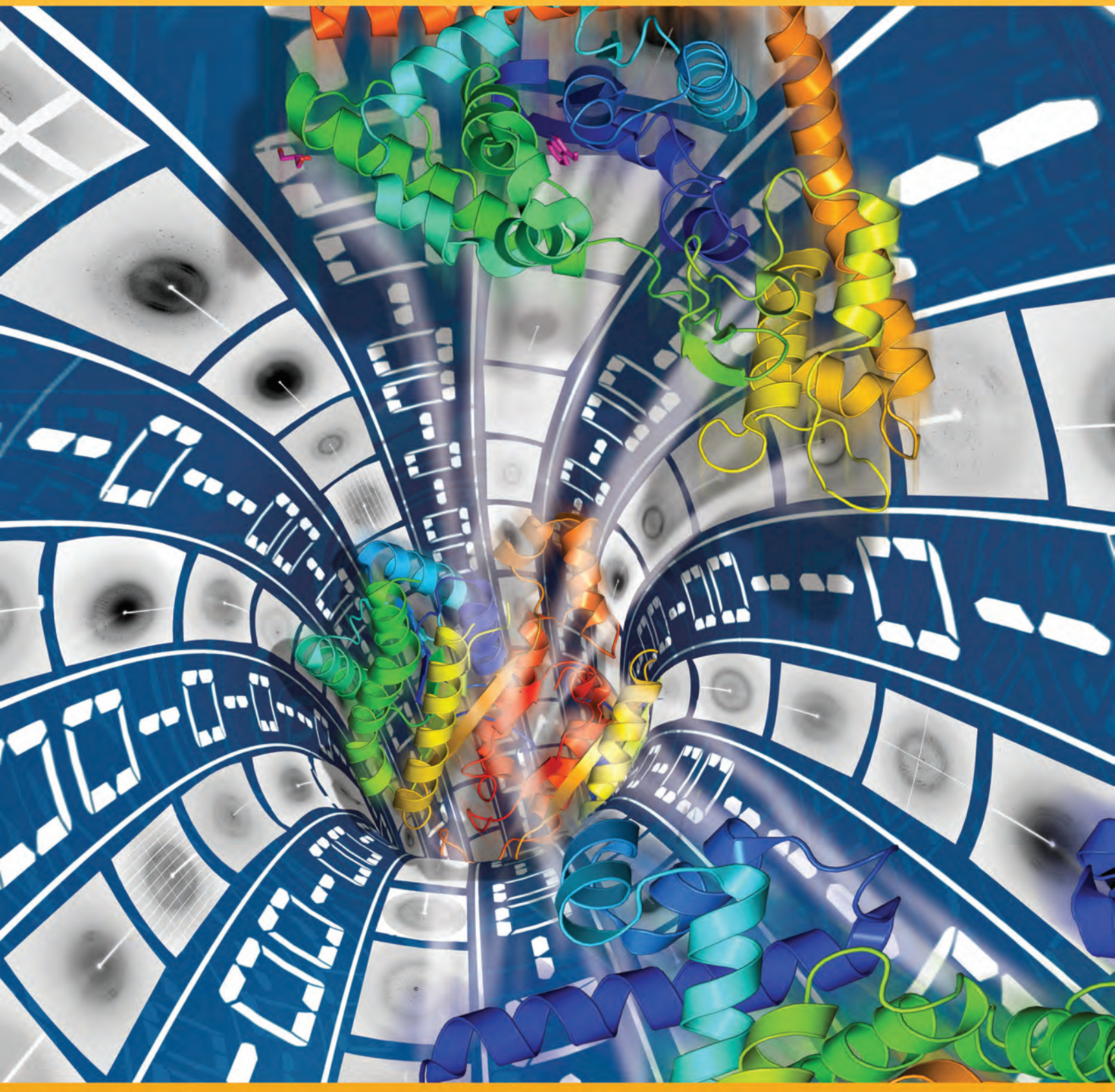
Acta Cryst D

Acta Crystallographica Section D
STRUCTURAL BIOLOGY

ISSN 2059-7983

Volume 72 | Part 11 | 1 November 2016



IUCr Journals | Wiley

STRUCTURAL BIOLOGY

ISSN: 2059-7983
journals.iucr.org/d

A public database of macromolecular diffraction experiments

Marek Grabowski, Karol M. Langner, Marcin Cymborowski, Przemyslaw J. Porebski, Piotr Sroka, Heping Zheng, David R. Cooper, Matthew D. Zimmerman, Marc-André Elsliger, Stephen K. Burley and Wladek Minor

Acta Cryst. (2016). D**72**, 1181–1193

STRUCTURAL BIOLOGY

ISSN 2059-7983

A public database of macromolecular diffraction experiments

Marek Grabowski,[a]‡ Karol M. Langner,[a]‡§ Marcin Cymborowski,[a] Przemyslaw J. Porebski,[a,b] Piotr Sroka,[a] Heping Zheng,[a] David R. Cooper,[a] Matthew D. Zimmerman,[a] Marc-André Elsliger,[c] Stephen K. Burley[d,e] and Wladek Minor[a]*

[a]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22904, USA, [b]Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences, Niezapominajek 8, 30-239 Cracow, Poland, [c]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 90237, USA, [d]RCSB Protein Data Bank; Center for Integrative Proteomics Research; Institute for Quantitative Biomedicine; Rutgers Cancer Institute of New Jersey; Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, and [e]San Diego Supercomputer Center and Skaggs School of Pharmacological Sciences, University of California, San Diego, La Jolla, CA 92093, USA. *Correspondence e-mail: wladek@iwonka.med.virginia.edu

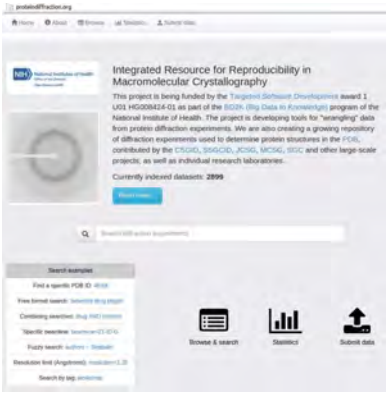
Received 23 June 2016
Accepted 17 September 2016

Edited by T. O. Yeates, University of California, USA

‡ The first two authors contributed equally.
§ Present address: Google Inc., Mountain View, CA 94043, USA.

Keywords: diffraction experiment; protein crystallography; repository; data; metadata; IRRMC.

The low reproducibility of published experimental results in many scientific disciplines has recently garnered negative attention in scientific journals and the general media. Public transparency, including the availability of 'raw' experimental data, will help to address growing concerns regarding scientific integrity. Macromolecular X-ray crystallography has led the way in requiring the public dissemination of atomic coordinates and a wealth of experimental data, making the field one of the most reproducible in the biological sciences. However, there remains no mandate for public disclosure of the original diffraction data. The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) has been developed to archive raw data from diffraction experiments and, equally importantly, to provide related metadata. Currently, the database of our resource contains data from 2920 macromolecular diffraction experiments (5767 data sets), accounting for around 3% of all depositions in the Protein Data Bank (PDB), with their corresponding partially curated metadata. IRRMC utilizes distributed storage implemented using a federated architecture of many independent storage servers, which provides both scalability and sustainability. The resource, which is accessible *via* the web portal at http://www.proteindiffraction.org, can be searched using various criteria. All data are available for unrestricted access and download. The resource serves as a proof of concept and demonstrates the feasibility of archiving raw diffraction data and associated metadata from X-ray crystallographic studies of biological macromolecules. The goal is to expand this resource and include data sets that failed to yield X-ray structures in order to facilitate collaborative efforts that will improve protein structure-determination methods and to ensure the availability of 'orphan' data left behind for various reasons by individual investigators and/or extinct structural genomics projects.



