THE AUSTRALIAN STORE.SYNCHROTRON DATA MANAGEMENT SERVICE FOR MACROMOLECULAR CRYSTALLOGRAPHY

Grischa R Meyer, Steve Androulakis, Philip Bertling, Ashley Buckle, Wojtek James Goscinski, David Groeneveld, Chris Hines, Anitha Kannan, Sheena McGowan, Stevan Quenette, Jason Rigby, Patrick Splawa-Neyman, James M Wettenhall, Monash University, Clayton, Australia
David Aragão, Tom Caradoc-Davies, Nathan Mudie, Synchrotron Light Source Australia, Clayton
Charles Bond, University of Western Australia, Crawley, Australia

Abstract

Store.Synchrotron is a service for management and publication of diffraction data from the macromolecular crystallography (MX) beamlines of the Australian Synchrotron. Since the start of the development, in 2013, the service has handled over 51.8 TB of raw data (~4.1 million files). Raw data and autoprocessing results are made available securely via the web and SFTP so experimenters can sync it to their labs for further analysis. With the goal of becoming a large public repository of raw diffraction data, a guided publishing workflow which optionally captures discipline specific information was built. The MX-specific workflow links PDB coordinates from the PDB to raw data. An optionally embargoed DOI is created for convenient citation. This repository will be a valuable tool for crystallography software developers. To support complex projects, integration of other instruments such as microscopes is underway. We developed an application that captures any data from instrument computers, enabling centralised data management without the need for custom ingestion workflows. The next step is to integrate the hosted data with interactive processing and analysis tools on virtual desktops.

INTRODUCTION

The deposition and the wide availability of raw diffraction data have far-reaching benefits for the structural biology community [1–4]. We recently created TARDIS, a suite of tools for the deposition of X-ray diffraction images in an open access repository to facilitate their deposition using federated institutional repositories [3]. Subsequent engagement with IUCR and closer relationships with the Australian Synchrotron prompted us to develop a new framework for raw X-ray data archival and dissemination — the Store.Synchrotron service, which is described in this manuscript.

An earlier but more detailed description of this service is available at [5].

THE OPERATION OF STORE.SYNCHROTRON FOR THE AUSTRALIAN SYNCHROTRON

The Store.Synchrotron service (https://store.synchrotron.org.au) has been deployed to receive diffraction data from the macromolecular crystallography (MX) beamlines at the Australian Synchrotron (AS) automatically.

When a user starts collecting data at the beamline, all raw diffraction data and automated data processing results are transferred in real-time and on the next day respectively to Store.Synchrotron without user intervention.

This system is accessible through any web browser. Data is available immediately at time of collection to all authorized users via the internet. If desired, the data can be opened to the public.

Figure 1 shows a schematic overview of the service.

Figure 1: Schematic overview of Store.Synchrotron

The Need for an Automated Data Management System

At the Australian Synchrotron (AS) up until recently the most practicable solution for data management has been to take portable hard drives to the beamlines. The use of portable drives is a problem, however, because these drives may remain the lab’s sole resource for raw data. After researchers have left, it may be difficult to correlate raw frames with datasets and structures, and a single disk is very prone to failure and data loss.

With Store.Synchrotron users now have for the first time a long-term, secure archive of their raw data, including metadata, which can be used to relate raw frames to projects and visits to the Australian Synchrotron.

In addition to the practical need for a robust data archive, new requirements by funding agencies for researchers to have a data management plan are conveniently solved by a

ISSN 978-3-95450-148-9

Data Management, Analytics & Visualisation
centralised solution that takes care of the data from collection to archival and/or publishing.

**The End-to-End User Experience**

The default way of accessing data on the Store.Synchrotron service is via the web through a web browser. As a user collects a diffraction image it is available on the Store.Synchrotron website within minutes (Fig. 2) including instrument specific metadata.

Data is organised into three tiers. An Experiment is related to an allocation on the beamline (Fig. 2A,B). The data is then separated into Datasets (Fig. 2C). Finally, the individual diffraction images are Datafiles in their respective datasets (Fig. 2D). All or selective elements of each tier can be downloaded via the web or via SFTP.

