Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Federated repositories of X-ray diffraction images

Steve Androulakis,[a] Jason Schmidberger,[a] Mark A. Bate,[a] Ross DeGori,[a] Anthony Beitz,[b] Cyrus Keong,[b] Bob Cameron,[b] Sheena McGowan,[a] Corrine J. Porter,[a,c] Andrew Harrison,[d] Jane Hunter,[e,f] Jennifer L. Martin,[g,h] Bostjan Kobe,[g,h] Renwick C. J. Dobson,[i] Michael W. Parker,[i,j] James C. Whisstock,[a,c,k] Joan Gray,[d] Andrew Treloar,[b,f,l] David Groenewegen,[b,f] Neil Dickson[b] and Ashley M. Buckle[a,k]*

[a]The Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Faculty of Medicine, Monash University, Clayton, Victoria 3800, Australia, [b]ARROW Project, Monash University Library, Monash University, Victoria 3800, Australia, [c]ARC Centre for Structure and Functional Microbial Genomics, Monash University, Clayton, Victoria 3800, Australia, [d]Monash University Library, Monash University, Victoria 3800, Australia, [e]School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Queensland 4072, Australia, [f]ARCHER Project, 700 Blackburn Road, Clayton, Victoria 3168, Australia, [g]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia, [h]School of Molecular and Microbial Sciences, University of Queensland, Brisbane, Queensland 4072, Australia, [i]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, 30 Flemington Road, Parkville, Victoria 3010, Australia, [j]Biota Structural Biology Laboratory, St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia, [k]Victorian Bioinformatics Consortium, PO Box 53, Monash University, Clayton, Victoria 3800, Australia, and [l]Australian National Data Service (Establishment Project), Monash University, Victoria 3800, Australia.

Correspondence e-mail: ashley.buckle@med.monash.edu.au

Received 25 April 2008
Accepted 23 May 2008

There is a pressing need for the archiving and curation of raw X-ray diffraction data. This information is critical for validation, methods development and improvement of archived structures. However, the relatively large size of these data sets has presented challenges for storage in a single worldwide repository such as the Protein Data Bank archive. This problem can be avoided by using a federated approach, where each institution utilizes its institutional repository for storage, with a discovery service overlaid. Institutional repositories are relatively stable and adequately funded, ensuring persistence. Here, a simple repository solution is described, utilizing Fedora open-source database software and data-annotation and deposition tools that can be deployed at any site cheaply and easily. Data sets and associated metadata from federated repositories are given a unique and persistent handle, providing a simple mechanism for search and retrieval *via* web interfaces. In addition to ensuring that valuable data is not lost, the provision of raw data has several uses for the crystallographic community. Most importantly, structure determination can only be truly repeated or verified when the raw data are available. Moreover, the availability of raw data is extremely useful for the development of improved methods of image analysis and data processing.

1. Introduction

The ability to reproduce experiments is a central tenet of the scientific process. Currently, in a typical report of a macromolecular structure determination, statistics are provided for crystallization, data collection, model building and refinement. Coordinates, and more recently structure-factor amplitudes and phases, are also deposited in the Protein Data Bank archive (wwPDB; Berman *et al.*, 2003). While electron-density maps can be calculated using the processed structure-factor amplitudes and calculated phases (Kleywegt *et al.*, 2004), this information is not sufficient to adequately reproduce the experiment, since amplitudes are computationally derived from the raw images and the process of model building inevitably introduces bias. Collecting diffraction data is effectively a 'one-time' destructive experiment, *i.e.* the crystal cannot be retained in perpetuity and often suffers severe radiation damage. Thus, for crystallography experimental reproduction can optimally be carried out using the original diffraction images.