1. Introduction

Issues with the reproducibility of published experimental results have recently attracted attention in many different scientific fields (Collins & Tabak, 2014). The lack of availability of original, primary scientific data represents a major factor contributing to reproducibility problems (Iqbal *et al.*, 2016). The structural biology community (led by protein crystallographers) has already taken significant steps towards making experimental data available. Currently, the main archive(s) of macromolecular structures, the PDB (Protein

Data Bank, 1971) and related projects (Joosten *et al.*, 2009, 2012; Touw *et al.*, 2015), contain not only the atomic coordinates for each published macromolecular structure, but also key intermediate data, including structure-factor amplitudes and a description of the diffraction experiment. In the 1990s, the PDB began requesting the submission of additional metadata, which were included in each PDB entry header, and since 2006 the deposition of structure-factor files for X-ray diffraction models has been mandatory (Berman *et al.*, 2006). The requirement to deposit this intermediate data is now regarded as one of the most important advances in archiving structural information and ensuring robust validation and reproducibility of the method and results (Terwilliger & Bricogne, 2014). In some cases, access to structure-factor amplitudes has enabled structures to be reinterpreted (Shabalin *et al.*, 2015; Raaijmakers & Romão, 2006; Choi *et al.*, 2008).

Data collected in a typical single-crystal experiment include a series of monochromatic two-dimensional diffraction images recorded while the crystal is rotated around a spindle. These images are then indexed, integrated, scaled and merged into a single file in a process that is usually called 'data reduction'. Each recorded reflection is assigned Miller indices (h, k and l), and its intensity $I(hkl)$ and estimated error $\sigma[I(hkl)]$ are recorded. During scaling, equivalent reflections, including those related by space-group symmetry, are scaled and merged together. The structure-factor amplitudes $[|F(hkl)|]$ are derived from these merged intensities. This process of reduction greatly decreases the size of the data set, at the (unfortunate) expense of losing potentially useful data in the merging process, such as diffuse scattering (Van Benschoten *et al.*, 2016). Owing to various restrictions, structure-factor amplitudes (*i.e.* reduced experimental X-ray diffraction data) are often the only data preserved.

Currently, only the coordinates and structure-factor amplitudes (SF) are required for the deposition of a structure in the PDB. Our analyses have shown that SF are not always produced in an optimal way. In many cases, even automatic reprocessing of raw diffraction images could lead to a set of SF with better resolution. Surprisingly, a substantial number of deposited structures report an $\langle I/\sigma(I)\rangle$ greater than 10 for the highest resolution data, far exceeding the common rule-of-thumb threshold of an $\langle I/\sigma(I)\rangle$ of about 2. Moreover, this rule of thumb may be too restrictive given recent advances in refinement algorithms (Diederichs & Karplus, 2013).

The availability of raw diffraction images for subsequent reprocessing could lead to the extraction of additional usable data, enable some previously deposited structures to be improved and/or permit a better interpretation of biomedical results (Shabalin *et al.*, 2015; Sato *et al.*, 2006; Ramachandraiah *et al.*, 2002). In general, raw diffraction images of macromolecular crystals are discarded owing to disk-space limitations or are lost owing to the obsolescence of digital storage media. As a consequence, the original experimental data are irretrievably lost. In the rare cases where the original data images are available they are usually distributed 'as is', and substantial expertise is needed to deduce a wide array of essential experimental information, including X-ray beam geometry, measurement protocols, X-ray detector and goniometer type and movement *etc*. All such metadata are required for successful re-processing.

Amassing a large set of diffraction experiments provides an opportunity for carrying out 'data-mining' studies to explore the impact of various parameters on structure-determination/refinement processes. Machine-learning tools may uncover hidden patterns and correlations, thereby providing a rational basis for recommending 'best practices' for future experimental design, such as data-collection strategies and protocols. Such analyses are also likely to yield improvements in automatic processing and scaling procedures with which to generate optimal sets of structure factors and refined structures.

The importance of retaining raw diffraction data has been emphasized numerous times (Jones *et al.*, 1996; Androulakis *et al.*, 2008; Baker *et al.*, 2008; Jovine *et al.*, 2008; Rupp, 2012; Domagalski *et al.*, 2014; Minor *et al.*, 2016). The International Union of Crystallography (IUCr) responded by forming the IUCr Diffraction Data Deposition Working Group (DDDWG) in 2011. The 2011–2014 DDDWG triennial report (http://forums.iucr.org/viewtopic.php?f=21&t=343) made several key recommendations regarding the preservation of raw diffraction images. These recommendations include (i) the creation of a new type of article in the *Journal of Applied Crystallography* for difficult experiments where the authors 'would describe the nature of the data set and in effect invite the community at large to work with these data'; (ii) the development of 'specifications for a centralized crystallographic repository of metadata describing and locating experimental data sets'; and (iii) that 'authors **should** provide a permanent and prominent link from an article to the raw data sets' used to produce a peer-reviewed publication. Placing the responsibility for a 'permanent' link with a single laboratory is, however, problematic for a variety of reasons, not limited to the lifetime of the originating research group. In contrast, larger resources, such as that described herein, assign DOIs (Digital Object Identifiers; International DOI Foundation, 2016), which should provide a reliable mechanism of locating the data, even if the URL or the maintainer of the data changes.

The potential benefits of archiving raw diffraction-image data (Terwilliger & Bricogne, 2014; Terwilliger, 2014; Meyer *et al.*, 2014; Kroon-Batenburg & Helliwell, 2014; Guss & McMahon, 2014) include the opportunity to improve existing PDB depositions, the provision of training and test sets for methods development, and the prevention of loss of data upon the closure of laboratories and collaborative programs.

We would argue that preserving raw diffraction data is beneficial regardless of whether or not a structure has been determined. Having the experimental data allows *de novo* redetermination of macromolecular structures for validation purposes, and may permit some future structure determination in cases where current methodologies have failed. Of particular importance is archiving high-quality data that may currently be 'unsolvable' owing to the lack of an appropriate

molecular-replacement model, a situation that will likely change as more structures are determined and alignment methods improve. Numerous studies have established that, on average, the quality of crystallographic structures in the PDB (as measured by a variety of validation metrics) is generally good and has steadily been increasing over time (Brown & Ramaswamy, 2007; Read & Kleywegt, 2009; Bagaria *et al.*, 2013; Domagalski *et al.*, 2014). This trend reflects advances in techniques of phasing, structure determination, refinement and validation. New validation tools (Kleywegt *et al.*, 2004; Read *et al.*, 2011; Westbrook *et al.*, 2003) provided by the wwPDB (Berman *et al.*, 2003) website (http://wwpdb-validation.wwpdb.org) generate standardized validation reports for all structures with available structure-factor data and provide statistical measures of structures, assessing how well structures match the corresponding electron density. Following the weekly release of new structures by the PDB, the *PDB_REDO* project automatically re-refines structures with deposited SF using current, state-of-the-art refinement algorithms, and in the vast majority of cases produces improvements in geometric validation criteria, as well as in R and $R_{\rm free}$ (Joosten *et al.*, 2009). However, at present, automatic re-refinement is likely to fail when the deposited atomic coordinates have significant errors. Moreover, automatic re-refinement does not correct situations in which ligands are misidentified and/or misplaced (Cooper *et al.*, 2011; Grabowski *et al.*, 2009; Zheng, Hou *et al.*, 2014). Recent analyses have shown that a significant number (around 15%) of metal ions reported in the PDB are misassigned and/or incorrectly modeled (Zheng, Chordia *et al.*, 2014). Improperly defined ligands may lead to wasted efforts in both academic and commercial research. Re-refinement cannot extend back to the diffraction experiment, but must rely on the deposited structure factors; for this reason, full reprocessing of the diffraction data may be necessary to successfully reinterpret the macromolecular structure (Shabalin *et al.*, 2015; Ramachandraiah *et al.*, 2002; Sato *et al.*, 2006).

