/display/DMGT/Data+Management and in Veseli *et al.* (2018).

```
api = FileCatApi(username, password)
files = api.getExperimentFiles(experimentName='smith-Jun2018')
for file in files:
    print('%s: %s,%s' % (file['fileName'],file['fileSize'],file['md5Sum']))
```

Figure 2
Example of DM API usage. More examples are available at https://confluence.aps.anl.gov/display/DMGT/Data+Management.

The DM Software is written in Python (REST web services and clients) and Java (user Web Portal). It is built around modern open source databases (MongoDB, https://www.mongodb.com; PostresSQL, https://www.postgresql.org), CherryPy web framework, http://cherrypy.org) and APIs (Java Persistence API and JavaServer Faces, http://www.oracle.com/technetwork/java/javaee/overview/index.html with Primefaces, http://primefaces.org).

The DM storage uses a 1.5 PB[1] Data Direct Networks (DDN) storage system with a high-performance GPFS file system (Schmuck & Haskin, 2002). The storage system has data redundancy enabled. The DM storage has a 10 Gbps network bandwidth capacity with two redundant network links. Read-only data access is provided via the *aps#data* Globus Online (Foster, 2011; Allen *et al.*, 2012) endpoint. User authentication is handled by the Globus Online MyProxy and Lightweight Directory Access Protocol (LDAP) servers. Periodic hardware updates are envisioned to keep pace with increasing data rates; for example, the DM storage will be updated in early 2019 to a system with larger file system (from 1 PB storage to 4 PB storage) and bandwidth capacity.

User authorization to access experimental data files is based on experiment LDAP group membership. Each experiment in the DM system is associated with a corresponding LDAP group that consists of users with access to the experiment data via *aps#data* Globus Online endpoint. The DM beamline managers can use the APS Experiment Safety Authorization Form database, APS General User Proposal database, or a simple list of DM user names to control the experimenter membership for an experiment.

3. DM deployment at the APS 1-ID and 6-BM beamlines

The APS Materials Physics and Engineering (MPE) group operates the 1-ID and 6-BM beamlines[2], which are used to investigate a wide range of polycrystalline materials (Park *et al.*, 2017). The 1-ID beamline provides high-energy monochromatic X-rays to support several experimental techniques such as micro-computed tomography (μ-CT), wide-angle X-ray scattering (WAXS), small-angle X-ray scattering (SAXS), and near- and far-field high-energy diffraction microscopy (NF- and FF-HEDM) (Wang *et al.*, 2003; Haeffner *et al.*, 2005; Suter *et al.*, 2006; Lienert *et al.*, 2011). The 6-BM beamline provides polychromatic X-rays to support energy-dispersive diffraction (EDD) and μ-CT. With the exception of the EDD technique, several types of area detector systems are employed for these techniques. Table 1 summarizes the available detectors and approximate data rates. It is worth comparing the WAXS and FF-HEDM techniques that use the same detector. While techniques such as diffraction/scattering tomography (Stock *et al.*, 2008) exist, WAXS is predominantly a grain-averaging technique requiring minimal sample rotation and data. FF-HEDM, on the other hand, is a grain-resolving technique that requires significant sample rotation and data. Hence, a WAXS data set is typically smaller than a FF-HEDM data set.

At each beamline, these experimental techniques are frequently conducted simultaneously to obtain complementary information about the sample at different length scales. They are also often combined with external stimuli such as thermo-mechanical loading to investigate the induced changes to the sample (Colas *et al.*, 2010; Varlioglu *et al.*, 2010; Shade *et al.*, 2015; Chatterjee *et al.*, 2016; Zhang *et al.*, 2016). Table 1

[1] As of August 2018.

[2] For the 6-BM beamline, the MPE group operates the 6-BM-A experimental hutch.

Table 1
Summary of detector systems used for experimental techniques at the APS 1-ID and 6-BM beamlines. The software listed here are examples of ones that can be used for the associated techniques and by no means exhaustive.