Over the past year, the sentiment for the need for structure-factor and diffraction-data deposition has been echoed by

many in the crystallographic community (Jones & Kleywegt, 2007; Joosten & Vriend, 2007; Jovine *et al.*, 2008) and has been the topic of fierce debate on the CCP4 bulletin board (http://www.ccp4.ac.uk/ccp4bb.php). Indeed, this same issue was the subject of a recent Editorial in *Acta Crystallographica Section D* (Baker *et al.*, 2008). In this article, the authors make the important point that the increasing reliance on automatic methods of data processing can lead to misinterpretations and incorrect models and this can only be rectified by re-evaluating the original diffraction images. In a recent letter published in *Science*, Jones & Kleywegt (2007) argue that in some cases the experimental results of a structure determination can only be validated when all raw protein crystallographic data, *i.e.* diffraction images, have been deposited in an appropriate database.

There are three key benefits to making raw diffraction data available to the crystallographic community.

Firstly, validation of coordinates deposited with the wwPDB may in some cases require access to the raw data (Janssen *et al.*, 2007; Ajees *et al.*, 2007). While it is possible to detect some errors in protein structures by careful inspection of coordinates, statistics and electron-density maps, other potentially serious errors may only be detected by analysing the raw diffraction data. Examples include mis-indexing and incorrect assignment of space group, inappropriate treatment of twinning, overestimation of data quality and resolution, treatment or otherwise of anisotropy in the diffraction patterns, secondary diffraction patterns and radiation damage and absorption corrections. Access to the diffraction images will also give an indication of issues such as diffraction quality throughout the lifetime of exposure: summarized statistics from processing programs can mask much of what has happened during processing. We believe that deposition of raw data is therefore of paramount importance, because the interpretation of the resulting structures in the context of their biological and chemical function relies heavily on the interpretation of electron-density maps and on the accuracy of coordinates.

Secondly, easy access to raw diffraction data will facilitate the development of new or improved methods of data reduction and scaling. Development of methods is particularly important in the context of high-throughput approaches pioneered by structural genomics consortia. In addition, data are often discarded because they cannot be processed using current algorithms (for example, in cases of high mosaicity and/or spot overlap owing to very large unit-cell dimensions or crystal disorder). Making such data available may allow their processing by improved methods in the future.

Finally, the availability of raw data will allow improvements in re-refining published structures as and when new methods become available (Ramachandraiah *et al.*, 2002).

The deposition of raw diffraction data is scientifically important and we believe that doing so will provide significant benefits to the structural biology and wider scientific community. We recognize that the support of scientific journals and the wwPDB will be required to encourage contributors to make raw data available upon publication.

The question then becomes: where and how should the raw data be stored? The options range from one central database to local storage at the researchers' laboratories. An obvious global home is the PDB archive, but such a centralized approach would be very costly and may not be feasible with current funding and resources. Raw data sets can be large, typically between 5 and 100 GB depending on format and the type of compression used. For this reason, data storage in a central repository is technically and economically challenging. At the other end of the spectrum, storage of raw data within individual laboratories offers a relatively simple solution to the problem. However, the nature of research groups (*e.g.* staff turnover, different processes over time, nonstandard operating procedures), unreliable media, the lack of URL persistence and accessibility issues (*e.g.* firewalls) represent serious impediments.

We argue that a federated system might address many of the obvious challenges outlined above for the traditional centralized approach and will satisfy the requirement to maintain local research data in a secure and readily accessible manner. Such a solution is already utilized by the astronomy community to share terabytes of data collected from radio telescopes (Szalay & Gray, 2001; Foster, 2005).

The protein crystallography researchers at our institutions have already made the decision to go down a federated route. We have taken advantage of the changing role of the modern university library, which increasingly archives electronic rather than print media. Thus, projects such as ARCHER (http://archer.edu.au) and ARROW (http://arrow.edu.au) have resulted in new collaborations between scientific research groups and libraries in Australian institutions. In many disciplines, the library is thus emerging as the logical medium-to-long-term caretaker of online repositories of scientific results, where 'results' will increasingly comprise compound objects that link traditional publications to raw and derived data sets and workflows. The ultimate aim is to publish compound scientific objects that encapsulate the complete set of information necessary to enable verification, reproducibility and re-use of a scientific experiment or discovery process.