In addition, collections of diffraction data sets will serve as training sets for the development of new algorithms and hardware. All instruments commonly used in macromolecular X-ray diffraction studies require calibration. For example, two-dimensional CCD-based detectors produce images that require corrections for spatial distortions and local non-uniformity in response to incoming X-rays. Both calibrated and uncalibrated diffraction images for a given detector are of value in the development of calibration procedures and new detectors. In addition, an archive of systematically processed and organized images is a valuable resource for the development of new algorithms and software for diffraction-image processing. For example, the traditional practice of 'cutting' diffraction data at the resolution where $\langle I/\sigma(I)\rangle$ in the highest resolution shell falls below 2.0 may be too conservative, and may result in the loss of usable high-resolution reflections. Two measures that have been proposed to better identify the optimal resolution limit for a given set of diffraction images are the CC* and $CC_{1/2}$ statistics (Diederichs & Karplus, 2013; Karplus & Diederichs, 2012), although these statistics were derived from limited data sets (Luo *et al.*, 2014). The public availability of a large collection of raw diffraction images constitutes a resource to facilitate the development of improved statistical methods for data analysis.

Last but not least, access to raw diffraction data can further improve the quality of macromolecular X-ray crystallography by facilitating the detection of errors and the identification of (potential) fraud. There have been cases where significant errors were discovered in the published crystal structures of macromolecules (Chang *et al.*, 2006; Matthews, 2007; Zaborsky *et al.*, 2012; Rupp, 2012). Some of these cases have resulted in notable retractions. In addition, examples of questionable interpretations of key components of crystal structures, such as ligands modelled into weak or non-existing electron density that cannot justify their presence, have been identified (Weichenberger *et al.*, 2013). The need for model verification has prompted the development of numerous advances in macromolecular structure validation, such as the now obligatory $R_{\rm free}$ factor (Brünger, 1992) and more recently the RSRZ test (Kleywegt *et al.*, 2004) in the new wwPDB validation tools, *MolProbity* (Davis *et al.*, 2007; Chen *et al.*, 2010) and other statistical quality indicators (Tickle, 2012). Nevertheless, a number of problems cannot be rectified from the atomic coordinates and SF data alone, such as errors in space-group assignment and an inappropriate choice of resolution-limit cutoff.

Several bottom-up initiatives by research groups and institutions have already started gathering diffraction images. Synchrotron-radiation sources, where most diffraction images are collected, are natural places to build repositories of raw data. One of the pioneers has been the Store.Synchrotron project at the Australian Synchrotron (https://store.synchrotron.org.au; Meyer *et al.*, 2014), which reports over 150 000 archived data sets and thousands of experiments, although only very few have been made publicly available (reportedly 35 as of May 2016). Other synchrotrons and/or individual beamlines have also set up archives of macromolecular X-ray diffraction data (with limited or no public availability), *e.g.* the MX archive at beamline 8.3.1 of the Advanced Light Source (ALS; Holton, 2012). Store.Synchrotron has been using an open-source, web-based image and metadata management system called MyTARDIS (http://mytardis.github.io/; Androulakis *et al.*, 2008). Kroon-Batenburg & Helliwell (2014) have published a collection of raw data sets online (http://rawdata.chem.uu.nl/) corresponding to structures reported previously (Tanley *et al.*, 2013). These initial explorations of raw image archiving, allowing scientists to experiment with data from other researchers, have been valuable; however, their impact has been limited owing to a paucity of data.

Two recently launched projects are building large-scale collections of publicly available diffraction experiments (raw diffraction data). One is the project described in this manuscript, which contains data for almost 3000 diffraction experiments. The other is the Structural Biology Data Grid (http://data.sbgrid.org; Meyer *et al.*, 2016), which has made available data from 193 structures determined by 59 affiliated

laboratories (as of June 2016). In parallel, the European Synchrotron Radiation Facility (ESRF) is introducing policies of long-term (at least five years) storage of diffraction data collected by users funded through public agencies (ESRF, 2016). Such data will be made publicly available after an initial embargo, which should result in tens of thousands of diffraction experiments becoming publicly available within the next few years.

Our own efforts in this area aim to encourage community interest by providing a useful resource that contains a significant amount of data from the outset. We are fortunate in having access to large stores of diffraction-image data for protein structures published by individual laboratories, the Center for Structural Genomics of Infectious Diseases (CSGID), the Seattle Structural Genomics Center for Infectious Disease (SSGCID), the Structural Genomics Consortium (SGC) and three extinct SG centers: the Midwest Center for Structural Genomics (MCSG), the Joint Center for Structural Genomics (JCSG) and the New York Structural Genomics Research Consortium (NYSGRC). Currently, our resource contains 2920 diffraction experiments (corresponding to 5767 diffraction data sets and 1.2 million diffraction images), all of which are publicly available for download. Data sets corresponding to these images may be browsed and queried based on general terms and crystallographic metadata. It is our hope that the crystallographic community will see the benefits of depositing diffraction data into our resource, not only to preserve their hard work and relinquish them of the burden of maintaining their own archives, but also to provide a more solid foundation for the future of crystallography.

2. Design and implementation

2.1. Overall architecture

In the current implementation of the IRRMC, metadata for all diffraction sets are stored in a central metadata server, in a searchable relational database, while the diffraction images themselves are stored separately on multiple file servers. Splitting the archive of images across distinct storage servers provides extra flexibility and scalability by allowing the distribution of many terabytes of data in potentially disparate physical locations, while simultaneously providing redundancy. To accommodate different storage locations and file-transfer protocols, which may be associated with heterogeneous distributed storage networks, the central database stores a Universal Resource Identifier (URI) of the diffraction experiment that identifies both the location of the image set and the protocol used to retrieve it. This resource-agnostic means of identifying an image set *via* a URI, which can be easily updated in a database, allows persistent storage of experimental data at multiple physical locations. Moreover, the central metadata database can easily be managed and migrated to a different server without affecting the data storage itself, thereby avoiding the migration of large amounts of data. Our distributed approach is also designed to ensure that the IRRMC is an easily maintainable and sustainable repository.

2.2. Harvesting the metadata

A set of diffraction images without the associated metadata describing how they were measured and what they represent is **not** useful. Diffraction images usually contain some metadata within the header of the image, which typically are limited to data-collection parameters, such as the geometry of the diffraction experiment, the equipment used for measurement, when and where it was collected *etc.* These may be sufficient to process the images and obtain an experimental electron-density map, but errors in the image header are not uncommon (Meyer *et al.*, 2016). Moreover, a description of the experimental sample, *i.e.* the crystal, is usually not contained in the header, thus an unambiguous identification of its content (macromolecule, ligand, buffer *etc.*) is not possible from the header information alone. For structural genomics (SG) targets, much of this 'upstream' metadata can be harvested from centralized databases such as TargetTrack (Gabanyi *et al.*, 2011) and/or the local databases of individual SG projects. However, in most other cases one must rely on information provided by the depositors to the PDB during the deposition process. Much of this information is not standardized and/or mandated, and frequently one must consult the laboratory notebooks/memory of colleagues that were involved in a particular diffraction experiment in order to correctly reproduce the original results. In addition, all available 'downstream' information represents very informative metadata, and should also be included.

