ISBN: 978-3-95450-237-0 ISSN: 2673-5512 doi:10.18429/JACoW-PCaPAC2022-FR023

EXPERIMENTAL DATA COLLECTION STANDARDS AT SESAME SYNCHROTRON

M. Alzubi*, A. Abbadi, A. Al-Dalleh, A. Aljadaa, A. Lausi, A. Mohammad, B. Aljamal, G. Iori, G. Kamel, M. Abdellatief, M. Genisel, M. Harfouche, R. khrais, S. Matalgah, Y. Momani
SESAME Synchrotron, Allan, Jordan

Abstract

Experimental data collection is the essential process of acquiring experimental raw data along with its associated metadata from SESAME beamlines. For data collection and processing; scanning modes, data and metadata formats, and data visualisation are only a few aspects in which individual beamlines differ from each other. In addition, the volume of experimental datasets every experimental day might range from a few gigabytes to many terabytes. Herein, the effectiveness of the experiments being conducted at SESAME depends heavily on the efficiency and reliability with which experimental data are collected. Each beamline at SESAME has its own Data Acquisition (DAQ) system that is intended and being developed primarily to meet beamline users' and scientists' expectations. It also ensures that experimental raw data and metadata are not randomly generated and are stored together in a stander and well-defined file formats in compliance with SESAME Experimental Data Management Policy. In this paper, we present the standards and features employed in SESAME's DAQ systems, as well as the experimental data creation, curation, storage, and accessibility pipeline currently being built for SESAME beamlines.

INTRODUCTION

SESAME [1] is a third-generation synchrotron light that has currently three beamlines in operation and serves the SESAME users' community; the XAFS/XRF (X-ray Absorption Fine Structure/X-ray Fluorescence) spectroscopy [2], MS (Materials Science) and IR (Infrared Spectromicroscopy) beamlines [3]. Two more beamlines are being installed and will be commissioned soon, namely, HESEB (HElmholtz-SEsame Beamline) [4] and BEATS (BEAmline for Tomography at SESAME) [5,6].

At SESAME, the first monochromatic beam was obtained in November of 2017 at the XAFS/XRF beamline. In August of 2018, the first user experiment was conducted on the same beamline. In June of 2020, the SESAME Council adopted its "Experimental Data Management Policy," which is a deliverable of the H2020 BEATS project that is a part of the EU research and innovation funding programme. This policy is harmonised based on the European Synchrotron Radiation Facility (ESRF) and PaNData data policy frameworks [7]. Late in the same year, the organisational structure of SESAME was reformulated and a new dedicated team was created, under the supervision of the scientific director, to obtain the experimental data from the beamlines in compliance with data policy. In January of 2021, the Data Collection and Analysis (DCA) team was founded with these objectives: i) enabling the beamlines' DAQ systems to generate experimental data aligned with SESAME's Data Management Policy, ii) increasing the productivity of experiments, iii) enhancing the scanning time and quality, iv) considering the maximum integration levels with control, motion, and scientific computing systems, v) minimising user experience gaps, vi) automating raw data and metadata collection, and vii) adhering to software engineering standards developing systems inasmuch as possible.

DATA POLICY AND DAQ STANDARDIZATION

The implementation of the data policy not only aids standardising DAQ systems and the data sets generated across beamlines, but also enhances the integration between DAQ and other systems in which they serve as a source of raw data and metadata, as well as the computing infrastructure used to store data and provide access and analysis services on it.

In nutshell, the data policy constitutes and describes the ownership, storage, access and management of the experimental data generated from SESAME beamlines [8]. It is applied on the experimental data for both users awarded beamtime proposals and in-house experiments. There is a three year embargo period after an experiment is completed, during which only the principal investigator of the beamtime proposal and the experimental team have access to the data. Passing the three years, the experimental data becomes openly accessible to SESAME registered users. There is a possibility to exceptionally extend this period by submitting a request to SESAME officials. The experimental data is kept at SESAME archiving for at most ten years on best efforts basis; the exact number of years will be defined for each beamline in reference to the data type and volume considering the financial and technical limitations. The policy implies sorting and archiving the experimental data in well-defined format where such data should be acquired from SESAME software and associated with unique key identifier. In contrast, this data policy does not apply to industrial research and collaborative research with SESAME scientists where other policies will be specifically developed.