In order to demonstrate the usefulness and practicality of our envisaged approach for storing raw crystallographic data images, we have utilized existing library Fedora-based (http://www.fedora-commons.org/) repositories run by the libraries at Monash University and the University of Queensland in Australia. We have built tools allowing diffraction images to be deposited in the local repositories and developed metadata-extraction software such that the data-collection experiment can be described in a semi-automated fashion. We call this initiative 'The Australian Repositories for Diffraction Images (TARDIS)' and have created a website (http://www.tardis.edu.au) where the deposition tools can be downloaded freely. The site will also function as a central portal allowing searching and browsing across all registered Australian repositories. Whereas other initiatives, such as the MEDSBIO project (http://www.medsbio.org/), CrystalGrid (http://www.crystalgrid.org) and the eCrystals federation (http://ecrystals.chem.soton.ac.uk), are actively engaging the

community with the issue of data archival and format standardization, we believe that our approach represents the first real repository solution for the protein crystallography community.

2. Implementation

An overview of the software tools and typical workflow is given in Fig. 1. *Dataset Tools* consists of four separate tools that allow a user to upload collections of diffraction images (termed 'data sets') into a repository-based persistent storage medium. Once projects are in a repository, it is intended that they are accessible to the outside world through the internet and able to be harvested by the upcoming TARDIS web application.

A typical workflow is as follows. A user would start by annotating their project data by opening *PROJECT DESCRIPTOR*. The user inputs basic details about an entire project, such as Project Title and Authors. Once executed, *PROJECT DESCRIPTOR* creates a Fedora-compatible XML description file conforming to the Metadata Encoding and Transmission Standard to be ingested into Fedora along with the data (Fig. 1).

Owing to the large size of data sets, several methods have been implemented to mould the data into a more repository-suitable format. *DATASET PACKAGER* is a program that performs several procedures on a set of images. The term 'packaging' when referring to a data set is the process of converting a set of diffraction images (a 'data set') into a repository-suitable format complete with technical metadata that describe the image set.

Firstly, the data set is packaged together into a single large file (using tar archiving) and then compressed using the bzip2 algorithm. Typically, bzip2 compression of a 4 Gb tar archive takes approximately 11 min on a high-performance workstation (2 × 3 GHz Quad-Core Intel Xeon CPUs). The bzip2 compression algorithm minimizes upload times to the repository and storage requirements on the server by reducing the package to as little as one third of its original size. However, compression alone is sometimes not enough to minimize file sizes, especially for very large data sets where the compressed file size may still be cumbersome. During testing, it was discovered that the repository software was prone to failure when dealing with individual files larger than 2.0 GB. To solve this problem, maximum file sizes are specified within *DATASET PACKAGER*. For example, if a compressed archived data set file is 8 GB and the maximum split file size is set to 1.8 GB, *DATASET PACKAGER* will produce five files as a split file set; the first four will be 1.8 GB and the fifth file will be the remaining 0.8 GB, ready to be deposited into a repository.

Once *DATASET PACKAGER* has packaged the diffraction images, technical information is extracted from the original diffraction image files and written as XML, conforming to a custom 'data sets' schema (Fig. 1). This XML metadata will be exposed by the repository, along with the

