Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

High-throughput sample handling and data collection at synchrotrons: embedding the ESRF into the high-throughput gene-to-structure pipeline

A. Beteva,[a] F. Cipriani,[b] S. Cusack,[b] S. Delageniere,[a] J. Gabadinho,[a] E. J. Gordon,[a] M. Guijarro,[a] D. R. Hall,[a] S. Larsen,[a] L. Launer,[c] C. B. Lavault,[b] G. A. Leonard,[a] T. Mairs,[a] A. McCarthy,[b] J. McCarthy,[a] J. Meyer,[a] E. Mitchell,[a] S. Monaco,[a] D. Nurizzo,[a] P. Pernot,[a] R. Pieritz,[a] R. G. B. Ravelli,[b] V. Rey,[a] W. Shepard,[a] D. Spruce,[a] D. I. Stuart,[d] O. Svensson,[a] P. Theveneau,[a] X. Thibault,[a] J. Turkenburg,[e] M. Walsh[c] and S. M. McSweeney[a]*

[a]ESRF, 6 Rue Jules Horowitz, 38043 Grenoble, France, [b]EMBL, 6 Rue Jules Horowitz, F-38042 Grenoble CEDEX 9, France, [c]MRC-France (BM14), 6 Rue Jules Horowitz, F-38042 Grenoble CEDEX 9, France, [d]Oxford Protein Production Facility and Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, and [e]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England

Correspondence e-mail: seanmcs@esrf.fr

Received 3 November 2005
Accepted 16 August 2006

An automatic data-collection system has been implemented and installed on seven insertion-device beamlines and a bending-magnet beamline at the ESRF (European Synchrotron Radiation Facility) as part of the SPINE (Structural Proteomics In Europe) development of an automated structure-determination pipeline. The system allows remote interaction with beamline-control systems and automatic sample mounting, alignment, characterization, data collection and processing. Reports of all actions taken are available for inspection *via* database modules and web services.

1. Introduction

Since the analogy was first made, it has become something of a cliché to describe the process of determining macromolecular crystal structures as a pipeline (Stevens & Wilson, 2001). Nevertheless, this analogy is appealing because macromolecular structure determination consists of a number of relatively simple steps. In many ways, the process may be thought of as a simple sequence of operations: crystallization, diffraction, phasing, model building, refinement and deposition of coordinates. That this sequence is well understood leads one to hope that all the steps of the pipeline may be readily automated. The reality is of course much more complicated since the process of proceeding from crystal to structure deposition is more labyrinthine than the analogy would suggest. However, the strength of the analogy is that one is forced to think about the critical actions and processes that contribute to the successful outcome of any particular section of the process. Thus, the most critical components of the pipeline can be identified and appropriate mechanisms and measures put in place to monitor their successful completion.

The development of a data-collection pipeline (DCP) forms the bulk of the deliverables for workpackage 6 'High-throughput Synchrotron Facilities' of the SPINE (Structural Proteomics In Europe) project. The goals of the workpackage were to automate data collection at synchrotron beamlines as well as putting in place protocols to automatically align optical elements, improve beamline diagnostics and optimize the usage of available beamtime. The developments described here have evolved as a collaborative effort between the main partners of this workpackage, EMBL and ESRF, as well as the involvement of other similar European and national initiatives, in particular the EU-funded BIOXHIT project, the UK Biotechnology and Biological Sciences Research Council (BBSRC) funded e-HTPX project (http://www.e-htpx.ac.uk), the UK-funded BM14 MAD beamline and the DNA collaboration (http://www.dna.ac.uk).

Provision of automated sample mounting realised by the design and implementation of a robotic sample changer achieved within the SPINE project has formed an important starting point for the implementation of the DCP and is described in more detail by Cipriani *et al.* (2006). The development of the DCP within SPINE was also initiated at a time when a number of national and international structural genomics programmes were under way. In the light of this situation, the DCP has, wherever possible, used standards for hardware (*e.g.* SPINE standards; Cipriani *et al.*, 2006), software (*e.g.* those developed within the DNA collaboration), data models elaborated by the EBI within e-HTPX (http://www.e-htpx.ac.uk) and BIOXHIT that have been agreed at a national or international level. Using such standards means that a scientist using our (ESRF) DCP should also be able to exploit DCP systems developed at other synchrotron facilities.

A DCP at a synchrotron facility is composed of a number of closely coupled steps (Fig. 1). Ideally, as a sample passes through the pipeline, human intervention should be minimal and should occur only in the transport of samples to and from the facility and in the loading/unloading of the samples in the beamline equipment (*i.e.* the sample changer). It is thus instructive to consider those portions of the DCP that may be automated and how this automation may be brought about. The DCP has its roots at the sample-preparation stages before any samples have been sent to the synchrotron. Safety approval of the proposed experiments is necessary before an experiment can be undertaken. With the development of laboratory information-management systems (LIMS), home laboratories will have the information necessary for safety assessments of the crystal samples available in electronic form. This information should be exploited so that robust sample safety validation is implemented in a flexible and automated fashion.