EXPERIMENTAL DATA COLLECTION

The majority of SESAME's beamlines use locally developed and maintained DAQ systems. They are the main

* mostafa.zoubi@sesame.org.jo

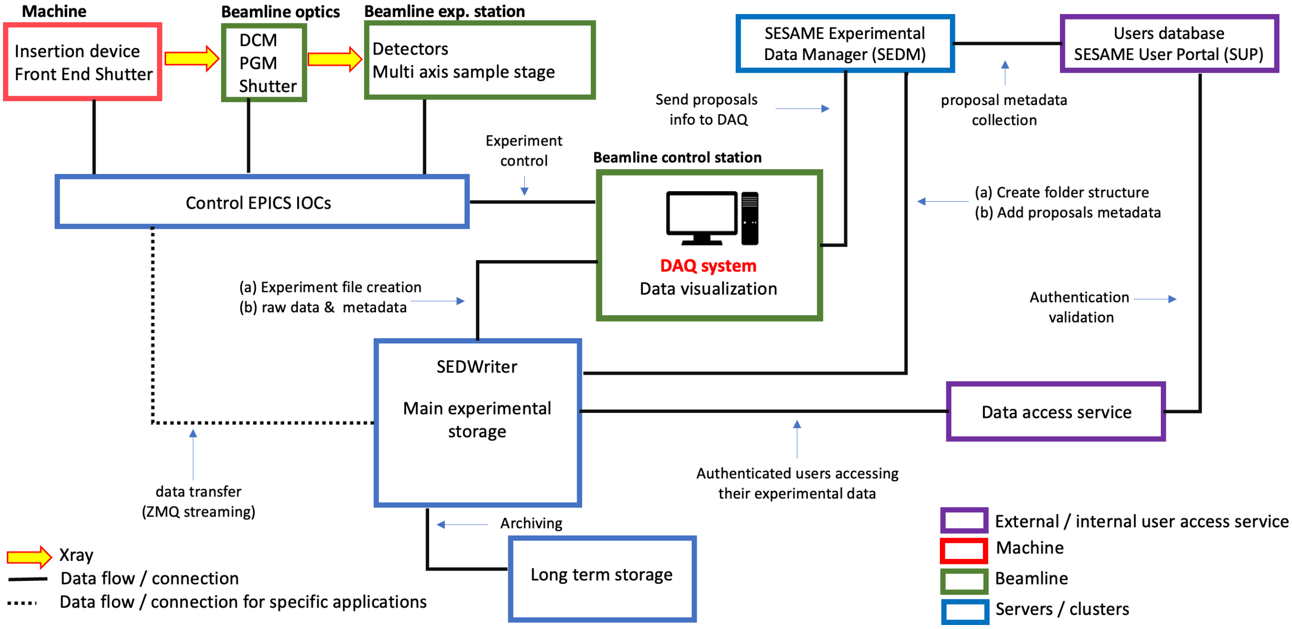

Figure 1: A General block diagram describes beamlines DAQ system at SESAME and its interaction with other systems. Normally, Double Crystal Monochromator (DCM) and Plane Grating Monochromator (PGM) are not combined together in the same beamline also not all beamlines have Insertion Device (ID) as a source, however, this indicative block diagram shows different beamline components that are controlled or provide data to DAQ through the Experimental Physics and Industrial Control System (EPICS) Input Output Controllers (IOCs) layer. On the other hand, it is very common to have more than one detector in the beamline where DAQ is supposed to drive them either individually or simultaneously depending on the beamline and applications.

beamline users' and scientists' interface to set up and control scans for acquiring and collecting experimental data. Each beamline at SESAME has its own and specific DAQ system, which is pure object-oriented and follows the modern software engineering development standards in terms of modularity, testing and validation, user experience, documentation, and code version control. Figure 1 shows a general block diagram that describes the DAQ pipeline. The DAQ system interacts with many other systems to create experimental data and makes it available for access, processing, and archiving services.