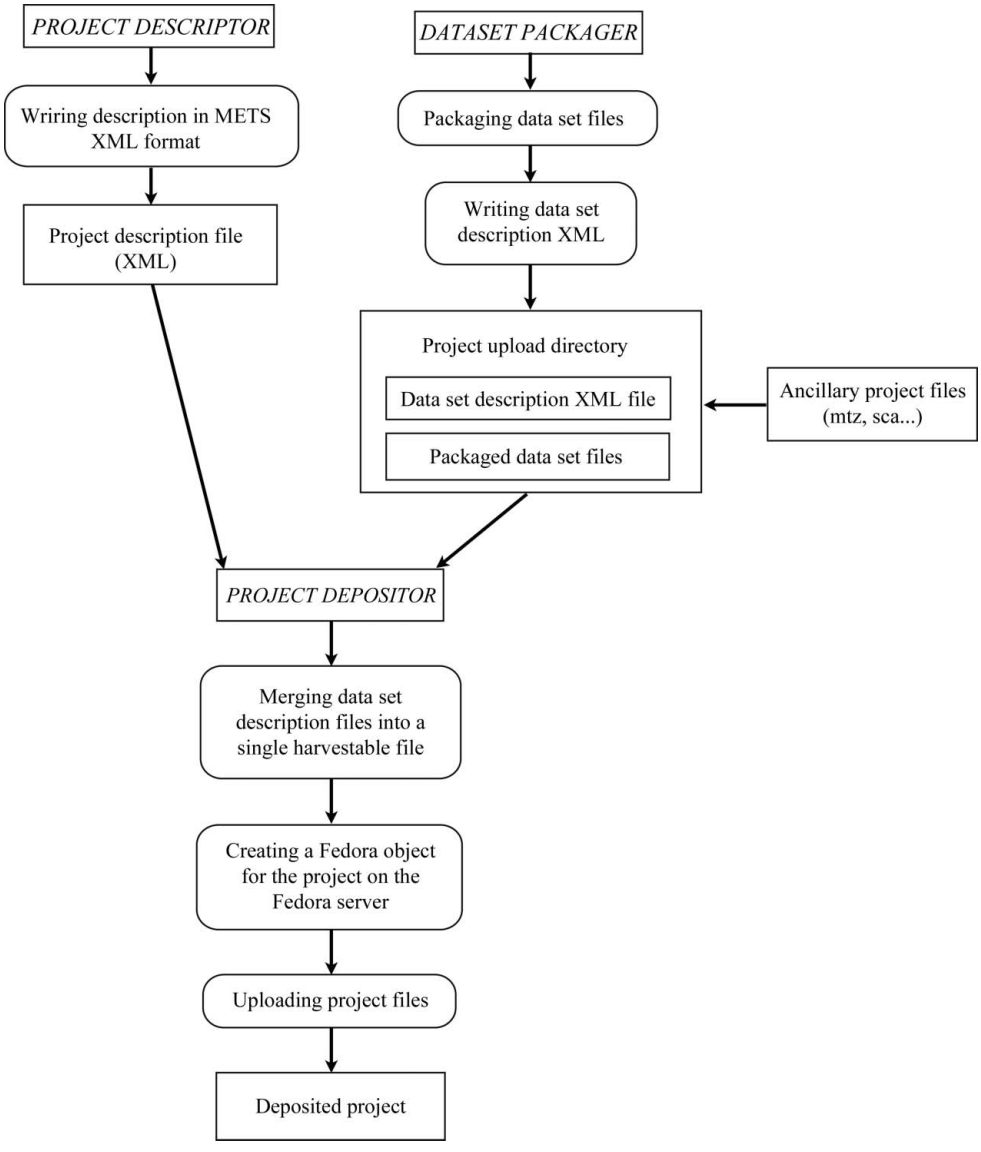

Figure 1
Overview of software tools and typical workflow.

more general project-based data, for indexing and searching through *TARDIS*. Metadata such as detector type, wavelength and rotation range are automatically extracted from each image using the program *DIFFDUMP*, which is part of the *XIA*2 software package (http://www.ccp4.ac.uk/xia/). Once this has been completed for a data set, *DATASET PACKAGER* is able to calculate derived information such as total rotation range and the number of images, ready to be written to XML. *DATASET PACKAGER* supports all image formats currently supported by *DIFFDUMP*. At this stage a user can move additional (smaller) project files that do not require packaging, such as unmerged intensities and ancillary files that the user wishes to deposit in the repository, so that they are uploaded along with the data sets. For example, users are encouraged to deposit data-processing log files, text files describing the experiment and a description of software versions.