In addition to diffraction datasets, Store.Synchrotron makes accessible the results of preliminary autoprocessing. Collection of the images triggers automatic indexing and full integration and scaling for complete datasets. All autoprocessing output is downloadable as well.

**Design and Implementation**

The Store.Synchrotron service is an implementation of MyTardis (http://mytardis.org); a web-based open-source data management platform. MyTardis began as the TARDIS [3] (http://tardis.edu.au) diffraction image repository but has since expanded to suit the storage and organisation of a wide range of scientific data. The interface allows browsing, data access, data sharing and publication.

The beamline registers its the raw images with Store.Synchrotron in real-time, triggered by the instrument control system. Different processes extract information and preview images, compute SHA-1 checksums, and copy the diffraction data to the final Store.Synchrotron-managed storage location. By the end of an allocation slot both raw data and autoprocessing results are archived and made available on Store.Synchrotron. For resilience all these processes are run using a message queue. All networked communication is encrypted.

The integrity of the files stored within the Store.Synchrotron service is of utmost importance. As the data are the results of often non-repeatable experiments, data is verified via checksums at several steps in the pipeline, starting with ingestion. If a checksum does not match, it can be re-sent if possible. All archived files are re-verified regularly to ensure the ongoing integrity of the data.

In a 12 month period since Store.Synchrotron began operation, only 8 out of more than 1,700,000 files had mismatching checksums. This means the vast majority of data (99.9995%) was transferred successfully in the first attempt. It also demonstrates the value of integrity checking as these 8 files could be re-transferred before the original was lost.

The Store.Synchrotron service receives metadata via its RESTful API. By following the REST standard we made it trivial for developers to interact with Store.Synchrotron programmatically, e.g. automatically or manually via different front ends.

Another important technology we use that allows space and time efficient archiving and retrieval of large datasets is SquashFS (http://squashfs.sourceforge.net/). SquashFS is an archive file standard that supports compression and direct random file access. Store.Synchrotron can transparently provide access to individual files inside SquashFS archives.

Store.Synchrotron was developed based on the web framework Django. It is running on a Ubuntu 14.04. It is hosted on the NeCTAR Project’s cloud computing service (http://www.nectar.org.au/research-cloud). To assist the maintenance and coordination of this service, we have used an orchestration and configuration management system (SaltStack http://www.saltstack.com/).

**Benefits from User Perspective**

Protein crystallography has benefited for many years by an open-access community based approach to the constant development and improvement of data processing, structure solution and refinement tools. As these tools evolve, it is often beneficial to revisit projects where data collected has proven intractable at the time, as has been done for example in [6].

A broadly systematic approach can be taken to processing old data for new results. Data processing programs are continuously evolving. As the Store.Synchrotron public repository grows we can envisage a system where as new methods to analyse raw data become available they are used against old data automatically.

**OPEN DATA FOR THE CRYSTALLOGRAPHIC COMMUNITY**

The Australian Synchrotron is the only facility of its type in Australia, and is in constant use. As such, the organisation has actively encouraged opening up of data produced there, to encourage efficiencies and reuse. This requires services that enable effective long term storage, curation and citation.

 Provision of services related to making data more easily available are timely, as Australian government research funders have been strengthening their language in this area [7, 8], bringing Australia more in-line with the global research environment and that of international funding agencies such as the US National Science Foundation (NSF) and the UK Medical Research Council (MRC).

 Store.Synchrotron also represents the logical extension of a long-standing effort in the macromolecular crystallography community to ensure that satisfactory evidence is provided to support the interpretation of structural experiments. This effort has included unrivalled requirements for validation of data interpretation processes [9–12].

 Other projects exist to deposit and host openly-accessible diffraction data on the web, such as DIMER (https://dimer.uq.edu.au/), the registration-required JCSG Repository of Crystallography Datasets (http://www.

jcsg.org/), and TARDIS (http://tardis.edu.au/). Store.Synchrotron differs from these services by automatically archiving all diffraction data produced from beamlines at time of data collection. This act eliminates the need for manual diffraction data deposition by the user as a step toward open crystallography data.