Beamline Scanning Techniques

All of the beamlines that use our DAQ system now make use of step scanning technique, which is easy to use, reliable, and operable even with basic hardware and control support. The main drawback with this technique is the long time it takes to complete a scan, as most of the time is spent waiting for mechanical parts to move and, once the movement is done, for mechanical vibration to settle [9, 10]. Moreover, detector readout time and the DAQ system's connection time are additional elements that impose delays. In order to get around this issue, several synchrotron facilities have introduced continuous scanning mode namely the on-the-fly mode. This technique has been successfully demonstrated and put into operation with a large number of beamlines across the facilities.

At SESAME, recent work carried out on the testing bench of the BEATS beamline, resulted in the development of a software-based continuous scan (no external triggering system). This was done in an attempt to synchronise the detector readout frames with the rotating stage speed. Visible light tomography [11] has been used to evaluate the effectiveness of the technique as the beamline is being constructed, and the results have been very encouraging.

DAQ Standers and Features for Beamlines

SESAME's DAQ system for any beamline is constructed on top of the EPICS control system. This necessitates the presence of EPICS device support for the various hardware components that are placed at the beamlines, the control team also employs this as their standard when installing new devices and components at SESAME. The DAQ system also employs standard EPICS records and drivers for certain and often used subsystems. More specifically, EPICS motor records are used for motion system, while areaDetector driver is used for any 2D imaging detector [12].

Graphical User Interface (GUI) is used for the DAQ system which helps the users to easily and quickly understand the experimental settings and the scanning parameters. Motif Editor and Display Manager (MEDM), EPICS Qt, and PyQt are now being used as GUI clients throughout the beamlines.

Automated energy calibration module has been recently added to DAQ of the XAFS/XRF beamline. The module adjusts the monochromator's Bragg angle (theta) based on selected and well known metallic references of some elements . The module enables users to perform a calibration scan, from which necessary data is automatically extracted and made available to different processing techniques via a GUI, before proceeding to the final step, which calculates the angular offset of and set it tothe theta motor as a correction between Bragg angle and the energy.

In order to ensure that all DAQ systems are in line with the data policy, the SESAME users' database should be integrated with all of them. The database contains metadata about users and their submitted beamtime proposals. In this context, recently, we were able to effectively link the XAFS/XRF beamline DAQ system with the database by means of a module that is developed and named SESAME Experimental Data Manager (SEDM). On each experimental day, this module pulls out all metadata pertaining to accepted and scheduled proposals, creates folder structures includes the experimental data path for each proposal in the main storage using a predefined naming convention scheme, and then sets the appropriate access and read/write permissions for those folder structures. SEDM can intelligently identify and communicate proposals' metadata to their assigned beamlines, allowing DAQ to inject this data into the experimental files. Before each scan, the DAQ system verifies all scan parameters to prevent human mistakes, with proposal number validation being the most critical, as it is used to uniquely identify the generated data sets. For beamtime allocated to accepted proposals, users are required to enter their proposal number to be validated with the SEDM metadata.

Figure 2 shows the validation process in an activity diagram. The system only accepts scheduled proposals for current experimental day, otherwise users must contact the beamline scientist, who has full access to the beamline calendar, in order to reschedule an accepted proposal on this particular experimental day.

Since unanticipated issues can crop up during scans, our DAQ system supports for what we call unattended scan, which implies that a scan may run, detect issues, and take corrective action without human intervention. Currently, one action has been implemented that the system does not acquire noise data when the readouts of beamline-specific specified parameters exceed predefined limits (i.e. beam availability and energy, shutters, detector readout, .. etc). In such scenario, the scan will be temporarily paused and resumes automatically when favourable conditions are satisfied.

The system is able to automatically switch between samples, each of which can have different scan parameters applied. The number of samples is dependent only on the capability of the sample holder used in the beamline and has no software limits. The user is responsible for assigning the name and position to identify each sample. The naming convention that DAQ adopts for generating experimental files necessitates that the name shall be descriptive of the sample. In most cases, the number of experimental files is typically proportional to the number of samples.