Experimental technique	Detector characteristics	Maximum data rate	Example analysis tools
WAXS and FF-HEDM	Four panels of 409.6 mm × 409.6 mm active area / 2048 × 2048 pixels / 14-bit pixel	Each panel generates 8 MB at 7 Hz (56 MB s^{-1}) / four panels generate 32 MB at 7 Hz (224 MB s^{-1})	Custom scripts, *Fit2d* (Hammersley, 1995, 2016), *GSAS2* (Toby & Von Dreele, 2013) and *MAUD* (Lutterotti, 2010) for WAXS; *HEXRD* (Bernier *et al.*, 2011), *MIDAS* (Sharma *et al.*, 2012*a*,*b*) and *Fable* package (Schmidt, 2014) for FF-HEDM
WAXS and FF-HEDM	One panel of 290.8 mm × 229.8 mm active area / 3888 × 3072 pixels / 14-bit pixels	23 MB at 10 Hz (230 MB s^{-1})	Custom scripts, *Fit2d* (Hammersley, 1995, 2016), *GSAS2* (Toby & Von Dreele, 2013) and *MAUD* (Lutterotti, 2010) for WAXS; *HEXRD* (Bernier *et al.*, 2011), *MIDAS* (Sharma *et al.*, 2012*a*,*b*) and *Fable* package (Schmidt, 2014) for FF-HEDM
SAXS	One panel of 62 mm × 25 mm active area / 1024 × 476 pixels / 15-bit pixels	0.8 MB at 100 Hz (80 MB s^{-1})	Custom scripts and *IRENA* (Ilavsky & Jemian, 2009)
NF-HEDM	One panel of 3.03 mm × 3.03 mm active area / 2048 × 2048 pixels / 12-bit pixels	8 MB at 4 Hz (32 MB s^{-1})	*Ice9* (Suter *et al.*, 2006) and *MIDAS* (Sharma *et al.*, 2012*a*,*b*)
μ-CT	One panel of 2.2 mm × 1.4 mm active area / 1920 × 1200 pixels / 12-bit pixels	4 MB at 100 Hz (400 MB s^{-1})	Custom scripts (Khounsary *et al.*, 2013), *TomoPy* (Gürsoy *et al.*, 2014), *Tomo-Recon* (Rivers, 2012)

also summarizes some of the data analysis tools that are available for respective techniques. The complexity of the experimental techniques and high data rates as well as expanding user base highlight the need for a streamlined data management and processing workflow mechanism. The APS DM system meets these needs; in the following sections, we describe the deployment and customization of the DM system for the 1-ID and 6-BM beamlines. We describe the data management and data processing workflows in §3.1 and §3.2, respectively.

3.1. Data management workflow at 1-ID and 6-BM beamlines

Fig. 3 shows the movement of experimental data for a particular experiment before the deployment of the DM system at the 1-ID beamline in 2015. Prior to 2015, the data from the suite of detectors were stored directly to a beamline storage system dedicated for 1-ID operation (approximately 30 TB in size) accessible through the APS network with underlying backup system which makes incremental backups of the data as they are created in the storage. For some area detectors (particularly for the WAXS/FF-HEDM detectors), their data rates were throttled to allow this 1-ID storage to keep up with the data generation rate. After the completion of an experiment, the users manually copied their data to a set of external hard drives. Because this beamline storage was approximately 30 TB in size, the user data were also backed up to external hard drives that resided with the MPE staff. The users and beamline staff predominantly shared processed data for evaluation and analysis through electronic mail and a File Transfer Protocol location.

Improvements to detector drivers, storage technology and network infrastructure at the APS allow 1-ID users to collect data at the maximum rates, particularly for the WAXS/FF-HEDM detector array. In this case, the area detector data are temporarily stored on detector local (solid state) drives

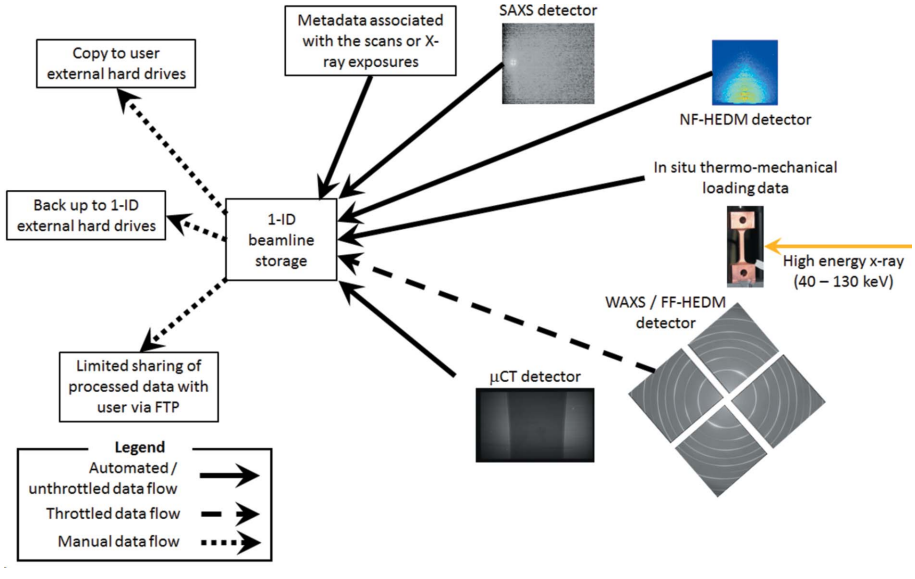


Figure 3
Schematic of the old data flow at the APS 1-ID beamline.

(approximately 1 TB in size), initially. Then the data files are moved using background processes to the beamline storage system. The metadata associated with the area detector data (such as scanning parameters and imposed stimulus on the sample) are also stored on the beamline storage system in a tabular form.