Table 1 lists the major categories of metadata associated with diffraction images that we attempt to harvest within the IRRMC. The extraction and curation of data are the first steps of the annotation process for newly deposited diffraction images. Custom-built tools extract, gather and perform some basic checks on the metadata, and organize the raw diffraction images into a standardized directory structure. Currently, metadata are obtained from four sources: the user, the headers of the image files, processed structure factors/scaling logs and molecular models (initial model or final structure). Auxiliary metadata pertinent to the diffraction data set are also gathered from external databases and included as needed.

Implementing metadata harvesting turned out to be the most difficult part of building the IRRMC system, as is often the case in projects involving large amounts of data. Our initial process gathered the metadata associated with diffraction experiments contributed by structural genomics centers from their internal databases. Subsequently, we built custom scripts to automatically extract metadata from the image files themselves, making use of the *HKL*/*HKL*-2000 (Otwinowski & Minor, 1997) and *HKL*-3000 (Minor *et al.*, 2006) program suites, which contain a number of algorithms with extensive heuristics for the accurate identification and processing of >250 different detector image formats. In some cases, new functionality was developed within the *HKL*-2000/*HKL*-3000 suite to permit the automatic execution of some program

Table 1
Metadata sources and basic parameters harvested by the IRRMC.

Metadata source	Metadata parameters
User	Identity and affiliation of the user
	Identities of the people who collected the data
	Location of data collection (beamline, home source *etc.*)
	Date of collection
	Identity of the protein (*e.g.* GenBank, UniProt identifiers)
	PDB identifier of solved structure (if deposited)
	Custom labels
Diffraction images	Detector type and serial numbers (S/N) and image format
	Goniostat type
	Data-collection parameters: number of frames, oscillation-step size, goniostat orientation angles, 2θ offset, detector distance
Structure factors/scaling logs	Integrated reflection data
	Nominal resolution cutoff
	Completeness, overall and highest resolution shell (HRS)
	Redundancy, overall and HRS
	Mean $I/\sigma(I)$, overall and HRS
	Software used to process diffraction images
	$R_{\rm merge}$, $R_{\rm meas}$, $R_{\rm p.i.m.}$
Automatic reprocessing	Validation of provided/extracted metadata
	Validation of space group
	Validation of merging statistics from deposited structure factors/scaling logs
	Estimation of radiation damage
	Estimation of crystal internal non-isomorphicity
	Presence/strength of anomalous signal
	Diffraction-image artifacts and other 'features' (background scattering, ice rings, diffuse scattering *etc.*)
Molecular models	Structure-determination methods (SAD/MAD/MR *etc.*)
	Programs used to determine the structure
	Structure-refinement methods
	Programs used to refine the structure
	$R/R_{\rm free}$
	Electron-density maps (calculated or extracted from the Uppsala Electron Density Server)
SG databases/LIMS	Sample-preparation data
	Target justification and selection criteria
	Crystallization conditions
External databases	PDB data items
	Protein data items
	PubMed data items

components. Our extraction system performs basic 'sanity' checks of data from different sources to ensure internal consistency. For example, the detector type identified in the file headers is checked against the collection location specified by the user. If the beamline is known to have used a different detector on the date of collection (as is ascertainable from the BioSync service; http://biosync.sbkb.org/) to the type identified in the headers, this inconsistency is signaled. More advanced techniques based on methodology developed for the LabDB LIMS (Zimmerman *et al.*, 2014) are used to flag other metadata inconsistencies. For the diffraction experiments contributed by SG projects, metadata and annotation extraction has been successful in about 95% of cases. Cases where this process failed were a result of missing or corrupted files, incomplete or inconsistent information in the headers or apparent contradictions between the data in the headers and the data in the PDB (*e.g.* 'mistaken identity' cases). A number of incorrectly labelled diffraction experiments have been manually reassigned to the correct deposition.

Various metadata models and levels of granularity can be used to organize the diffraction metadata. At the top level, diffraction images are usually grouped into an ensemble comprising data collected for a particular project or 'diffraction experiment' (*e.g.* representing a particular PDB deposition). This ensemble may contain images collected at different wavelengths (*e.g.* MAD or SAD experiments), but also on different beamlines or multiple physical samples, *i.e.* multiple crystals *etc.* At the most detailed level, metadata can be extracted from each individual diffraction image. In virtually all cases, however, individual images can be grouped into 'diffraction data sets' (framesets), which are series of images collected during a single goniometer rotation for a single sample using a particular equipment configuration. Typically, filenames for images in such framesets are numbered consecutively within a given frameset. In our current data model, these 'diffraction data sets' constitute the most basic entity. It is usually not possible to ascertain the role of individual framesets in the process of structure determination based on their content alone. This information may be of great benefit for the future reinterpretation of structures, so we plan to extend the deposition dialog to request annotations pertaining to the original processing of raw data (*e.g.* processing logs). The metadata collected for the diffraction sets can be exported in the macromolecular Crystallographic Information File (mmCIF) format (Hall *et al.*, 1991; Westbrook & Bourne, 2000; Bernstein *et al.*, 2016), as defined by the IUCr.

2.3. Data storage

Our goal was to rapidly establish a useful service using the simplest possible infrastructure. Thus, the necessary disk space was provided by expanding the storage capabilities of several existing servers. On each individual server, diffraction data are stored on a RAID array. Backups and redundancy are facilitated by this distributed data-storage system architecture, as more than one URI can be assigned to data sets. Diffraction experiments are stored in two ways: the originally uploaded tar files representing a single 'diffraction experiment' and repacked, standardized tar files that contain unprocessed diffraction images and a directory that contains the results and the log files of data processing and scaling (when present). Diffraction experiments range in size from several hundred megabytes to 80 GB (uncompressed). Experiments are composed of data sets, and may contain one data set or as

many related data sets as necessary for structure determination and/or refinement. Currently, the largest experiment contains 12 data sets for one structure determination, but advances in multi-crystal experiments could dramatically increase this number. Individual data sets range from around 100 MB to 14.9 GB in size, with an average size of 4.3 GB. The total size of the complete uncompressed diffraction data housed at the IRRMC is currently around 25 TB. The total storage required for the original and repacked data sets as well as their backups is ∼100 TB (uncompressed). Compression reduces the storage requirements significantly and speeds up file-transfer times, since even simple gzip compression typically reduces a data set to 25–30% of its original size. However, using a compressed file (*i.e.* processing diffraction data) requires an additional decompression step, which can take much longer than the reduction in the file-transfer time. For this reason, compression may not always be practical.

2.4. User interface

We employ a standard Apache server with a web server gateway interface (WSGI) to host the website (http://proteindiffraction.org; Fig. 1). In our initial design, the server was built using the minimalist web.py library (http://webpy.org) and the Bootstrap display framework (http://getbootstrap.com), which are readily available tools that allow a web service to be built quickly. In response to a larger than expected number of diffraction experiments and the corresponding increased requirements for metadata search and presentation capabilities, we have developed a more robust implementation that takes advantage of the exceptional scalability, rapid development capabilities and large code repositories of the Django framework (http:/djangoproject.org).