A European project, PaN-Data (http://pan-data.eu/) has also created infrastructure for the automatic archival and dissemination of raw diffraction data from facilities such as the Diamond Light Source and the ESRF. This is achieved through a combination of an information management and real-time data monitoring system ISPyB [13], metadata store ICAT (http://icatproject.org/) and its data-browsing and access web front-end TopCat (https://code.google.com/p/topcat/). To date, access to data has been private to users who collect it. Store.Synchrotron extends this functionality by providing both an automatic archival mechanism from the time of data collection and a publishing interface for open access to collected data.

To facilitate the publication process the Store.Synchrotron has developed a publication form to assist researchers in identifying and describing the raw data associated with their experiment and adding metadata such as related PDB entries. After a final check by Synchrotron staff the Experiment link is made publicly available and discoverable via the Store.Synchrotron site, with a Digital Object Identifier (DOI) minted for citing the data.

Considerations for Use of the Store.Synchrotron Model in Other Facilities

If this system is to serve as a model for evaluation by other synchrotrons and instrument facilities around the world, it is important to consider potential points of difficulty that may arise with the application of our approach. Certainly, the most difficult infrastructure requirement for Store.Synchrotron is the expense of storage and compute servers. When operating a data intensive service, a significant challenge arises: Who will pay to keep this data alive?

Monash University is a partner operator of VicNode (http://www.vicnode.org.au), the local node of the Research Data Service (RDS, http://www.rds.edu.au). VicNode provides a multi-location, tape-archived, large volume cloud data storage service that is utilised for all the data stored at the Store.Synchrotron service. Storage initially
provided to Store.Synchrotron by the VicNode is over 200 terabytes in size and is a combination of disk and tape, with an off-site backup.

The minting and ongoing maintenance of Digital Object Identifiers (DOIs) represent an ongoing cost. Store.Synchrotron uses the federally-run Australian National Data Service (ANDS) who provide a service that allows the free minting and maintenance of DOIs resolving to Australian research data.

As discussed above, the MyTardis data management and instrument integration application is free, open-source, and able to be deployed on a compute cloud or web server in an automated fashion. We envision that our service could offer an attractive model elsewhere for raw data deposition, access, archival and dissemination.

Integration with Current and Future Instrumentation

Structural biology increasingly harnesses and integrates multiple approaches to attack challenging problems. In addition to protein crystallography, these include electron microscopy, X-ray scattering, and multiscale computational modelling. This creates a real need for integrated, end-to-end data management solutions. Similar services are also being operated that manage data from over 25 instruments at other facilities, including Monash University Micro Imaging platform (Electron Microscopy), the Monash University Medical Proteomics Facility (Mass Spectroscopy) as well as instruments at the Australian neutron source ANSTO.

All this work is built upon the MyTardis data management platform and thus forms a common interface and data organisation strategy.

ACKNOWLEDGEMENTS

Auto-processing and the Store.Synchrotron was partially supported by the National eResearch Collaboration Tools and Resources (NeCTAR - grant “Monash_Synchrotron_Storage_Service”) project (www.nectar.org.au), and the VicNode/RDSI (grant 2013R2.1).

D.A. would like to thank the Australian National Health and Medical Research Council (APP1047004) for partial funding.

Development of Store.Synchrotron was supported by staff at the Australian Synchrotron MX beamline and Scientific Computing groups.

The publication workflow was funded by the Australian National Data Service (ANDS).

S.M. is an Australian Research Council Future Fellow (FT100100690) and she thanks both the Australian Research Council and National Health and Medical Research (Project grants 1025150; 1063786; 1062216) for funding support.

AMB holds a research fellowship from the National Health and Medical Research Council (NHMRC), Australia. Aspects of this work were funded by the Australian National Health and Medical Research Council (DA, APP1047004; CSB, APP513880 and APP1050585).

REFERENCES