A key section of the DCP is the ability to automatically screen and assess the quality of a consignment of samples. This process requires considerable book-keeping. The DCP developed at the ESRF utilizes a LIMS developed within the SPINE project to keep track of individual samples using a unique sample-holder identifier (Cipriani *et al.*, 2006). Moreover, the development of the LIMS has been kept generic and in keeping with this a large collaborative effort is in place between SPINE, BIOXHIT, e-HTPX and DNA to accelerate developments following an agreed data model. Information on any sample must be linked to experiment requirements (initial screening, MAD, SAD, ligand studies, minimum diffraction limit *etc*.), experiment planning, screening output (particularly data-collection strategies) and any eventual data-collection and processing results. If many samples are to be screened, a robotic sample changer (Snell *et al.*, 2004; Cohen *et al.*, 2002; Ohana *et al.*, 2004) and software capable of automatic alignment of samples to the X-ray beam (Andrey *et al.*, 2004) are essential and development of these has been funded by SPINE. Contained within the screening process are a set of decision-making points that must be based on user input. Ultimately, the DCP should be able to act on this advice and screen and assess samples to find those most suited to the experimental requirements and priorities, optimize data-collection strategies and then carry these out automatically. However, in the near future it is unlikely that such a complex system will work without manual intervention. It is much more likely that a hybrid system will be used where sample screening is performed and the results returned to the home laboratory where an assessment is made and updated data-collection priorities are assigned. Data collection would only commence once this manual assessment of results had been made. One might imagine an initial delay of up to 2–3 days between sample screening and data collection (see §4.1). Further enhancements of the DCP will certainly arise as links with software packages developed with input from SPINE (see Bahar *et al.*, 2006) as well as other initiatives that automatically reduce crystallographic data, produce phasing information and subsequently build a three-dimensional model of the target under investigation. Such software packages including *SOLVE*/*RESOLVE* (Terwilliger & Berendzen, 1999; Terwilliger, 2004), *PHENIX* (Adams *et al.*, 2002, 2004) and *Auto-Rickshaw*/*ARP*/*wARP* (Panjikar *et al.*, 2005; Lamzin & Perrakis, 2002) are increasingly complete, efficient and rapid. An assessment of the current status of this part of the structure-determination pipeline is given in Bahar *et al.* (2006).

An important part of the X-ray diffraction experiment is of course the provision of the X-ray beam. Many advances are being made in the establishment of control systems which will allow the automatic delivery of appropriately configured X-ray beams. This work is the subject of active development, but is too complicated for adequate discussion here. The interested reader is referred to a recent review on the subject (Arzt *et al.*, 2005). In this paper, we confine ourselves to an elaboration of the crystallographic DCP implemented on ESRF macromolecular crystallography (MX) beamlines. A number of other initiatives have successfully automated portions of the DCP we report, including the X-ray beam provision (Arzt *et al.*, 2005; Pohl *et al.*, 2004; Gaponov *et al.*, 2004), sample exchange (Snell *et al.*, 2004; Cohen *et al.*, 2002), data processing (Leslie *et al.*, 2002; Ferrer, 2001; Kroemer *et al.*, 2004; Otwinowski & Minor, 1997) and the logging of the sample information (Ueno *et al.*, 2004). The layout of this paper is as follows: §2 presents results related to the implementation of a crystallographic DCP at the ESRF and §3 presents the methods deployed to create various components of the DCP, while §4 discusses plans and ideas for several future developments for the DCP.

2. Results

A DCP has been developed and deployed on seven insertion-device beamlines and one bending-magnet beamline at the ESRF. The principle components of the DCP and the interconnections between these are shown schematically in Fig. 2.

2.1. The DCP information-management system

Management of experimental data in all areas of scientific research is an important part of ensuring greater efficiency in

the workplace. Many LIMS have been developed that allow sample tracking and reporting of sample information and these have already had a positive impact on structural biology research. For example, a LIMS for protein production (PIMS) is currently in development (Pajon *et al.*, 2005) and a LIMS has been developed for protein crystallization (Mayo *et al.*, 2005). Completing the pipeline series is a LIMS for diffraction data collection which has been developed at the ESRF in collaboration with the SPINE and e-HTPX projects and given the acronym ISPyB (**I**nformation **S**ystem for **P**rotein Crystallography **B**eamlines).