In conjunction with these improvements, the DM system was first deployed at the 1-ID beamline in 2015 to streamline the consolidation and distribution of raw and processed data as well as creating archives of these data.[3] Fig. 4 shows the data flow after the deployment of the DM system at the 1-ID beamline. At the onset of an experiment, a data directory associated with the experiment is created on the beamline storage system. All area detector data are piped from respective local drives to this data directory through background processes. The data directory on the beamline storage system is monitored by the DM data acquisition service. As new area detector data files show up in this directory, they are uploaded to the DM Storage in quasi real-time. As the data are uploaded, they are also catalogued by the DM beamline metadata catalogue service. Pre-processed data such as intensity versus plane spacing data (which can still be large but significantly smaller than the raw area detector data) can also be uploaded to the DM Storage if the users choose to take advantage of data processing workflows during their experiment.

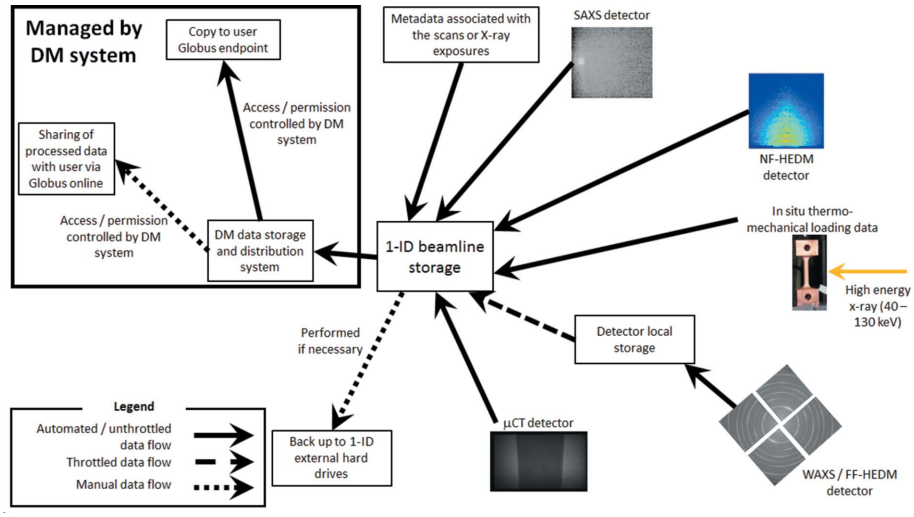


Figure 4
Schematic of the new data flow at the APS 1-ID beamline.

In addition to standard features of the DM system, a set of customized command line scripts for the 1-ID and 6-BM beamlines was built on top of the the DM command line tools. These have the following functions to address the specific needs of the 1-ID and 6-BM beamlines and highlights the flexibility of the DM system:

(i) Consolidates the data that were initially stored in various locations (local drives) to the 1-ID or 6-BM beamline storage at the end of an experiment.

(ii) Uploads the consolidated data set from the beamline storage to the DM Storage. While the data files are uploaded to the DM Storage in quasi real-time during the user experiment, the consolidation and upload of the full data set at the end of the experiment ensures that all the files are accounted for.

(iii) Checks the integrity of the data set by comparing the full data set in the APS centralized storage (original) and uploaded data set in the data distribution system (copy) so that the original can be deleted to make space for the next set of experiments. To check the integrity of the data, the MD5 checksum value for each original file is compared with the copy's MD5 checksum value stored in the DM catalogue. In the rare case where the MD5 checksum values of the originals and copies are different, the beamline staff are alerted so that the discrepancy can be reconciled before the original is removed.

(iv) Compresses the data set to reduce its footprint in the DM data distribution system. The *bzip2* (http://www.sourceware.org/bzip2/) compression program was selected after performing tests on different compression programs to assess their performances on representative experimental data sets collected at the 1-ID beamline. Using this compression program, the approximate reduction in file size for typical WAXS data was 70%, SAXS was 80%, FF-HEDM data was 75%, and tomography data was 60%.

Since the deployment of the DM system in 2015, the 1-ID beamline has added more than 230 user experiments with a total of more than 340 TB of user experiment data.[4] For the 6-BM beamline, approximately 40 user experiments were added since the deployment of the DM system in 2017; they used 383 GB and 666 GB of DM storage space in 2017 and 2018, respectively.[4] With the tomographic imaging capabilities coming online at the 6-BM beamline in 2018, it is anticipated that more data will be collected at the 6-BM beamline in the future.

For the 1-ID beamline and 6-BM beamline users, the majority of the data sets were accessed using Globus Online with endpoints predominantly located in North America, Europe and Asia. The deployment of the DM system at the beamline has streamlined and simplified the users' access to their experimental data. In the past, a user group made a copy of the data (several TB large) to external hard drives to carry

[3] The APS does not have an official data retention policy as of August 2018. However, the DM system is able to provide some form of back-up capabilities. We anticipate that the data store in the DM system will be retained for at least a year and more with data compression and DM storage expansions. As the data become obsolete or not actively used, they can be transferred to a tape archive. Published data can be retained with archiving systems like the Materials Data Facility (Blaiszik *et al.*, 2016).

