DOI: 10.17199/NOBUGS2016.57

Scientific data lifecycle at Elettra-Sincrotrone Trieste.

Milan Prica, Fulvio Bille`, Roberto Borghes, Valentina Chenda, Alessio Curri, Daniele Favretto, Georgios Kourousias, Roberto Pugliese, Martin Scarcia, Michele Turcinovich

Elettra-Sincrotrone Trieste S.C.p.A. Strada Statale 14 - km 163,5 in AREA Science Park 34149 Basovizza, Trieste, Italy

milan.prica@elettra.eu, roberto.borghes@elettra.eu, george.kourousias@elettra.eu

Abstract. Elettra-Sincrotrone Trieste comprises Elettra synchrotron and FERMI free-electron laser. Combined, the two facilities serve 34 beamlines. The Scientific Computing Team supports the full data lifecycle. Proposal submission and evaluation are handled in the Virtual Unified Office. Data acquisition and experimental control are built on top of the TANGO control system on all but the oldest synchrotron beamlines. Data reduction, on- and off-line analysis workflows and visualisation are supported by a common framework. Data is stored and catalogued with access through the web portal. Elettra implements the PaNdata-like data policy since 2014.

1. Introduction

Synchrotrons and especially free electron lasers facilities produce enormous quantities of scientific data during their normal operations. With advances in modern instrumentation and in particular with the newest, fastest, high resolution detectors, the problem of data deluge often arises. Besides, facilities handle a large quantity of sensitive user information and invaluable intellectual property contained in the research proposals. Managing this high volumes of scientific data and their ownership crosses different application systems, databases and storage media, often beyond the facility boundaries. The data lifecycle at Elettra facilities [1] is a complex process that begins with a research proposal and follows through a series of steps including laboratory access, data acquisition, meta-data correlation, pre-processing, storage, setting access rights, retrieval, analysis and publications.

2. A brief overview of the facilities

Elettra-Sincrotrone Trieste runs two advanced light sources: a third-generation synchrotron Elettra and a single pass Free Electron Laser (FEL), FERMI. Since the major upgrade of the facility in 2010, Elettra SRF operates in a top-up mode at two energies: 2.0 GeV for enhanced extended ultraviolet performance and spectroscopic applications (75% of user time) and 2.4 GeV for enhanced x-ray emission and diffraction applications (25% of user time). A vast range of research in physics, chemistry, biology, life sciences, environmental science, medicine, forensic science, and cultural heritage is carried out at 28 beamlines. A substantial upgrade of the Elettra machine is planned [2] in the near future.

The FERMI FEL design is based on an external seeding scheme that improves the output pulse coherence, central wavelength control and spectral bandwidth. The FEL output is tunable in power, wavelength, temporal duration, and polarization. With a peak brightness of about 6 orders of magnitude higher than third generation sources, and pulse lengths of the order of a picosecond or less, FERMI supports scientific investigations of ultra-fast and ultra-high resolution processes in material science and physical biosciences. FERMI is running at 50 Hz pulse rate since 2015. Currently six beamlines are operational.

3. User registration and proposal submission

Virtual Unified Office (VUO) is the main users' interface to the facilities. Constantly evolving since 1997, the web portal manages all the steps that together form an experiment, beginning with the user registration and a proposal submission and following through the elaborated results and publications. Besides Elettra and FERMI, the VUO portal manages also the facilities of the CERIC-ERIC [2] consortium.

The process starts with the creation of the user account. A user's login credentials will be valid not only on the web portal but also on the acquisition workstation, on the data storage devices, on the data analysis computational cluster, on the remote operations tools and for the Wi-Fi access. The authentication and authorization system dedicated to the scientific users is separated from the main company one. Umbrella authentication system is supported on the web portal. The proposal evaluation process is entirely managed in the portal, from proposal submission to the beamtime scheduling for the accepted ones.

Elettra was among the first light source facilities to adopt an official data policy in 2014. Users must agree to the data policy as part of the proposal submission procedure. The Elettra implementation is a soft version of the PaNdata proposal. A three-year embargo period is easily (automatically) extendable and while no data have been deleted to date, no hard guarantee is offered for long term data preservation. DOIs on the datasets are not issued at the present moment but that is likely to change in the future.

4. Data acquisition

The beamline end-stations of Elettra and FERMI are complex instruments, each consisting of a large number of interconnected components. Each experiment requires a long sequence of operations on most of these components that include among others valves and shutters, vacuum pumps, positioning motors, and a large variety of sensors, cameras and detectors. The data acquisition software should be capable of commanding all of these components while at the same time enabling the user to monitor the process through a clear graphical interface. A flexible, extensible and easily adaptable end-station control software based on the Tango distributed control system has been developed [3]. Its key components regard the automatization of the experimental sequence (Executer) and a fast acquisition system (FDAQ).

The Executer application is a highly flexible PyTango device that executes generic Python functions from scripts located in external files. These scripts can call external programs and functions written in other languages. Input variables and result viewers are handled as dynamic Tango attributes configurable via an XML configuration file. Each experimental sequence may be coded as an Executer's function.