Viewed from the LIMS perspective, the process of collecting X-ray diffraction data from single macromolecular crystals at a synchrotron beamline has been divided into four steps. Firstly, safety approval is obtained and then the samples (as well as sample information pertinent to the actual experiment) are dispatched to the synchrotron. Diffraction data are then acquired and analysed and finally these are returned to the home laboratory. ISPyB, as currently deployed at the ESRF, allows the logging of information concerning sample data, sample shipping and data collection and allows the harvesting of data-reduction statistics from *DNA* (Leslie *et al.*, 2002; §3.4).

Sample data can be input into ISPyB at different levels of sophistication depending on the user's needs.

(i) Direct manual input *via* a web-based browser interface.

(ii) Input of data using a personal digital assistant (PDA) device that runs a pocket version of ISPyB developed to provide a portable desktop solution for storing data as you mount your crystal. Data is stored on the PDA for future upload to ISPyB. The PDA-based software uses a wireless barcode reader and provides a simple way to link upstream information stored on the sample, such as crystallization conditions, to the ISPyB LIMS.

(iii) Robotic crystallogenesis systems can produce large numbers of samples and LIMS exist to manage these data. To minimize manual sample data-input web services have been employed to allow users to automatically transmit relevant information from their LIMS to ISPyB. Web services provide a means for computers to interact with each other over the internet using standard protocols. Security problems and access through firewalls are overcome by using the Hypertext Transfer Protocol. Currently, at the ESRF web services are available to allow users to transmit crystal details, send information of a shipment of samples *via* courier, submit a diffraction plan and to retrieve diffraction experimental results from ISPyB.

(iv) Once a user logs into ISPyB, he/she can the retrieve sample information stored in the ESRF User Office database.

The sample-shipping module of ISPyB (Fig. 3) allows the user to prepare a shipment of samples to the synchrotron for data acquisition. A shipment consists of one or more dry-shipper dewars which can be filled with samples that have been mounted in sample holders which in turn are held in specified containers (*e.g.* a cryocane or basket for use with a robotic sample

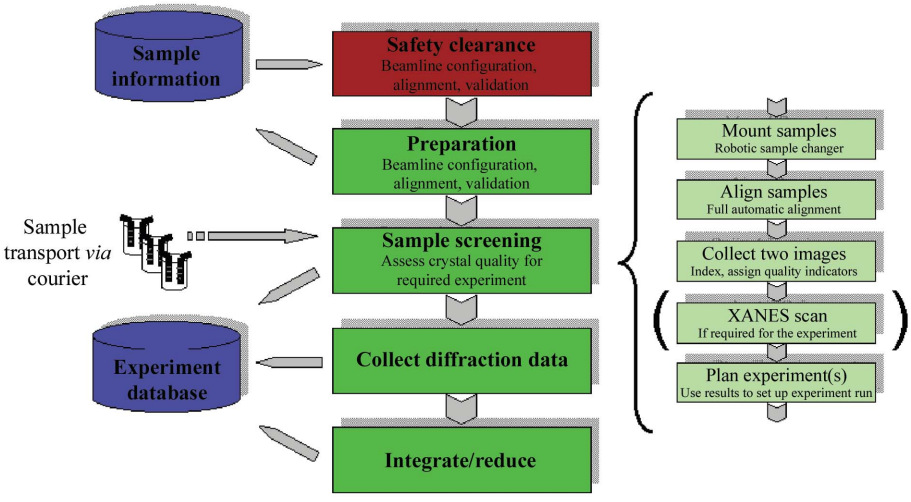

Figure 1
The typical MX experiment. Schematic representation of an experiment from beamline configuration through to final data deposition, with the core parts of the DCP highlighted in green, experimental safety control being the final control before data collection (highlighted in red). MX experiments normally follow a well established format which permits a systems approach to integrating and automating the steps. Data tracking will mean the integration of home-laboratory databases for pipelining information to and from the synchrotron.

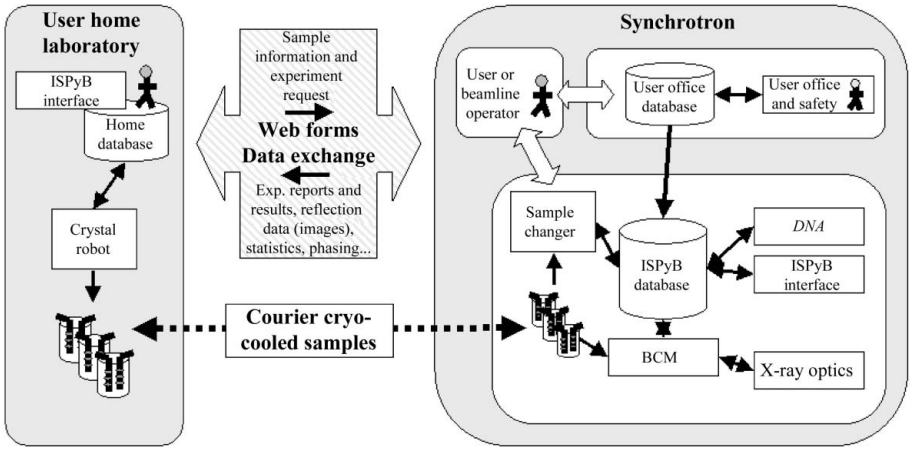

Figure 2
Schematic of DCP for a user experiment. The schematic shows how a typical user of the DCP may interact with synchrotron facilities using web services to transmit and receive XML-packaged data describing samples and experimental requirements. BCM, beamline-control module.