[4] Information as of August 2018.

back with them or made arrangements to be shipped. This process was slow, not secure, and needed significant human intervention and efforts. Furthermore, the user group would make copies of the copy to the computers used for analysis at their home institutions. If a user group were a collaboration of multiple institutions, multiple copies of the raw data had to be made and shipped separately.

With the new DM system at the beamline, the experimental data can be accessed by the users with appropriate permissions through the DM system (Veseli *et al.*, 2018) and Globus Online. Three layers of authentication and permission checks are necessary to access the data files generated from a particular experiment:

(i) An experimenter needs to have access to Globus Online through an institutional authentication or Globus ID authentication.

(ii) The experimenter also needs to have an APS user identification to access the DM storage system. This is provided to all experimenters who have allocated APS beam time and have site access to Argonne National Laboratory.

(iii) The experimenter needs to be associated with a particular experiment that generated the data files. The association between a particular experiment and an experimenter is controlled by the beamline staff using the APS General User Proposal database and APS Experimental Safety Assessment Form database. These APS databases hold the identifiers of the experimenters associated with a particular experiment.

When an experimenter satisfies all three conditions, only then the experimenter can access and download the data; the ability to change the data stored in the DM storage system is not granted to the experimenters. With a prearranged Globus Online endpoint at their home institutions, the off-site experimenters can access the data over the Internet for visualization or analysis. They can interact with the on-site experimenters and participate in the experiment. The on-site experimenters can also initiate the data download before leaving the APS and have the full data set waiting for them at their home institutions once they return to their home institutions. It is also worth noting that experimenters can take advantage of various Globus Online transfer options to satisfy their time and security needs. Fig. 5 shows an instance of user access to the data in the DM storage.

The DM catalogue of the experiments can provide useful information for the beamline staff and administrators. Figs. 6 and 7 show the number of allocated user groups and the data storage usage since calendar year 2015 for the 1-ID beamline. Note that the DM system was first installed and tested at the 1-ID beamline in late 2015 and the statistics in these two figures are as of August 2018. Fig. 6 shows that approximately 60 user groups visited the 1-ID beamline each year since 2016. Of these, approximately 55% of the user groups performed WAXS/SAXS type of experiments and approximately 45% of the user groups performed HEDM type of experiments. Fig. 7 shows that the 1-ID beamline collected approximately 120 TB of data each year between year 2016 and 2017 and will most likely hit that level in 2018 as well. Looking at the two figures, HEDM experiments produce relatively larger data sets when

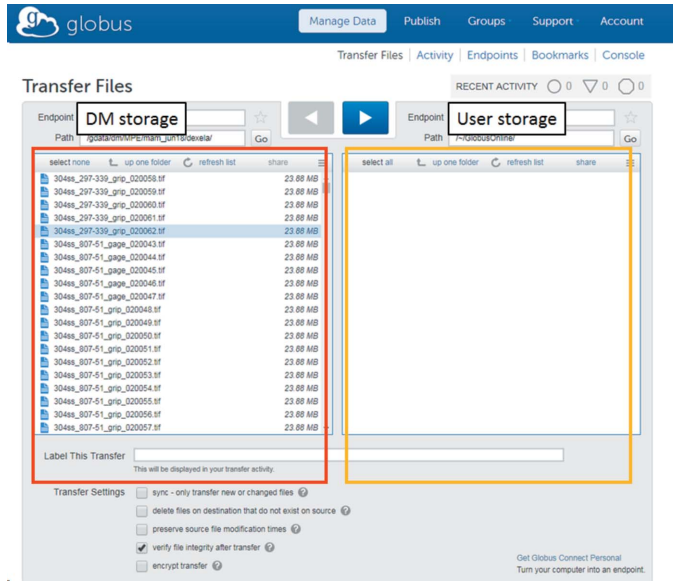


Figure 5
Example of user access to the DM storage and data using the Globus Online website. A user can only browse and download his/her data and access permissions are regulated by the DM system.

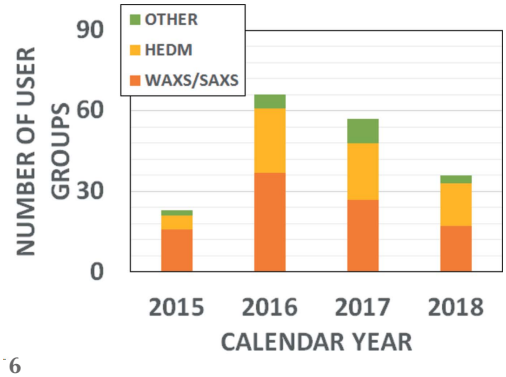


Figure 6
Number of allocated user groups since calendar year 2015 for the 1-ID beamline obtained from the DM system catalogue. Note that the DM system was first installed at the 1-ID beamline in 2015. The plotted data are as of August 2018 and we anticipate similar user group numbers for 2018 as in 2017.