The main user interface (UI) provides tools for searching/browsing diffraction experiments and displays data-set information. The UI also provides meaningful statistics for individual or multiple data sets and provides a convenient mechanism for data download. Any word or parameter in the metadata can be used as a search criterion, which will present the resulting data-set collection together with statistics relevant to the particular query. The UI search tool utilizes a single text box for user queries. A simple keyword query retrieves all records in which any of the textual metadata matches the keyword. Individual fields such as authors, resolution, beamline, experimental method *etc.* may be searched by more complex queries with the operators '=', '<', '~' (fuzzy search) *etc.* These searches may be further linked using the Boolean operators 'AND', 'OR' and 'NOT', and grouped using parentheses (with spaces within keywords acting as an implicit 'AND'). The search results are listed below a small summary containing basic statistics for the resulting data-set collection compared with all data sets. The query language parser has been implemented using the *pyparsing* module (McGuire, 2008). Fig. 2 shows an example of an advanced search to generate a collection of structures with resolution less than 2.0 Å or more than 3.0 Å, determined with methods other than SAD or MAD, and matching both the keywords '*Homo sapiens*' and 'JCSG'.

The search function can also be used to identify promising candidates for manual reprocessing, meaning structures with a high biomedical impact that appear to have been sub-optimally processed. For example, reprocessing of all data

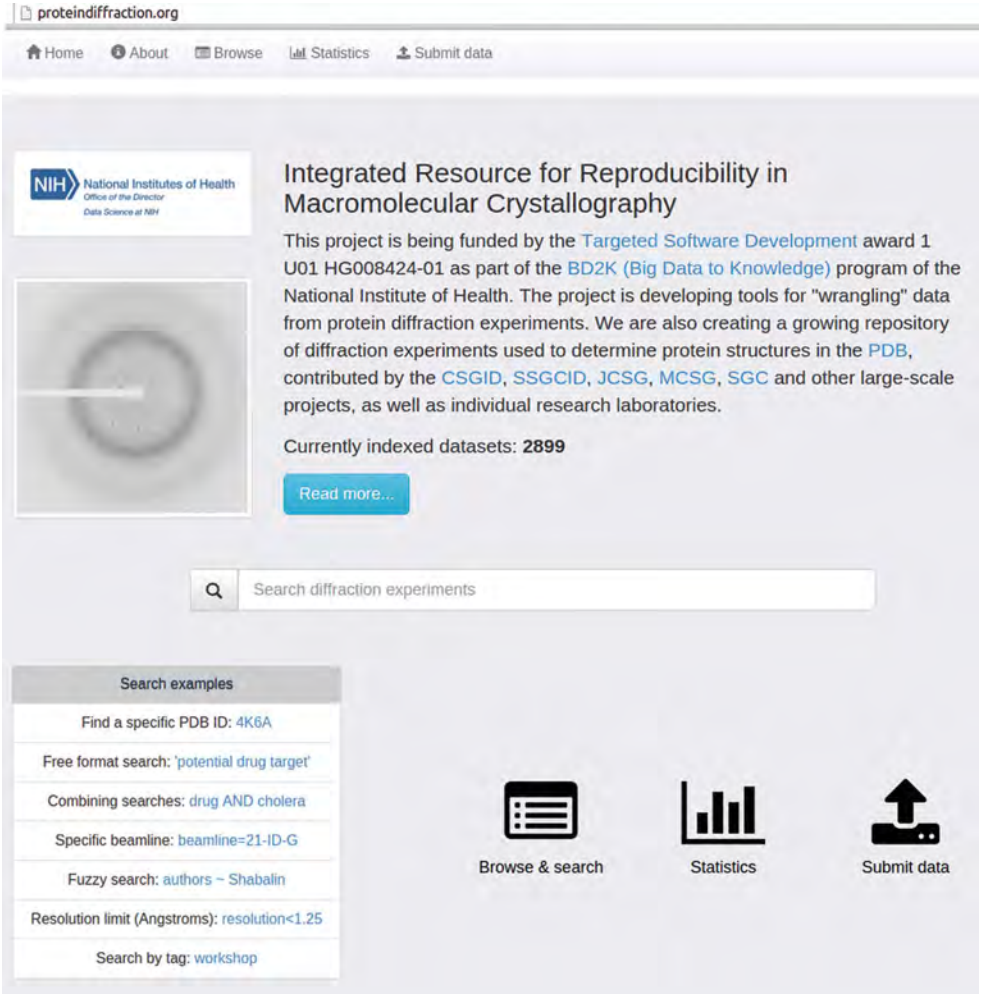


Figure 1
Home page of the http://proteindiffraction.org web portal.

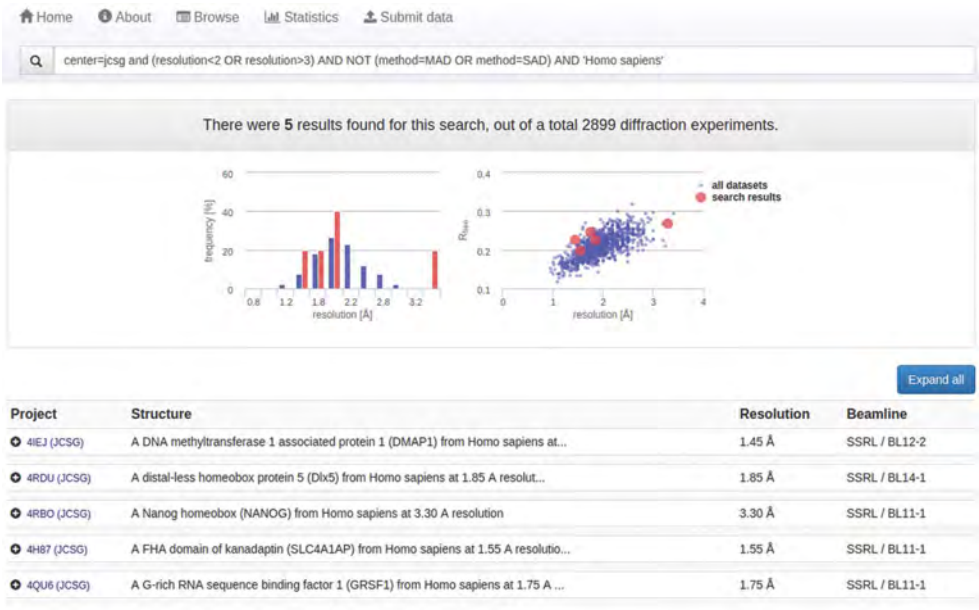


Figure 2
Example of a search of the http://proteindiffraction.org web portal.