changer). In cases where the users utilize the SPINE standard sample holder (see http://www.spineurope.org; Cipriani *et al.*, 2006), ISPyB can be provided with the unique sample-holder identifiers.

The data-collection module of ISPyB harvests experimental results for each data collection performed (Fig. 4; a data collection is considered to be one or more exposures). In addition to the experimental parameters (beamline used, start and end points for data collection, oscillation range, exposure time, wavelength, crystal-to-detector distance *etc.*), JPEG thumbnails linked to high-resolution JPEGs are also generated. A typical summary of data collections or test shots for a particular project is shown in Fig. 4. At all stages, summaries can be generated and output in PDF format for further reference.

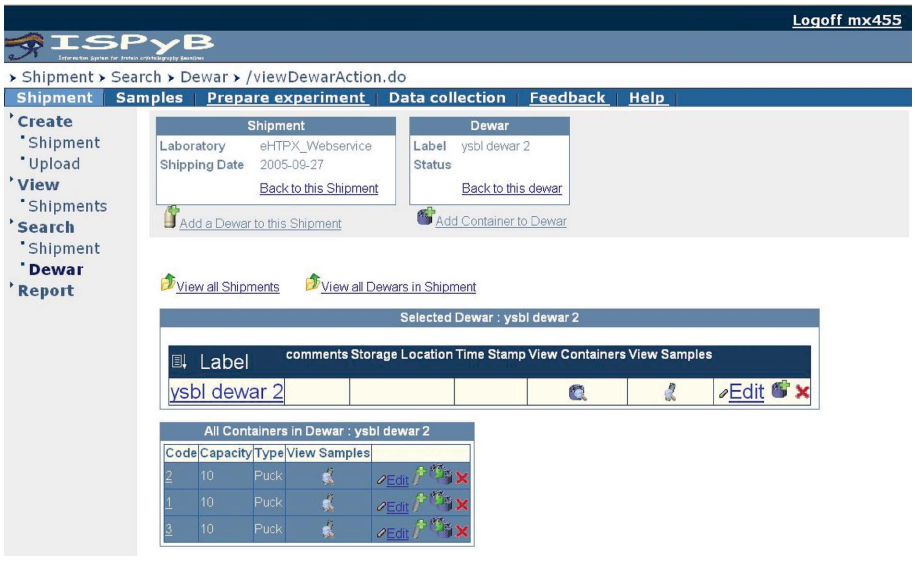

Figure 3
The ISPyB dewar-shipment interface. The program allows the experimenter to organise the sample information for a number of shipping dewars. Provision is made for shipments containing both sample-changer pucks and crystals on canes.