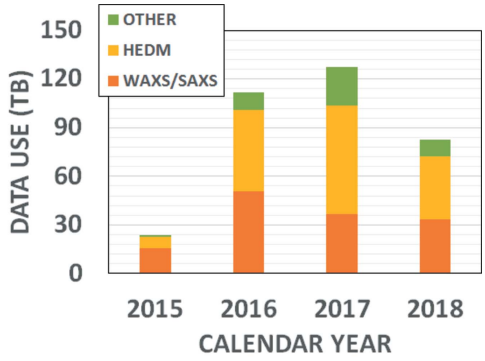


Figure 7
Volume of experimental data collected since calendar year 2015 at the 1-ID beamline obtained from the DM system catalogue. Note that the DM system was first installed at the 1-ID beamline in 2015. The plotted data are as of August 2018 and we anticipate similar data levels for 2018 as in 2017.

we consider that the HEDM experiments and the WAXS/SAXS experiments are allocated approximately evenly. As alluded to in §3, the WAXS technique is not as data intensive as the FF-HEDM technique even though they use the same detector and the trends observed in Figs. 6 and 7 are consistent with this expectation. Such quantitative information can be furnished with the use of the DM system combined with appropriate searchable keywords for the DM database. One benefit of such information is that computing infrastructure for future beamlines and facilities hosting similar experiments or techniques can be estimated and planned.

3.2. Data processing workflow for WAXS at 1-ID

A DM workflow is defined as a set of processing steps that are executed in order. Each step involves an arbitrary command or a script, and can contain both input and output arguments. Whenever possible, the DM Processing Service automatically parallelizes execution of those steps (Veseli *et al.*, 2018).

There are numerous analysis software and scripts available to analyze the area detector data collected at the 1-ID beamline as illustrated in Table 1. The DM workflow tool allows us to deploy these to take advantage of various computing resources available to the users at the APS ranging from local workstation with multiple cores, local high-performance computing cluster supported by the APS, high-performance computing resources available around the nation, including systems at the Argonne Leadership Computing Facility. The goal of the DM workflow tool is not to (re)invent another analysis tool. As listed in Table 1, there are already many software packages suitable for various experimental techniques. The goal of the DM workflow tool is to provide an interface such that existing scriptable analysis tools can be sequenced and automated to read necessary inputs (such as detector calibration or experiment configuration information) and executed on available or allocated computing resources. In this sense, the approach we take is similar to ones pursued by the EDNA framework (Incardona *et al.*, 2009), Directly Programmable Data Analysis Kit (DPDAK) (Benecke *et al.*, 2014), and Information System for Protein crystallography Beamlines (ISPyB) (Delagenire *et al.*, 2011) to name a few other similar developments.

We demonstrate use of the DM workflow tool for a set of MATLAB scripts used to process and analyze WAXS data. Typically, the raw WAXS area detector data are processed using the following steps (Fig. 8):

(i) Extract individual frames from the area detector data file that contains multiple frames in a single file. These individual frames can be summed for better statistics or have time-dependent information taken at up to 7 Hz.

(ii) Subtract dark field and remove bad pixel data from each frame.

(iii) Integrate to produce intensity versus lattice spacing data.

(iv) Fit the peaks of interest using a suitable peak profile function.

These steps can be automated into a workflow and computation jobs can be submitted to the computing resources available to the 1-ID users.

In our WAXS workflow example, a set of MATLAB scripts for these procedures were available (https://github.com/junspark/matlab_tools) and have been used in various experiments measuring strain pole figures and preferred crystallographic orientation (Miller *et al.*, 2008; McNelis *et al.*, 2013; Park *et al.*, 2013, 2016). These scripts need a set of parameters (such as image names, corresponding dark field file name, and detector calibration information). For each procedure, a wrapper command line script with the following functionality was generated:

(i) Gathers and checks the input parameters.

(ii) Connects to a computer available to the users (in this case a high-end workstation at the beamline).

(iii) Launches MATLAB (or MATLAB Runtime if MATLAB is not available at the designated computer).

(iv) Runs the job based on the parameters.

Users generate the parameters and submit the job using a graphical user interface (Fig. 9). Each job submission obtains a unique alphanumeric identifier to keep track of its progression. Successful or unsuccessful completion of a job are distinguished by the output message upon termination of the job. For each job submission, its progress, job summary and termination messages are piped to a text file for debugging or data provenance needs. The reduced data output (in this case several diffractograms from different frames in an area detector file)