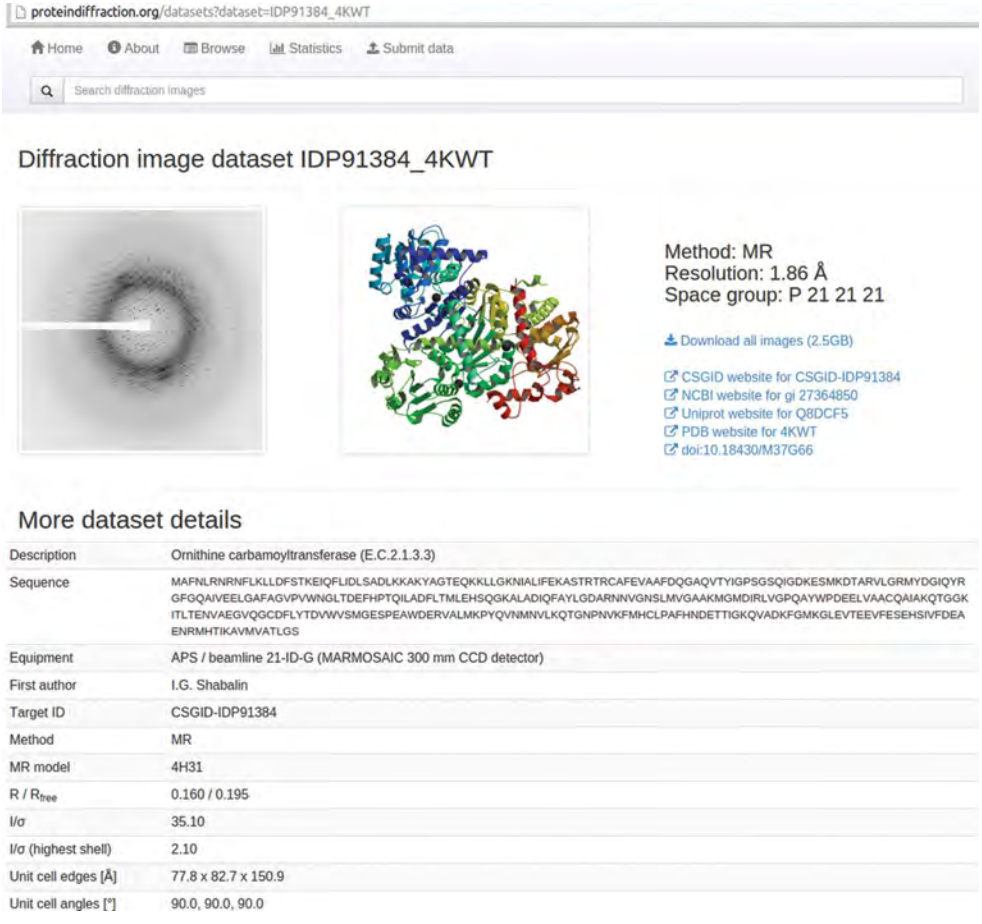


Figure 3
A view of a specific diffraction experiment.

sets with $\langle I/\sigma(I)\rangle$ greater than 10 in the highest resolution shell may significantly improve the useful resolution, which will be especially beneficial for medically relevant protein–ligand complexes.

The search results are presented as rows of identifying data that can be 'expanded' to view thumbnail views of the structure and the first diffraction image, information about an associated structure, connections to external resources and links for downloading the data. Larger thumbnails and more detailed information about an individual data set are available from a link in the expanded view (Fig. 3). We are continually updating these views and developing new functionality, while preserving quick download links to the entire data set and essential crystallographic details, and links to external resources wherever possible. For example, PDB identifiers are linked to the PDB resource, protein accession codes are linked to the UniProt and NCBI databases, target identifiers are linked to SG websites, electron density is linked to the Uppsala Electron Density Server (Kleywegt *et al.*, 2004) *etc.*

Several informative plots are available in the 'Statistics' tab. For example, Fig. 4 shows a scatterplot of the average $\langle I/\sigma(I)\rangle$ against the same ratio in the highest resolution shell. Such plots are especially useful for the rapid assessment of overall dataset quality and outlier identification. For example, in Fig. 4 the highlighted data set used to determine the structure with PDB entry 1vkb is clearly an outlier, with an $\langle I/\sigma(I)\rangle$ value of 27 in the highest resolution shell.

A 'Submit data' tab allows registered users of the IRRMC to upload diffraction data *via* a web form employing the DropzoneJS library. Registration is open to anyone interested in

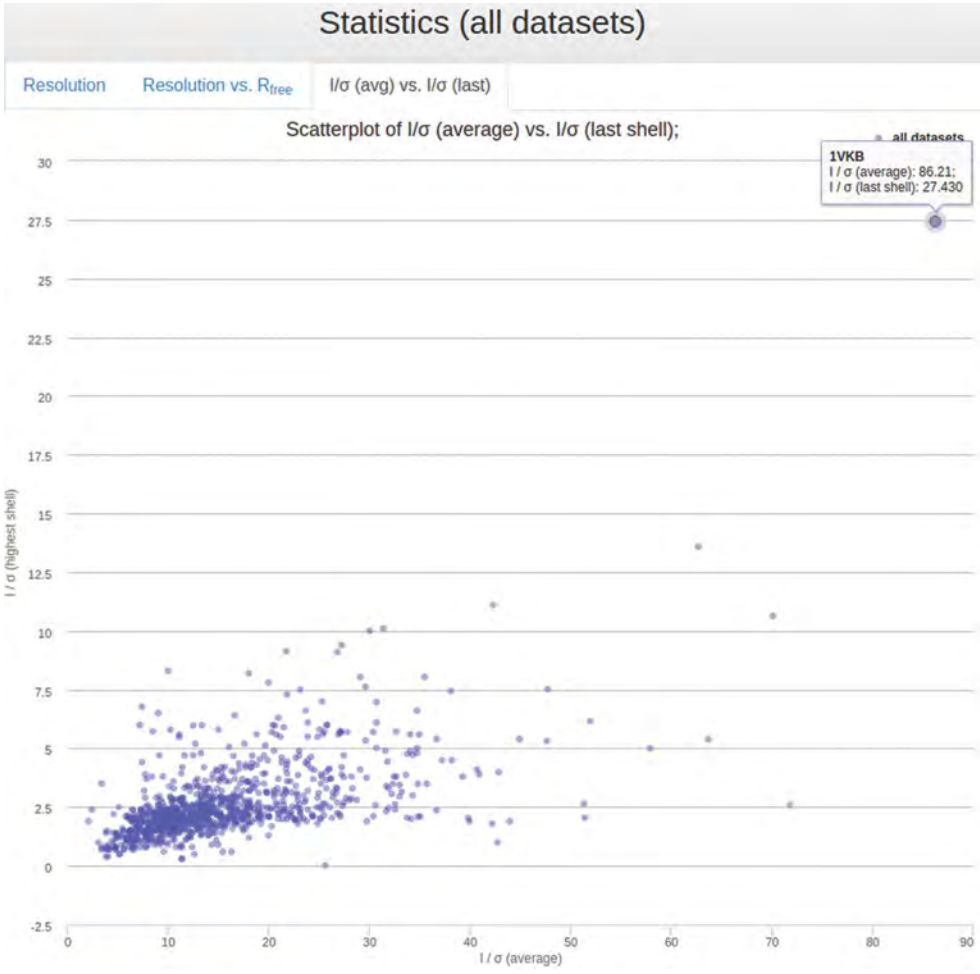


Figure 4
Statistical tools on the IRRMC website include a scatter plot of the average $\langle I\rangle/\langle\sigma(I)\rangle$ *versus* $\langle I\rangle/\langle\sigma(I)\rangle$ in the highest resolution shell.

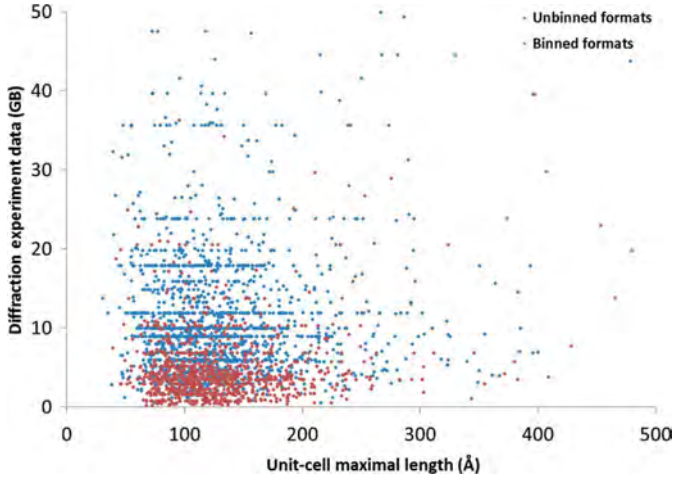


Figure 5
Scatter plot of the total size of uncompressed diffraction experiment data (GB) *versus* unit-cell size (as measured by the longest unit-cell length in Å).

contributing to enlarging the body of publicly available protein diffraction data.