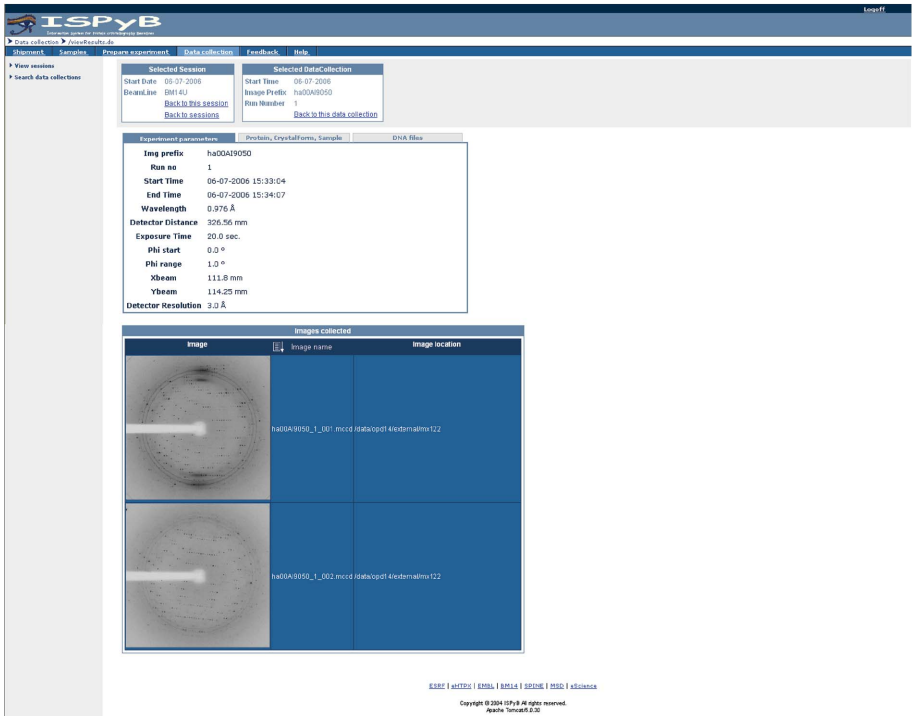

Figure 4
During a data collection, ISPyB maintains a gallery of thumbnails of the diffraction images. In addition, a link to the *DNA* output is available, in case closer examination of the software decision chain is needed.

2.2. Instrument-control system

The creation of a control software framework, which brings together all the components in a form which is simple to control and efficient, was a major part of the development of the DCP. This overarching control software distributes the management of hardware over several computers whilst maintaining a principal control point on one machine where all the necessary information can be displayed and accessed. The overall product desired was an 'industrial-style' environment allowing both local and remote control for high-throughput experiments. These developments have also benefited again from collaborations between ESRF, SPINE, BIOXHIT and e-HTPX. The control software was written to make it easily distributable across all ESRF MX beamlines with minimum re-configuration and addresses the following aspects of instrument control.

(i) Automation of beam delivery to the sample, including beamline alignment, monochromator optimization and mirror focusing.

(ii) Automation of data collection for fixed- and multiple-wavelength experiments with XANES scans of absorption edges as necessary.

(iii) Information flow and bookkeeping for the high-throughput environment linked to ISPyB.

(iv) Automatic analysis of data by interaction with *DNA*.

(v) Manual pipeline operation combining automatic sample mounting and centring, sample screening, multi-sample collection, on-line data analysis and database recording.

(vi) Beamline diagnostics in real time (for example, shutter-synchronization information) or in a record mode (for

example, historical database) that allows off-line analysis of events.

(vii) Full DCP operation using a sample changer. This daisy-chains automatic sample mounting and dismounting, automatic sample centring, automatic sample screening and automatic on-line data analysis and logging of results.

2.3. Sample changer and standardization of sample holders

Considerable investment has been made into the provision of automatic sample changers (SCs) on all the beamlines covered by this report. These SCs offer a different means of sample exchange from other devices that have been developed (Cohen *et al.*, 2002; Ohana *et al.*, 2004) and they also benefit from the availability of a SPINE standard sample holder (pin and cap) and have been optimized for use with these systems. For further details, see Cipriani *et al.* (2006).

2.4. Experience gained through operation of the DCP

When in DCP mode, the beamline hardware-control system initiates the pipeline process with a scan of the contents of the sample changer, during which it identifies all samples using the two-dimensional barcode on the bases of the SPINE sample holders. Relevant data are then extracted from the ISPyB database, a check is made of safety status and pertinent data are then displayed in the relevant GUI. On command each sample is mounted, aligned, a diffraction test made using *DNA* and the relevant results stored. By a process of beamline crash testing and the availability of a dedicated team of software developers and MX beamline scientists a DCP has been established that aims to be robust and user-friendly. However, in practice there are a number of key areas where the pipeline can break and a failure can occur. Careful analysis of the hundreds of hours of SC operation and pipeline evaluation have led to the identification of the following chief causes of pipeline failure or leakage.

(i) Sample jamming and mount/dismount failures. This is caused primarily by vial jams inside the sample changer, normally owing to the use of non-SPINE standard caps and/or misalignment of the sample changer. The use of SPINE standard caps and vials is compulsory for secure operation of the sample changer. Protocols to maintain the alignment of the sample changer have been put in place.

(ii) Crystal is not found in the loops. Simple loop centring is robust and fast, but determination of a crystal's position within a loop is more difficult (Andrey *et al.*, 2004). In general, crystal recognition operates at better than 80% reliability (Lavault, personal communication). The remaining 20% of problems occur mostly with large loops and small crystals surrounded by large amounts of cryoprotectant solution. It is difficult to see how to avoid these problems completely, although appropriate mounting of the sample greatly increases the success rate. Investigations have been made into the utility of ultraviolet illumination as a general means of visualizing protein crystals (Pohl *et al.*, 2004; Vernede *et al.*, 2006).

(iii) Insufficient diffraction quality. The criteria for successful indexing by *DNA* are intentionally high, the aim being both to avoid the collection of poor data and to correctly identify when human intervention is needed. With further experience, these criteria will be refined.

Nevertheless, at the end of this development period a DCP has been put in place to allow automatic beamline operation using components that conform to agreed standards and which provides the following components.

(i) Web services for sample submission.

(ii) Web services to allow beamtime adminstrative tasks and safety assessment.