Once all packaged data sets and additional files are in one directory on the local file system and a project-description file has been created, the project is ready to be deposited into a Fedora repository. *PROJECT DEPOSITOR* allows a user to specify the project-description file and the directory to be uploaded. Once executed, a new object in Fedora will be created based on the ingested project-description file. The files will then be automatically uploaded as Fedora 'datastreams' within the created object and all technical metadata will be compiled together into a format able to be exposed for harvest. Upon completion, the browser will automatically launch, showing the Fedora index page for the deposited project.

Data sets that are re-downloaded from the repository need to be unpackaged again using *DATASET UNPACKAGER* to restore them to their original format.

It is possible to add ancillary files to existing projects using the official Fedora client (http://prdownloads.sourceforge.net/fedora-commons/fedora-2.2.1-installer.jar). In such cases, an existing project in the repository can be selected, allowing new files (contained in Fedora as 'datastreams') to be added.

3. Technical aspects

The four desktop applications *PROJECT DESCRIPTOR*, *DATASET PACKAGER*, *PROJECT DEPOSITOR* and *DATASET UNPACKAGER* were developed in Java using the *Java Development Kit* 1.5, making them platform-independent.

The applications were designed to deposit packaged annotated data into a Fedora repository server. Fedora repository software is a free open-source solution for digital storage used by many university libraries and provides a flexible extensible back-end storage solution exposed as a set of web services (Lagoze *et al.*, 2006). Several applications exist to provide usable front-ends to the server such as *Fez* (http://sourceforge.net/projects/fez) and *VITAL* (http://www.vtls.com/products/vital).

All software tools are open source and hosted on SourceForge (http://sourceforge.net). Additionally, a user guide and a video are provided to guide users through the process of using the tools and also for setting up a compatible Fedora repository.

3.1. Metadata schema

A standards-based approach to description, preservation and access to the data sets has been implemented for this investigation. Initially, standards such as the CCLRC (Council for the Central Laboratory of the Research Councils; http://epubs.cclrc.ac.uk/bitstream/485/csmdm.version-2.pdf) scientific metadata model as well as more universal standards such as Dublin Core (http://dublincore.org/), MARC (MARC Standards, Library of Congress; http://www.loc.gov/marc/), PREMIS (Preservation Metadata: Implementation Strategies Working Group; http://www.oclc.org/research/projects/pmwg/premis-dd.pdf) and JHOVE (JSTOR/Harvard Object Validation Environment; http://hul.harvard.edu/jhove/) were examined as part of this investigation.

Metadata automatically captured when the data files are stored conforms to a native XML schema. A sample XML file based on a subset of the CCLRC schema was produced for describing projects/data sets. In addition, sample mapping and transformation from CCLRC to Dublin Core and MARC were created to show that this relevant data could be extracted.

As we were attempting to describe more complex objects (an aggregation of data sets) with multiple related components requiring their own descriptions (aggregations of data files), we found that the METS standard (Metadata Encoding and Transmission, Library of Congress; http://www.loc.gov/standards/mets/) provided an excellent way to encode and package these objects for ingest into repositories.

4. Current use and future development

Currently, there are ten data sets in TARDIS, representing >80 GB of raw data. The TARDIS system currently only manages X-ray diffraction images. In reality, the process of solving a protein crystal structure begins much earlier with target selection, high-throughput cloning and protein expression and purification. The TimTam system (http://www.itee.uq.edu.au/~eresearch/projects/crystallography/prototypes.html) under development at the University of Queensland is a laboratory information-management system that captures all of the experimental data and laboratory information that occurs prior to crystallization and X-ray diffraction. One future aim is to link the experimental data captured in TimTam to the crystallographic image archive to provide an end-to-end data management system for protein crystallographers. Specific work in progress is described below.