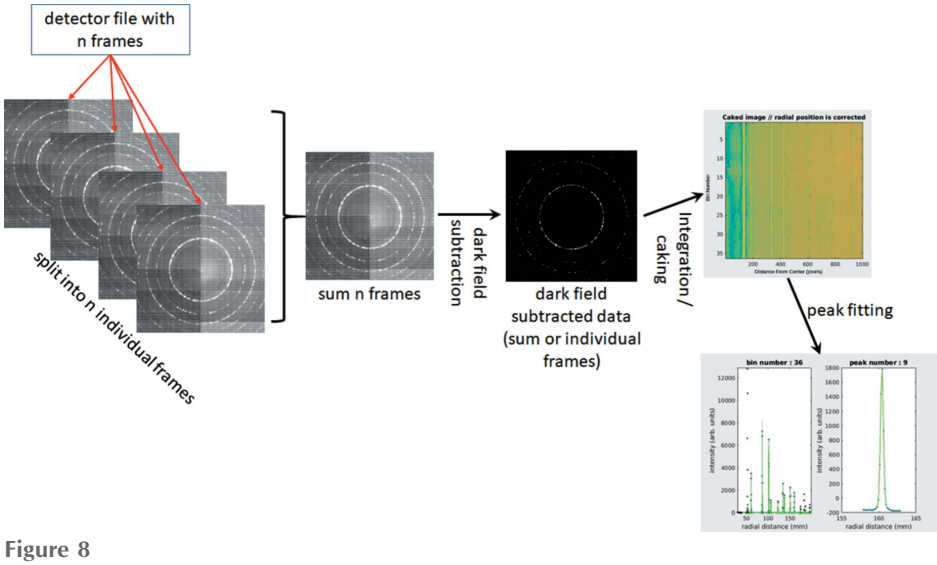


Figure 8
Schematic of the WAXS data processing workflow.

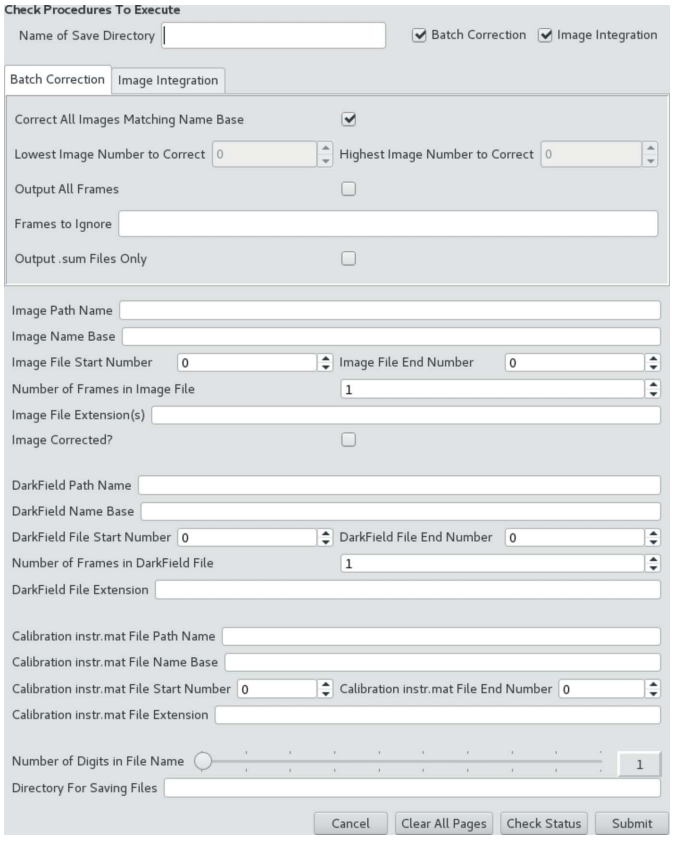


Figure 9
Example instance of the GUI for the WAXS data processing workflow.

contains the following to keep track of the data provenance:

(i) Intensity versus plane spacing data.

(ii) X-ray exposure conditions such as exposure time and intensities of the incoming and outgoing X-ray beam measured by ion chambers or diodes.

(iii) Scripts and tools used to reduce the data and associated calibration parameters.

(iv) State of the sample such as applied stimulus information.

(v) Sample translation and rotation information.

While this is a simple example that uses pre-existing MATLAB scripts, the idea shown here can be expanded to data processing or analysis scripts or routines written in any programming language that can be launched from a command line and is available at the desired computing resource.

4. Summary and future work

In this article, we demonstrated the deployment of the APS DM system to the 1-ID and 6-BM beamlines of the APS. In §3.1, the data management and distribution capabilities were illustrated. The DM system combined with the well planned infrastructure and data acquisition strategy can provide a robust, efficient data management architecture. The DM system's customization tools can be used as building blocks to enhance its functionalities and meet the needs of individual beamlines. The DM system's cataloguing capabilities can be manipulated to provide various statistics as long as appropriate keyword-value pairs used. In §3.2, the worflow capabilities of the DM system were demonstrated using a set of existing WAXS data reduction scripts in MATLAB. With the addition of simple wrapper scripts, these existing MATLAB scripts could be executed via the DM workflow service to reduce the area detector data into diffractograms.

While the initial deployment of the DM system at the 1-ID and 6-BM beamlines was successful, there are several areas of improvements.