(iii) Automated X-ray beam delivery and experiment configuration.

(iv) Sample changing utilizing standardized sample holders and containers.

(v) Loop and crystal-centring system allowing automatic sample positioning.

(vi) Software for the evaluation of diffraction quality and estimation of optimal data-collection strategies.

(vii) Storage of all experiment information in a database for remote observation of experiment progress.

Whilst the data-collection environment will undoubtedly continue to improve, increased exposure to the DCP will allow increased reliability of operation and will lead to new developments that will allow exploitation of the tools available.

3. Methods

3.1. ISPyB

The development of ISPyB was carried out with several constraints to ensure ease of its implementation and use and a long product lifetime. To achieve this, ISPyB has a web-based user interface which uses a look and feel that is common to other applications. The functionalities of the system are presented in an intuitive and easy-to-use manner. Importantly, the web-based interface facilitates integration into existing systems and ease of accessibility from outside a firewall-protected network. The software architecture itself follows a standard three-tier model. The information is stored in a MySQL (http://www.mysql.com/) relational database, the application logic is developed in Java according to the J2EE model (Singh *et al.*, 2002) and the presentation layer is made through Java Server Pages (JSP).

3.2. Instrument-control system

The software controlling MX beamline optics and experiments is based on the classic three-level approach used at all the ESRF beamlines and other facilities (McPhillips *et al.*, 2002). This includes the following.

(i) Front-end software: a first layer devoted to instrument control. In this layer, software is distributed over different beamline computers since performance and specialization are crucial. Programs at this level are mainly written in C/C++ and run on Linux or Windows platforms depending on the availability and convenience of the device drivers. Remote access is obtained through ESRF-developed communication layer

abstractions called Taco (older devices) and/or Tango for newer systems (http://www.esrf.fr/Infrastructure/Computing/).

(ii) Sequencer software and servers: an intermediate layer where integration of the different instruments takes place through exposure of the essential features of the hardware. Concentration of access to devices through scripting languages allows easy and flexible definition of data-collection sequences. The main program at this level is *SPEC* (Certified Scientific Software, Cambridge, MA, USA). Some sequencing also uses programs written in Python (http://www.python.org).

(iii) User-level software: user interfaces in the form of graphical applications for control or web-based applications for local and remote access to information are custom-built using Python and the Qt tool kit (http://www.trolltech.com/products/qt).

3.3. Automatic sample centring

Once a sample has been mounted on the goniometer, the crystal must be placed on the intersection of the X-ray beam and the crystal rotation axis. This task has been the topic of considerable research, but as yet no completely reliable system has been developed. The procedure undertaken at the ESRF follows a two-step approach (Andrey *et al.*, 2004). Firstly, the loop holding the crystal is brought onto the centre of rotation and into the X-ray beam. The loop is recognized by image-analysis techniques which, because of the high contrast around the loop, are robust and quick. With the loop centred, a more detailed analysis of the images at various (increasing) optical magnifications is undertaken (Andrey *et al.*, 2004). As mentioned above, this approach proves to be remarkably robust, with 80% of the crystals correctly identified. It should not be forgotten that at sites with relatively large focal spot sizes the initial step of loop centring and alignment may well be sufficient for an initial diffraction screening.

3.4. Analysis of diffraction quality and ranking of samples of the same type

The use of *DNA* (Leslie *et al.*, 2002) is central to our DCP. In order to assess diffraction quality, two images are collected (90° apart in phi). These initial images are examined to determine the effective resolution limit and to check for the presence of strong, ice rings and other unwanted diffraction features. If the quality of diffraction is acceptable, the images are autoindexed with *MOSFLM* (Leslie, 1992) to determine the unit-cell parameters and possible space groups. The accuracy of the indexing solution is checked by comparing the predicted diffraction patterns with those observed on the collected images. This provides an opportunity for detecting the presence of a second (weaker) lattice. Assuming successful autoindexing, the diffraction limit is found and the mosaic spread can be estimated. The data-collection strategy (total rotation angle and oscillation angles) is worked out and an exposure time is chosen using *BEST* (Popov & Bourenkov, 2003; Bourenkov & Popov, 2006), which assumes that the probability density functions for diffraction intensities derived by Wilson (1949) are applicable.

Owing to the speed with which data can be collected on undulator beamlines, it is unlikely that feedback from downstream data-quality monitors can be activated before the data collection has ended. Thus, the use of *DNA* is critical to optimizing the data extracted from a sample. With the availability of reliable automatic sample changers, it becomes natural that an experimenter would like to check and rank the diffraction properties of several crystals of the same macromolecule. After ranking the samples with respect to one another, data collection should proceed using the best (or most appropriate) crystal, with online integration and scaling with *MOSFLM* (Leslie, 1992) and *SCALA* (Evans, 2006); work is in progress to also allow integration with *XDS* (Kabsch, 1988). At the moment, it is not possible to have a unique and universal methodology to rank samples since ranking depends on the data-set characteristics and on the experimental aims. Thus, implementing flexible ranking procedures is a key feature of a good sample-screening scheme. This flexibility is required (i) to change the ranking strategy 'on the fly', (ii) to rank different data sets by different methodologies, (iii) to compare different ranking methods and (iv) to allow evolutionary development of ranking schemes.