The system for the data acquisition at FERMI had to be designed with the maximum regard for the high speed and the pulsed nature of a FEL source. The FDAQ application is PyTango device consisting of multiple threads (one for each data source) that collect and store shot-to-shot data and the related meta-data from a large number of Tango sources. Data and meta-data sources are configurable through an XML file. Elettra beamlines typically present much simpler, custom developed data acquisition systems. Except for the oldest beamlines, the end-station controls are Tango based. The experimental data and metadata are organized and saved in the HDF5 binary scientific data format. The HDF5 structure is custom at each beamline and no beamline has adopted NeXuS at the present moment. At FERMI, a typical experimental dataset is composed of multiple HDF5 archives containing data relative to contiguous FEL shots. FDAQ stores data in a fast local storage called scratch.

The main graphical user interface (GUI) to the data acquisition software presented to scientists has been developed in QTango, a C++ based framework. Just like the other key components of the system, it is highly customizable through an XML configuration file that defines both the content and the appearance of the front-end panel. The main section of the panel is a tabbed panel relative to the experiment. Each tab contains input and output variables and the command buttons of the corresponding Executer script. Another section is relative to the FDAQ application and contains a dynamic and checkable list of the available data sources. Finally, there is a common zone for data viewers and the additional command buttons. To access the panel, users must login into the acquisition system with their VUO credentials and choose a proposal to which the data collection is related.

4.1. Remote access

A number of tools are available for safe remote access to selected beamline controls and data analysis tools. Remote access allows for collaboration between researches present at the site and those following the experiments from their home institutes and guarantees prompt interventions from beamline scientist in case of necessity. However, the remote access presents high security risks and must be handled with utmost care. VUO integrates a Java-based tunneling tool with the authentication and authorization mechanism in support of RDP, NX and VNC.

4.2. End-user tools

The Scientific Computing team has built a collection of software tools [4] with a high level of customization reusable in a multitude of situations in support of user-oriented applications at the end-stations. These tools include an electronic logbook built on top of a HTML-based, WYSIWYG text editor that supports screenshot insertions and integrates a data server for upload of images and log messages from remote machines. Files are saved automatically and exportable to PDF format. Another important tool in the collection is a standard visualizer that supports a number of common scientific image formats like CBF, TIFF, MAR345, etc. It provides functionalities like region-of-interest zooming, line profiling, automatic or manual contrast adjustments and background subtraction. The visualizer tool is used on many beamlines of Elettra and FERMI for online data analysis of images from Dectris Pilatus detectors, mar345 imaging plates, Princeton Instruments CCDs, Basler and Andor cameras. A common necessity to investigate a new experimental techniques inspired the development of a tool for fast prototyping of data collection sequences and implementation of data analysis pipelines. Built as an extension module to Python Spyder IDE it integrates an interactive help system and numerous custom libraries developed and tested by the specialized staff. Plug-in based extensibility is common to all of the above tools.

5. Storage and data access

At some point, the collected data has to be moved from the fast scratch storage to the accessible online storage. This step may be performed manually from the VUO, but it is most often done automatically as part of the data acquisition pipeline. Some data preprocessing like the lossless compression on HDF5 files may be executed at the same time. Data is moved with *rsync* and the transfers are coordinated by the dedicated Tango devices. Tango devices take care of the creation of the necessary directories and setting of the POSIX access rights. Once moved to the online storage, the raw data files cannot be further modified. Selected users may access data in the online storage following a simple proposal / dataset / datafile scheme. The principal investigator and the responsible beamline scientists may enable anyone with a valid VUO account to access the data of the proposal. Data are accessible through WebDAV and through web browsers. Multiple files or entire datasets (folders) may be downloaded as a single zip file. The directory tree created provides a placeholder for elaboration results to which files may be uploaded. The E-logbook file generated during the beamtime is stored in the PDF format with the dataset. If and when necessary, data may be moved to long term (offline) storage at the CINECA supercomputing center. CINECA system is based on iRODS. Since the beginning of the PaNdata-ODI project, Elettra has been following the ICAT metadata catalog development. A move to ICAT will likely be necessary to implement open data access but the final decision is yet to be made. A purchase of some additional storage space for the online storage is imminent.

6. Conclusion

Implementing a complete data lifecycle in a large experimental physics facility, as is the case with Elettra, is a very complex and expensive task. A large number of heterogeneous systems have to work in concert while new pieces are constantly added to the puzzle. There are numerous attempts from international collaborations to provide solutions that address common challenges related to the Big Data and the open data access. The planned Elettra 2.0 upgrade will provide us an important window of opportunity to implement some of those advances.

References

[1] Billè F, Borghes R, Brun F, Chenda V, Curri A, Duic V, Favretto D, Kourousias G, Lonza M, Prica M, Pugliese R, Scarcia M, Turcinovich M Data Lifecycle In Large Experimental Physics Facilities: The Approach Of The Synchrotron ELETTRA And The Free Electron Laser FERMI *Proceedings of ICALEPCS2015,* Melbourne, Australia

[2] Karantzoulis E, Elettra 2.0 – The next machine, *Proceedings of IPAC2015,* Richmond, VA, USA

[3] Borghes R, Chenda V, Curri A, Kourousias G, Lonza M, Passos G, Prica M, Pugliese R, A common software framework for FEL data acquisition and experiment management at FERMI@Elettra, *Proceedings of ICALEPCS2013*, San Francisco, CA, USA

[4] Borghes R, Chenda V, Kourousias G, Lonza M, Prica M, Scarcia M A Flexible System for End-User Data Visualisation, Analysis Prototyping and Experiment Logbook, *Proceedings of ICALEPCS2015,* Melbourne, Australia