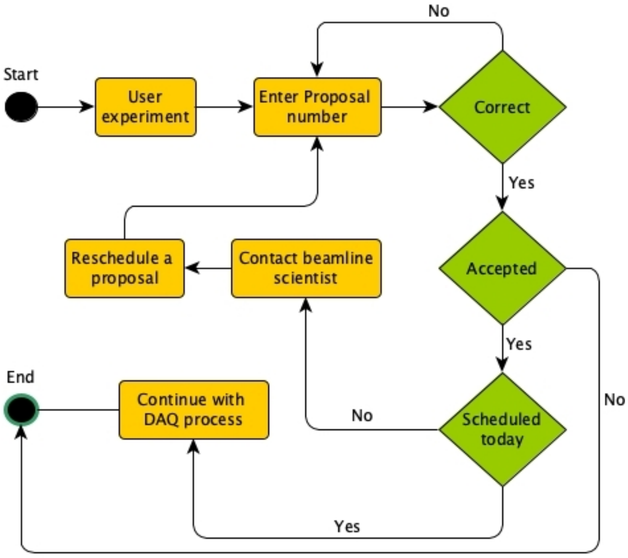

Figure 2: Activity diagram illustrates the proposal validation process on the experimental day before start collecting data sets.

A live data visualisation module has been added to the DAQ system, allowing users to see live raw data or live data representations in various graphs. The module also shows scan progress statistics, and includes basic analysis functions.

The DAQ system coordinates and tracks experimental data flow from its origins (IOCs) to its destination in the main storage. The dataflow route can either go through the DAQ itself or direct streaming from the IOCs to the storage, with the choice depending on the dataset volume size being sent, the Operating System (OS) hosting the IOC driver, and the application needs for a particular beamline. Experimental data goes through DAQ at most of the beamlines. Recently, we implemented a direct streaming solution on BEATS testing bench, which streams the data from the detector driver to the main storage at 1.1 Gigabyte per second, avoiding the limitations imposed by protocols transferring bigdata between different files systems. The publisher-subscriber sockets connection type of the asynchronous messaging library ZeroMQ (ZMQ) was used achieving this solution [5].

On the main storage, the customized module we built, SESAME Experimental Data Writer (SEDWriter), is responsible for writing the experimental data set into experimental files including raw data and metadata. In order for SEDWriter to function, it needs to know the origin of this data, type of experiment whether in-house or user experiment, the file format it has to create and some other beamline-specific settings. The needed information to the writer is provided by DAQ system of each beamline. At the time being, our writer generates experimental data files in three formats, namely X-ray absorption spectroscopy Data Interchange (XDI), Hierarchical Data Format version 5 (HDF5) - Scientific Data Exchange (DXfile) and data file (DAT). XDI is the stander experimental file for XAFS/XRF and HESEB beamlines. It is an open-source ASCII-based file that can be opened in any text editor, self-documented, flexible with adding metadata, and is being developed and used in similar laboratories [13]. The DXfile format will be utilised to record experimental data for BEATS beamline and it has been successfully tested on its testing bench. The DXfile file is open-source HDF5 based, self-describing with a tree structure, and supported in the file plugin of areaDetector driver. The file includes a full description about the proposal, experiment, beamline, sample, sample stage and detector [14].

On the other hand, to be in compliance with our data policy standardising SESAME experimental data files, we are evaluating DXfile format to be adopted as a universal experimental file format for all beamlines at SESAME. Diffraction data and other metadata associated with the MS beamline are recorded in DAT files.

Every time an experiment is run using our DAQ system, all the scan parameters, choices, and settings are recorded in plain-text configuration file. This configuration file is sent to SEDWriter so that any available metadata may be extracted, and the file itself can be used again to import its contents into DAQ GUI for editing to run another scan.

Within our system, we have implemented a simple online logging module that provides both a timestamp and colour labelling for each entry. This functionality assists us in error troubleshooting, makes it simpler to debug, and allows us to track which part of DAQ is now being executed. The logs are also captured in a log file for each experiment, and this log file is stored next to the experimental data files so that users and scientists have access to them in the event that they are required.

With each DAQ system that we provide for a beamline, we used to write a straightforward documentation that is aimed at the scientist who would be operating the system. We recently made the switch to a professional open-source documentation service, namely “Read the Docs” that caters to beamline scientists, users, and developers. This was necessitated by the fact that we are a users’ facility and also utilise GitHub for version control, which is already integrated with “Read the Docs”.