The software architecture used to implement the required flexibility in our DCP is based on a 'pure virtual mechanism' developed in an object-oriented programming manner (Booch, 2002). This virtual mechanism allows new ranking methodologies to be coded and added to the data-ranking software without any change in the main architecture. The module design is integrated into the main *DNA* package but may also can be used as an external tool or as an independent module.

4. Future developments of the DCP

4.1. Remote access

Taken together, the components of the DCP provide a platform for external user groups to access synchrotron-based macromolecular crystallography beamlines remotely. In its fully refined form, the DCP will automatically carry out all the steps usually performed manually by experimenters: crystals will be loaded onto (and unloaded from) the host goniometer and centred in the X-ray beam, the *DNA* software will characterize samples, provide detailed information on individual crystals (unit-cell parameters, resolution limits, mosaicity, spot shape *etc.*), choose an optimum data-collection strategy and exposure time based on experimental requirements contained in the ISPyB LIMS, use a ranking mechanism (if required) to choose the best of several crystals and automatically collect, integrate and scale diffraction data. In principle, there will no longer be any need for external user groups to be physically on the beamline while data from their crystals are being collected. Indeed, remote access to beamlines is already possible at some synchrotron sites; for example, SSRL or the SGX CAT at APS. These services correspond to remote control of the beamline

(SSRL) or data collected according to the (remote) users' wishes (SGX). At the ESRF we are considering several possibilities for remote data collection.

(i) Local control, local decisions. Here, users will ship their samples to the ESRF beamlines and ESRF staff (with the help of the DCP) will collect data based on diffraction plans contained in the ISPyB LIMS. ESRF staff will control the beamlines and take all decisions concerning data collection. This method of access to beamlines is analogous to the MXpress system already available at the ESRF (http://www.esrf.fr/Industry/Applications/MX/MXpress).

(ii) Local control, remote decisions. Users ship samples to the ESRF and the DCP is used to perform automatic screening. The results of this are returned (*via* ISPyB) to the users. Based on these, users decide from what crystals data should be collected, how it should be collected and communicate this to ESRF staff *via* ISPyB. The only input of ESRF staff to the experiment is to load samples into the sample changer and to control the beamline in question.

(iii) Remote control, remote decisions. Users ship samples to the ESRF. ESRF staff load them into the sample changer, ensure that the beamline is ready to be used and are on-call in case of problems. External users then control the whole beamline remotely and use of the DCP would not be obligatory.

A further option we plan to make available to external users, at least in the short term, is the option of remote sample screening followed by experimental sessions at which the users will be present and where full data collections will be carried out. In all cases security is ensured by password protection of all accounts and the use of secure servers for data transfer.

4.2. More complex data-collection systems and feedback

Current systems and data models manage only linear stepwise data collections. A more complete data model encompassing information beyond data collection (for example, results from data reduction, phasing results or ligand fitting) would facilitate the development of intelligent software to optimize the experimental strategies for data collection dynamically. A significant number of experiments, especially with challenging samples, would benefit from a more complete and detailed experimental strategy. For example, a kappa goniometer can be invaluable for MAD or SAD data collections or for data collection from crystals with a very large unit-cell parameter. Many samples also diffract to very high resolution and today's generation of X-ray diffraction detectors have insufficient dynamic ranges to handle the very strong low-resolution data concurrently with weak high-resolution data. The data-collection pipeline should be able to recommend suitable data-collection sweeps to obtain complete data from such crystals. There is increasing interest in microfocus X-ray beams which entail their own particular set of strategies. For example, many microcrystals in one loop (entailing uniquely identifying those used for data collection) or multiple parts of the same crystal(s) may be used to compile a complete data set. This means software should be able to take into account previously collected data and calculate an optimized strategy for completion of the data set. A classic example of dynamic feedback is the instance of radiation damage. DCP systems must be able to identify when a sample is suffering radiation damage that will prevent the successful outcome of the experiment (this is crucial in *de novo* phasing measurements). Such feedback would need to be returned after data evaluation (post-scaling and indeed post-phasing trials) and would require data-processing and scaling protocols capable of keeping pace with the ESRF data-collection rates (currently up to 45 frames min^{-1}) to be implemented.