3. Preliminary analysis

The set of diffraction experiments currently available on the server at http://proteindiffraction.org comprises data collected at over 35 different experimental stations, primarily from the Advanced Photon Source (APS), Stanford Synchrotron Radiation Lightsource (SSRL) and Advanced Light Source (ALS), plus ∼200 experiments conducted using 'home source' X-ray systems. Many different detectors and data formats are represented therein.

A full analysis of these experiments is beyond the scope of this paper; however, preliminary analysis shows that the raw data contain more useful information than was initially expected. Even a quick survey of the diffraction experiments deposited in the IRRMC shows that most of the data were collected in binned mode. Generally, the size of binned data sets is between 500 MB and 10 GB, while the sizes of most unbinned data sets are between 2 and 40 GB (Fig. 5). Rationally, one would expect that the data-set size would increase for larger unit cells, which necessitates smaller oscillation steps; however, this trend is not observed for two potential reasons: (i) the most popular oscillation step size of 1° is rarely modified, even for excessively large unit-cell dimensions (Fig. 6), and (ii) it may become unrealistic to collect many frames from the same crystal owing to factors such as radiation damage. Indeed, there is a weak correlation between the size of the unit cell and the experimental oscillation step (Fig. 5), although over 80% of the data were collected with an oscillation-step size of either 0.5 or 1°. For unit-cell edge lengths of <150 Å, there are 150 data sets with an oscillation-step size as large as 1.5° or even 2°. Analysis of the metadata and a limited sample reprocessing suggests that major gains can be achieved through the optimization of data-collection protocols, rather than by re-processing existing data. While reprocessing may improve $R_{\rm merge}$ statistics and can extend the usable resolution limit, it cannot overcome suboptimal experimental design. For example, metadata analysis of the oscillation angle shows that the oscillation step is rarely adjusted to match the mosaicity

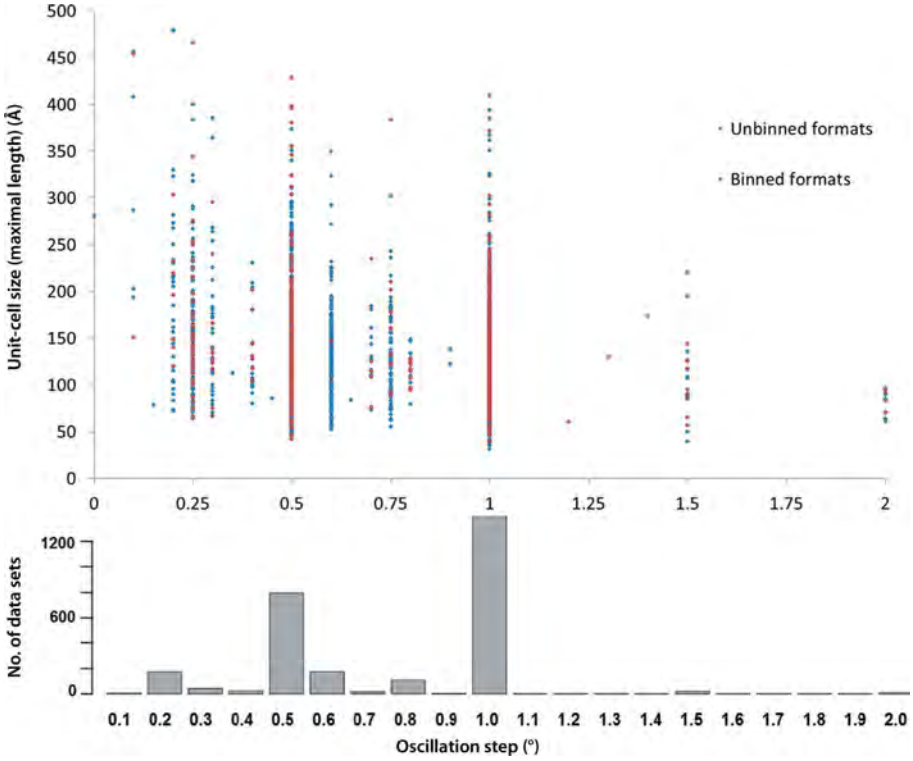


Figure 6
Scatter plot of unit-cell size (as measured by the length of the longest unit-cell edge in Å) *versus* the experimental oscillation step.

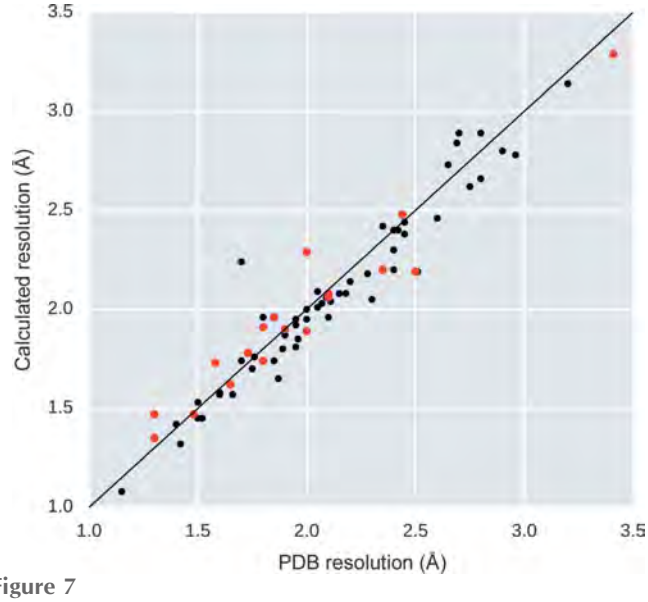


Figure 7
Resolution reported in the PDB *versus* that calculated from reprocessed diffraction images. Resolution was calculated from the reprocessed diffraction images as the resolution in the highest resolution shell that had $\langle I/\sigma(I)\rangle$ higher than 2 as reported by *SCALEPACK*. Each PDB deposition is represented as a dot. Red dots represent diffraction data sets for which the reprocessed $\langle I/\sigma(I)\rangle$ was higher than 3 in the highest resolution shell corresponding to the resolution 'at the edge' of the detector. The differences in the resolutions may be attributed to the adoption of different processing strategies for original processing and reprocessing, including different regions of the diffraction image included in the processing ('to the edge' *versus* 'to the corner' of the detector), inclusion/omission of parts of data sets and differences in integration/scaling programs and processing parameters.

and unit cell, but is rather fixed at the ever-popular value of 1° as described above (Fig. 5). Too large an oscillation step may result in a significant number of overlapping adjacent reflections, yielding incomplete data. Analysis of automatic reprocessing shows that sometimes the resolution of the data was limited by the sample-to-detector distance rather than by the diffraction power of the crystal. On the other hand, the data resolution estimated by automatic reprocessing was, for most data sets, only marginally better than that reported in the PDB (Fig. 7).

Automatic reprocessing of some data sets was quite straightforward, provided that the correct X-ray beam position was present in the diffraction-image header. Unfortunately, analysis of data from two different synchrotron stations revealed insufficient metadata to support automatic processing. In one case, a fixed value was used for the beam center and automatic processing was only possible with recourse to historical beam positions extracted from the *HKL* site files stored in our laboratory databases (Fig. 8). This pitfall illustrates not only the importance of metadata extraction and preservation, but also that the reliability of a local database may be greater than that of the databases used by some synchrotron stations. Anecdotal experience from some authors suggests that a blackboard/whiteboard/Post-It note is sometimes the most accurate (albeit insecure) source of metadata available at some experimental facilities.

4. Discussion and conclusions

Our experience with the IRRMC resource demonstrates that with relatively modest resources it is possible to build a large, searchable, web-accessible archive of protein crystallography diffraction images organized according to the metadata.