DATA ACCESS AND ARCHIVING

Each beamline’s DAQ system has distinct needs for storing experimental data. From storage administrator standpoint, the dataset volume, the number of files, and the desired writing speed performance define out our storage solutions. Currently, the primary experimental storage system is comprised of two high-end disc arrays: i) General Parallel File System (GPFS) and ii) Storage Area Network (SAN) storages. The Server Message Block (SMB)/Common Internet File System (CIFS) and Network File System (NFS) protocols make both arrays accessible for services, users and beamline scientists. The access and permissions are managed by SESAME’s active directory server. On the other hand, for cluster-to-cluster access or clients with high throughput demands, we use the native GPFS client, which allows transferring data at the maximum performance of the storage.

Access to experimental data is a vital aspect of completing the experiment life cycle, where dataset exporting activities are performed at two levels to give all possible data manipulation options for beamline scientists and users. On the basic level, the generated experimental data is downloadable by an authorized access service using SMB and NFS clients. This level is designed for on-premises use and the transmission of relatively modest datasets. For the advanced level, we are currently implementing ICAT [15] service, which is an open-source large-scale data management system, developed and maintained by scientific communities, and it has all necessary features needed to securely enable the web access to our experimental data. ICAT implementation will be focused on adapting SESAME’s data policy, it is already a data catalogue service and the core elements of FAIR data (findability, accessibility, interoperability, and re-usability) are already there. ICAT service will be integrated with SESAME storage systems and the universal experimental file format that will be used. Through ICAT, the users will also be able to post analysed results from their experiments and link those results to their publications using the service.

Undoubtedly, the fast growth of experimental data as well as the adoption of the SESAME data policy will make it necessary to establish a long-term data life cycle service. To achieve this, a data archiving solution should be serving in place. In accordance with our data policy, this will be the primary location where experimental data will be stored for a period of 10 years. At the moment, we are evaluating different technologies to implement this service. Cloud, disc, and tape based storages are compared from a variety of perspectives, with emphasis on reliability, data availability, and initial and running costs. Early findings suggest that tape-based storage might be the optimal solution for us.

CONCLUSION

This paper describes our DAQ systems, standards, and procedures for the majority of SESAME beamlines. Not one beamline’s DAQ in particular is addressed in any great detail. Table 1 provides a summary of available and planned DAQ system features. All systems are operational and fulfil the needs for the time being. However, for the upcoming period, it will be essential to implement continuous on-the-fly scanning mode in order to significantly reduce the scan times. Given that BEATS will be the first massive data-generating beamline at SESAME, it is advantageous to have a data policy in place prior to the expected increase in data volumes. As a result, the DAQ systems on all beamlines, including BEATS, are currently confirmed to a set of standards to assure policy compliance.

ACKNOWLEDGEMENTS

We thank the EU-H2020 BEATS project and European Synchrotron Radiation Facility (ESRF) for sharing assisting and helping us developing our data policy.

Table 1: DAQ Features Implementation Summery/Plan: Completed (C) – Under Development (D) – Planned/Under evaluation (P) – To be improved (T) – Not Applicable (NA)

Feature	XAFS/XRF	MS	HESEB	BEATS
EPICS motor records	C	C	C	P
areaDetector	NA	D	NA	C
GUI	C	P	C	C
Auto energy calibration	C	T	T	NA
Users' DB integration	C	D	C	P
Step scan	C	C	C	C
Cont. scan	P	P	P	C
Unattended scan mode	C	C	C	P
Auto sample changing	T	P	T	P
Data visualisation	T	T	C	C
ZMQ streaming	NA	D	NA	C
Common file format	C	P	C	C
Exp. config file	C	C	C	C
Logging	C	C	C	C
Code version control	C	C	C	C
Documentation	D	P	C	P
Feature	**Institution level, for all beamlines**			
Universal file format	P (dxFile as initial finding)			
Long-term archiving	P (tape-based storage as initial finding)			