(i) Deploying other data reduction and analysis tools (tabulated in Table 1) to the DM workflow service. An example of WAXS workflow was illustrated §3.2. Table 1 lists a suite of other experimental techniques and associated data reduction and analysis tools. We plan to add these data reduction and analysis tools to the DM data processing workflow in the near future.

(ii) Automated data reduction using DM workflow services. The ability to obtain results using real-time data analysis — albeit preliminary results — is particularly useful with *in situ* measurements where decisions on the applied stimuli must often be made in real time. This ability is also useful for high-throughput setups such as ones being used or planned at several APS beamlines. At these high-throughput setups, the data acquisition is envisioned to be automated to interrogate many samples. A key component for the success of these setups is how efficiently the large volumes of raw data can be reduced and analyzed. Furthermore, this ability provides more standardized data reduction and analysis procedures, thereby lowering the barrier for new users and reducing the instances of misinterpretation or overinterpretation of the experimental data.

(iii) Reducing data footprint. The area detector data generated from many of the techniques described in Table 1 are sparse. In the case of FF-HEDM, for example, the diffraction spots only occur when the diffraction conditions are met and diffraction spots only show up at very specific positions on the area detector. The rest of the pixels in the area detector typically contain no useful data. Taking advantage of this, various data compression strategies can be considered to reduce the data footprint. For example, pre-processing the WAXS area detector data with a dark field subtraction can significantly increase the compression ratio. In the case of FF-HEDM, one can also consider storing only the pixel intensities associated with Debye–Scherrer rings if the data processing and analysis workflow is robust.

(iv) Longer term data management. Simply adding more storage space to the DM Storage or compressing experimental data is insufficient. We are investigating a tiered approach such that older or published data are stored in slower but more cost-effective storage systems like tape drive systems and newer or unpublished data are stored in faster storage system with better accessibility. Data sets (both raw and processed/analyzed) that have been used in publication(s) can be transferred to other archiving systems like the Materials Data Facility (Blaiszik *et al.*, 2016).

Acknowledgements

The authors acknowledge the APS-IT group working behind the scenes to allow these different systems to work seamlessly. JSP would like to thank the APS 1-ID and 6-BM user communities for transitioning to the new DM system and their patience during that transition.

Funding information

Funding for this research was provided by: US Department of Energy, Office of Science (contract No. DE-AC02-06CH11357).

References

Allen, B., Pickett, K., Tuecke, S., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M. & Martin, S. (2012). *Commun. ACM*, **55**, 81.

Arkilic, A., Allan, D. B., Caswell, T., Li, L., Lauer, K. & Abeykoon, S. (2017). *Synchrotron Radiat. News*, **30**(2), 44–45.

Arnold, O., Bilheux, J., Borreguero, J., Buts, A., Campbell, S., Chapon, L., Doucet, M., Draper, N., Ferraz Leal, R., Gigg, M., Lynch, V., Markvardsen, A., Mikkelson, D., Mikkelson, R., Miller, R., Palmen, K., Parker, P., Passos, G., Perring, T., Peterson, P., Ren, S., Reuter, M., Savici, A., Taylor, J., Taylor, R., Tolchenov, R., Zhou, W. & Zikovsky, J. (2014). *Nucl. Instrum. Methods Phys. Res. A*, **764**, 156–166.

Basham, M., Filik, J., Wharmby, M. T., Chang, P. C. Y., El Kassaby, B., Gerring, M., Aishima, J., Levik, K., Pulford, B. C. A., Sikharulidze, I., Sneddon, D., Webber, M., Dhesi, S. S., Maccherozzi, F., Svensson, O., Brockhauser, S., Náray, G. & Ashton, A. W. (2015). *J. Synchrotron Rad.* **22**, 853–858.

Benecke, G., Wagermaier, W., Li, C., Schwartzkopf, M., Flucke, G., Hoerth, R., Zizak, I., Burghammer, M., Metwalli, E., Müller-Buschbaum, P., Trebbin, M., Förster, S., Paris, O., Roth, S. V. & Fratzl, P. (2014). *J. Appl. Cryst.* **47**, 1797–1803.

Bernier, J. V., Barton, N. R., Lienert, U. & Miller, M. P. (2011). *J. Strain Anal. Eng. Des.* **46**, 527–547.

Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S. & Foster, I. (2016). *JOM*, **68**, 2045–2052.

Chatterjee, K., Venkataraman, A., Garbaciak, T., Rotella, J., Sangid, M., Beaudoin, A., Kenesei, P., Park, J.-S. & Pilchak, A. (2016). *Int. J. Solids Struct.* **94–95**, 35–49.

Colas, K., Motta, A., Almer, J., Daymond, M., Kerr, M., Banchik, A., Vizcaino, P. & Santisteban, J. (2010). *Acta Mater.* **58**, 6575–6583.

Delagenire, S., Brenchereau, P., Launer, L., Ashton, A. W., Leal, R., Veyrier, S., Gabadinho, J., Gordon, E. J., Jones, S. D., Levik, K. E., McSweeney, S. M., Monaco, S., Nanao, M., Spruce, D., Svensson, O., Walsh, M. A. & Leonard, G. A. (2011). *Bioinformatics*, **27**, 3186–3192.