5. Conclusion

The SPINE project has made a central contribution to implementing a DCP at the ESRF which brings together a number of existing components. The formal framework of an agreed data model has been crucial to enabling the work to progress. The inter-synchrotron and inter-institute collaborations made around all of the DCP components have been invaluable and form the basis for common ground allowing developments to be shared. A functional DCP will be a major step forward for MX and will lead to a radical change in the way beamlines are used. In particular, experiments will not be constrained by the available beamtime but by the availability of samples for screening. It is likely that real improvements in data (and structure) quality will be obtained, as it will be possible to search for the best crystal rather than as is currently the case the most acceptable given the time available. The DCP will also enable 'user-less' functioning of beamlines with initial screenings of crystals being carried out automatically for routine samples. Whole beamlines could be dedicated to this mode of function whilst other beamlines become more specialized to experiments that are less amenable to automatic operations. While there remains potential for development of the DCP described in this paper, it already has demonstrated functionalities that will improve the throughput, ease of use, efficiency and quality of data collected at the ESRF MX beamlines. It is expected to be available at the ESRF to the MX user community soon.

This work formed part of the SPINE (Structural Proteomics In Europe) project, contract No. QLG2-CT-2002-00988, funded by the European Commission under the Integrated Programme 'Quality for Life and Management of Living Resources'. Additional support was provided by ESRF, EMBL and MRC-France (BM14).

References

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C. & Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 53–55.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Andrey, P., Lavault, B., Cipriani, F. & Maurin, Y. (2004). *J. Appl. Cryst.* **37**, 265–269.
Arzt, S. *et al.* (2005). *Prog. Biophys. Mol. Biol.* **89**, 124–152.
Bahar, M. *et al.* (2006). *Acta Cryst.* D**62**, 1170–1183.
Booch, G. (2002). *Object-Oriented Analysis and Design with Applications*, 2nd ed. Boston: Addison-Wesley.
Bourenkov, G. P. & Popov, A. N. (2006). *Acta Cryst.* D**62**, 58–64.
Cipriani, F. *et al.* (2006). *Acta Cryst.* D**62**, 1251–1259.
Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *J. Appl. Cryst.* **35**, 720–726.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Ferrer, J. L. (2001). *Acta Cryst.* D**57**, 1752–1753.
Gaponov, Y., Igarishi, N., Hiraki, M., Sasajima, K., Matsugaki, N., Suzuki, M., Kosuge, T. & Wakatsuki, S. (2004). *J. Synchrotron Rad.* **11**, 17–20.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kroemer, M., Dreyer, M. K. & Wendt, K. U. (2004). *Acta Cryst.* D**60**, 1679–1682.
Lamzin, V. S. & Perrakis, A. (2002). *Acta Cryst.* A**58**, C56.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
Leslie, A. G. W., Powell, H. R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. & Nave, C. (2002). *Acta Cryst.* D**58**, 1924–1928.
McPhillips, T. M., McPhillips, S. E., Chiu, H.-J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M. & Kuhn, P. (2002). *J. Synchrotron. Rad.* **9**, 401–406.
Mayo, C., Diprose, J., Walter, T., Berry, I., Wilson, J., Owens, R., Jones, E., Harlos, K., Stuart, D. & Esnouf, R. (2005). *Structure*, **13**, 175–182.
Ohana, J., Jacquamet, L., Joly, J., Bertoni, A., Taunier, P., Michel, L., Charrault, P., Pirocchi, M., Carpentier, P., Borel, F., Kahn, R. & Ferrer, J.-L. (2004). *J. Appl. Cryst.* **37**, 72–77.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pajon, A. *et al.* (2005). *Proteins*, **58**, 278– 84.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.
Pohl, E., Ristau, U., Gehrmann, T., Jahn, D., Robrahn, B., Malthan, D., Dobler, H. & Hermes, C. (2004). *J. Synchrotron Rad.* **11**, 372–377.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.
Singh, I., Stearns, B. & Johnson, M. (2002). *Designing Enterprise Applications with the J2EE Platform*, 2nd ed. Boston: Addison-Wesley.
Snell, G., Cork, C., Nordmeyer, R., Cornell, E., Meigs, G., Yegian, D., Jaklevic, J., Jin, J., Stevens, R. & Earnest, T. (2004). *Structure*, **12**, 537–545.
Stevens, R. & Wilson, I. (2001). *Science*, **293**, 519–520.
Terwilliger, T. C. (2004). *J. Synchrotron Rad.*, **11**, 49–52.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Ueno, G., Hirose, R., Ida, K., Kumasaka, T. & Yamamoto, M. (2004). *J. Appl Cryst.* **37**, 867– 873.
Vernede, X., Lavault, B., Ohana, J., Nurizzo, D., Joly, J., Jacquamet, L., Felisaz, F., Cipriani, F. & Bourgeois, D. (2006). *Acta Cryst.* D**62**, 253–261.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.