REFERENCES

[1] SESAME Synchrotron-light, https://www.sesame.org.jo

[2] M. Harfouche, M. Abdellatief, Y. Momani, A. Abbadi, M. Al Najdawi, M. Alzubi, B. Aljamal, S. Matalgah, Lu. Khan, A. Lausi, and G. Paolucci, "Emergence of the first XAFS/XRF beamline in the Middle East: providing studies of elements and their atomic/electronic structure in pluridisciplinary research fields", *J. Synchrotron Radiat.*, vol. 29, pp. 1107-1113, 2022. doi:10.1107/S1600577522005215

[3] G. Kamel, S. Lefrancois, T. Moreno, M. Al-Najdawi, Y. Momani, A. Abbadi, G. Paolucci, and P. Dumas, "The first infrared beamline at the Middle East SESAME synchrotron facility", *J. Synchrotron Radiat.*, vol. 28, pp. 1927-1934, 2021. doi:10.1107/S1600577521008778

[4] W. Drube, MF. Genisel, and A. Lausi, "SESAME Gets Soft X-Ray Beamline HESEB", *Synchrotron Radiat. News*, vol. 35, iss. 1, pp. 22-22, 2022. doi:10.1080/08940886.2022.2043710

[5] G. Iori, S. Matalgah, C. Chrysostomou, A. Al-Dalleh, and M. Alzu'bi, "Data Acquisition and Analysis at the X-ray Computed Tomography Beamline of SESAME", *IEEE Jordan Inter. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT'21)*, pp. 134-139, 2021. doi:10.1109/JEEIT53412.2021.9634151

[6] M. Abdellatief, M. A. Najdawi, Y. Momani, B. Aljamal, A. Abbadi, M. Harfouche, and G. Paolucci, "Operational status of the X-ray powder diffraction beamline at the SESAME synchrotron", *J. Synchrotron Radiat.*, vol. 29, pp. 532-539, 2022. doi:10.1107/S1600577521012820

[7] R. Dimper, A. Götz, A. de Maria, V.A. Solé, M. Chaillet, and B. Lebayle, "ESRF Data Policy, Storage, and Services", *Synchrotron Radiat. News*, vol. 32, iss. 3, pp. 7-12, 2019. doi:10.1080/08940886.2019.1608119

[8] SESAME Experimental Data Management Policy, https://www.sesame.org.jo/for-users/user-guide/sesame-experimental-data-management-policy

[9] Shang-Wei Lin, Chia-Feng Chang, Robert Lee, Chi-Yi Huang, Chien-I Ma, Liang-Jen Fan and Hok-Sum Fung, "On-the-fly Scan: Improving the Performance of Absorption Spectrum Measurement", in *Proc. 11th Int.l Conf. Synchrotron Radiat. Instrum., J. Phys.: Conf. Ser.*, vol. 425, p. 122002, 2013. doi:10.1088/1742-6596/425/12/122002

[10] S. Zhang, Y.M. Abiven, J. Bisou, G. Renaud, G. Thibaux, F. Ta, S. Minolli, F. Langlois, M. Abbott, T. Cobb, C.J. Turner "Pandabox: A Multipurpose Platform for Multi-Technique Scanning and Feedback Applications", in *Proc. 16th Int. Conf. on Accelerator and Large Exp. Control Sys. (ICALEPCS'17)*, Barcelona, Spain, 2017. doi:10.18429/JACoW-ICALEPCS2017-TUAPL05

[11] LINXS, https://www.linxs.se/news/2021/10/11/from-kitchen-tomography-to-a-cutting-edge-neutron-imaging-beamlinenbsp

[12] EPICS, https://epics.anl.gov

[13] XAS Data Interchange, https://github.com/XraySpectroscopy/XAS-Data-Interchange

[14] DXfile, https://dxfile.readthedocs.io/en/latest/

[15] D. Flannery, B. Matthews, T. Griffin, J. Bicarregui, M. Gleaves, L. Lerusse, R. Downing, A. Ashton, S. Sufi, G. Drinkwater, K. Kleese, and D. Limited, "ICAT: Integrating data infrastructure for facilities based science", 2009 Fifth IEEE International Conference on e-Science, Oxford, UK. doi:10.1109/e-Science.2009.36