4.1. CCLRC schema adaptation

In future releases, *Dataset Tools* will hold all of its data in an adaptation of the CCLRC Scientific Data Model XML schema. From this, all specialized metadata for repositories, data harvesting and also data sets will be derived from the data

held within this model. This represents an advance in managing data in a way that is generic, repository-agonistic and standardized.

4.2. SWORD implementation

It is proposed in future releases of *Dataset Tools* that the SWORD (Simple Web-service Offering Repository Deposit; http://www.ukoln.ac.uk/repositories/digirep/index/SWORD) project's *Deposit API* will be used to deposit data. SWORD allows deposition of data into many different kinds of repository software, removing the exclusive ties of *Dataset Tools* to the Fedora repository platform and increasing its compatibility with more institutional library systems.

4.3. TARDIS portal

The TARDIS website will also function as a central portal that allows browsing and searching of raw crystallographic data images across all registered Australian repositories. TARDIS will routinely gather and index data-set information in a central database. Indexing will be achieved through data exposed by Fedora using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) framework. Users can access the TARDIS website and search for data sets based on any number of project or data-set variables. From returned results, TARDIS will allow the user to download the data sets directly, without needing to navigate to an external website. This will enable new repositories (with crystallographic data sets) to be indexed and their metadata harvested, enabling their content to be searchable by the crystallography community.

5. Conclusions

We have created 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. The availability of diffraction images to the macromolecular crystallographic community will ensure that valuable data are not lost, enable a structure-determination procedure to be truly reproduced and facilitate the development of improved methods of image analysis and data processing. The need for the deposition of raw data has recently been intensely debated within the crystallographic community and we trust that scientific journals and the wwPDB will encourage researchers to make such data available.

6. Documentation and availability

All software and documentation can be accessed at the TARDIS website (http://www.tardis.edu.au/)

We thank Ruby Law, Neil Saunders and Anil Thakur for helpful discussions. We thank the NHMRC, ARC, Victorian Partnership for Advanced Computing, the Victorian Bioinformatics Consortium, Monash e-Research Centre, and the state government of Victoria (Australia) for funding and support. ARROW and ARCHER are funded by the Australian Commonwealth Department of Education, Science and Training (DEST) through the Systemic Infrastructure Initiative (SII), a part of Backing Australia's Ability – An Innovation Action Plan for the Future. AMB and JLM are NHMRC Senior Research Fellows. JCW is an ARC Federation Fellow and Monash University Logan Fellow. JH is an ARC Professorial Research Fellow. BK and MP are ARC Federation Fellows and NHMRC Honorary Research Fellows. CJP is an NHMRC (Peter Doherty) Training Fellow.

References

Ajees, A. A., Gunasekaran, K., Narayana, S. V. L. & Krishna Murthy, H. M. (2007). *Nature (London)*, **448**, E2–3.
Baker, E. N., Dauter, Z., Guss, M. & Einspahr, H. (2008). *Acta Cryst.* D**64**, 337–338.
Berman, H. M., Henrick, K. & Nakamura, K. (2003). *Nature Struct. Biol.* **10**, 980.
Foster, I. (2005). *Science*, **308**, 814–817.
Janssen, B. J., Read, R. J., Brunger, A. T. & Gros, P. (2007). *Nature (London)*, **448**, E1–2.
Jones, T. A. & Kleywegt, G. J. (2007). *Science*, **317**, 194–195.
Joosten, R. P. & Vriend, G. (2007). *Science*, **317**, 195–196.
Jovine, L., Morgunova, E. & Ladenstein, R. (2008). *J. Appl. Cryst.* **41**, 659.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Lagoze, C., Payette, S., Shin, E. & Wilper, C. (2006). *Int. J. Digit. Libr.* **6**, 124–138.
Ramachandraiah, G., Chandra, N. R., Surolia, A. & Vijayan, M. (2002). *Acta Cryst.* D**58**, 414–420.
Szalay, A. & Gray, J. (2001). *Science*, **293**, 2037–2040.