Filik, J., Ashton, A. W., Chang, P. C. Y., Chater, P. A., Day, S. J., Drakopoulos, M., Gerring, M. W., Hart, M. L., Magdysyuk, O. V., Michalik, S., Smith, A., Tang, C. C., Terrill, N. J., Wharmby, M. T. & Wilhelm, H. (2017). *J. Appl. Cryst.* **50**, 959–966.

Foster, I. (2011). *IEEE Internet Comput.* **15**, 70–73.

Gürsoy, D., De Carlo, F., Xiao, X. & Jacobsen, C. (2014). *J. Synchrotron Rad.* **21**, 1188–1193.

Haeffner, D., Almer, J. & Lienert, U. (2005). *Mater. Sci. Eng. A*, **399**, 120–127.

Hammersley, A. P. (1995). ESRF Internal Report ESRF97HA02T. ESRF, Grenoble, France.

Hammersley, A. P. (2016). *J. Appl. Cryst.* **49**, 646–652.

Ilavsky, J. & Jemian, P. R. (2009). *J. Appl. Cryst.* **42**, 347–353.

Incardona, M.-F., Bourenkov, G. P., Levik, K., Pieritz, R. A., Popov, A. N. & Svensson, O. (2009). *J. Synchrotron Rad.* **16**, 872–879.

Khounsary, A., Kenesei, P., Collins, J., Navrotski, G. & Nudell, J. (2013). *J. Phys. Conf. Ser.* **425**, 212015.

Lienert, U., Li, S. F., Hefferan, C. M., Lind, J., Suter, R. M., Bernier, J. V., Barton, N. R., Brandes, M. C., Mills, M. J., Miller, M. P., Jakobsen, B. & Pantleon, W. (2011). *JOM*, **63**, 70–77.

Lutterotti, L. (2010). *Nucl. Instrum. Methods Phys. Res. B*, **268**, 334–340.

McNelis, K., Dawson, P. & Miller, M. (2013). *J. Mech. Phys. Solids*, **61**, 428–449.

Miller, M., Park, J.-S., Dawson, P. & Han, T.-S. (2008). *Acta Mater.* **56**, 3927–3939.

Park, J.-S., Lienert, U., Dawson, P. R. & Miller, M. P. (2013). *Exp. Mech.* **53**, 1491–1507.

Park, J.-S., Okasinski, J., Chatterjee, K., Chen, Y. & Almer, J. (2017). *Synchrotron Radiat. News*, **30**(3), 9–16.

Park, J.-S., Ray, A. K., Dawson, P. R., Lienert, U. & Miller, M. P. (2016). *J. Strain Anal. Eng. Des.* **51**, 358–374.

Rivers, M. L. (2012). *Proc. SPIE*, **8506**, 85060U.

Schmidt, S. (2014). *J. Appl. Cryst.* **47**, 276–284.

Schmuck, F. & Haskin, R. (2002). *Proceedings of the FAST 2002 Conference on File and Storage Technologies*, 28–30 January 2002, Monterey, CA, USA, pp. 231–244.

Shade, P. A., Blank, B., Schuren, J. C., Turner, T. J., Kenesei, P., Goetze, K., Suter, R. M., Bernier, J. V., Li, S. F., Lind, J., Lienert, U. & Almer, J. (2015). *Rev. Sci. Instrum.* **86**, 093902.

Sharma, H., Huizenga, R. M. & Offerman, S. E. (2012*a*). *J. Appl. Cryst.* **45**, 693–704.

Sharma, H., Huizenga, R. M. & Offerman, S. E. (2012*b*). *J. Appl. Cryst.* **45**, 705–718.

Stock, S. R., De Carlo, F. & Almer, J. (2008). *J. Struct. Biol.* **161**, 144–150.

Suter, R. M., Hennessy, D., Xiao, C. & Lienert, U. (2006). *Rev. Sci. Instrum.* **77**, 123905.

Toby, B. H. & Von Dreele, R. B. (2013). *J. Appl. Cryst.* **46**, 544–549.

Varlioglu, M., Lienert, U., Park, J.-S. & Jones, J. L. (2010). *Text. Stress Microstruct.* **2010**, 1–10.

Veseli, S., Schwarz, N. & Schmitz, C. (2018). *J. Synchrotron Rad.* **25**, 1574–1580.

Wang, X.-L., Almer, J., Liu, C. T., Wang, Y. D., Zhao, J. K., Stoica, A. D., Haeffner, D. R. & Wang, W. H. (2003). *Phys. Rev. Lett.* **91**, 265501.

Zhang, X., Almer, J., Benda, E., Kenesei, P., Mashayekhi, A., Park, J.-S., Westferro, F., Chen, Y., Li, M., Wang, L. & Xu, C. (2016). *Rev. Sci. Instrum.* **88**, 015111.