In a recent paper, Guss & McMahon (2014) formulated eight essential attributes of an 'Image Archive', representing an ideal entity managing the storage of protein diffraction images for the crystallographic community. These are the following.

(i) Long-term availability of data.

(ii) Persistent identifiers are assigned for all data sets.

(iii) The status of data sets is trackable *via* persistent identifiers.

(iv) All data sets are accessible *via* persistent identifiers.

(v) Restricted-access data sets are discoverable *via* persistent IDs.

(vi) Bidirectional links exist between the data sets and the scientific publications that use them.

(vii) The archive is searchable by a wide variety of criteria.
(viii) Data sets are validated.

These attributes informed the development of our IRRMC resource (http://proteindiffraction.org). We have been assigning Digital Object Identifiers (DOIs) registered and maintained by the International DOI Foundation (IDF) as persistent identifiers. During the creation of a DOI for a diffraction experiment, basic metadata are collected, including the title and authors of the data set and the URL where the data set can be accessed. Through a combination of network resolution mechanisms and 'social infrastructure' (responsibilities of registering institutions) DOIs are 'network actionable' (*i.e.* can be located through the World Wide Web) and persistent. Once assigned, DOIs cannot be destroyed and their basic metadata remains accessible through the IDF 'resolvers', but maintenance of the URLs providing the status of the data as well as ensuring the availability of actual data remains the responsibility of the archiving resource. By maintaining up-to-date links within IRRMC, and by keeping the DOI metadata updated, our resource addresses requirements (ii)–(iv). While at present there are no instances of data sets with restricted access, our method of DOI generation would satisfy requirement (v) as well. As for the establishment of bidirectional links between data and publications (vi), the IRRMC provides links to the PDB and associated publications for diffraction experiments that were used in PDB depositions. Future publications using diffraction images collected by the IRRMC would be able to link to the original data using DOIs (which are also linked from the PDB web portal). The current IRRMC implementation also satisfies the 'searchability' requirement (vii). The validation requirement (viii) is currently implemented in our resource through a check of image headers, as described in §2.2. Even rudimentary validation flagged several dozen cases of 'mistaken identity', wherein the diffraction images clearly could not have come from the experiment described by the depositor. Resolution of these problems was performed by direct communication whenever possible. Further developed automated processing will eventually provide more complete data verification.

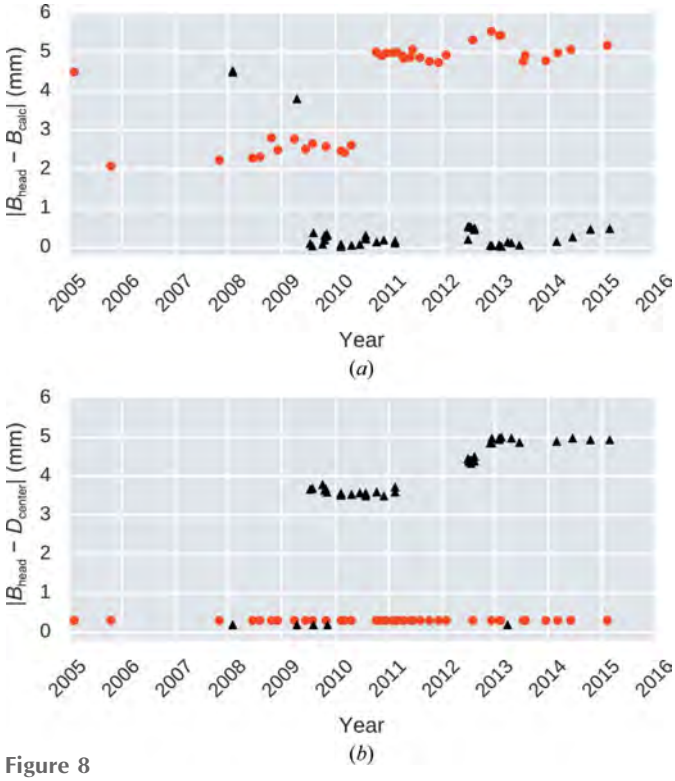


Figure 8
The beam position extracted from the image headers is not always meaningful. Each panel shows the data from two beamlines in red and black. (*a*) shows the distance between the beam position extracted from the header (B_{head}) and that calculated during automated data integration (B_{calc}). (*b*) shows the distance between the beam position extracted from the header (B_{head}) and the detector center. While the beam position reported in the headers from the black beamline fluctuates and changes over time, the beam position for the red beamline is fixed. The lack of a meaningful beam position led to problems with automatic reprocessing of the images owing to a huge discrepancy between the reported and the actual beam positions. It was possible to process data from the red beamline by using historic beam positions from *HKL* site files stored in our laboratory. This is a vivid example of how the extraction, validation and preservation of diffraction-experiment metadata is necessary.

Of all the postulated attributes of the ideal 'Image Server', long-term availability (i) is the most difficult to guarantee, owing to the inherently transitory nature of grant funding. Covering the costs of maintaining a repository of protein diffraction images, and especially of keeping it abreast of the rate at which new data are being collected, will be nontrivial. The establishment of a large public repository of diffraction images has traditionally been considered to face two major challenges: (i) the prohibitive costs of storing and transferring an extremely large repository and (ii) the difficulty in extracting and managing the semantic 'metadata' required for effective use of the raw images, not only by experts but also by the broader community of biomedical researchers. One could also debate the vexing question of return on investment, given that the PDB already requires the deposition of both atomic coordinates and SF data. What would it cost to detect and resolve one incorrect deposition to the PDB archive and how much is this really worth to the scientific community *versus* funding other endeavors aimed at promoting reproducibility elsewhere in the biomedical sciences?

Raw diffraction-image data sets are several orders of magnitude larger than the reduced sets of structure factors. Depending on the particular detector and the number of images collected, a complete set of diffraction images typically ranges from hundreds of megabytes to hundreds of gigabytes. Despite the continuous growth of data-storage technology, the fraction of publicly available data sets is quite low. This is mainly because the organization of the 'Augean stables' (Apollodorus, *Bibliotheca*, ∼200 BC, as cited in Frazer, 1921) filled with hundreds of thousands of diffraction experiments remains a Herculean task. To complicate matters, commercial detectors use a plethora of different image formats (the *HKL*/*HKL*-2000 suite currently recognizes more than 250), which are very poorly standardized and change over time. Recurrent efforts to encourage detector vendors to support standardized file formats or even image headers have not succeeded in the last 30 years, and in our opinion there is little chance that they will do so in the future. Therefore, relying on the metadata

reported in the 'header' of the diffraction image formats is not an option, not to mention that the experimental setup in the header may be improperly recorded or not recorded at all. The only available option is then to build an information system that can provides seamless access to a desired subset of all diffraction images and metadata.

The costs of storing diffraction images in a repository are declining steadily year on year. Currently, our estimates suggest that hardware-related expenditures amount to around $2 per data set to set up and maintain a dedicated storage server. Hosting on cloud-based services is comparable in cost over the short term; for example, on Amazon.com's S3 storage service it would cost about $1.80 per year to host a typical 5 GB data set. In addition, file compression using gzip or bz2 algorithms can further reduce hardware-related costs. Westbrook (2012) argued that the bulk of the costs of establishing and maintaining diffraction-image storage repositories are the human resources necessary for harvesting and curating raw diffraction data. The experience of the IRRMC confirms that most of the effort is not related to setting up hardware and software infrastructure for physical storage, but rather to the harvesting, validating, annotating and indexing of data and the associated metadata. To move forward responsibly, a full accounting of the costs required to establish and manage diffraction-image repositories will have to be balanced against the benefits to the structural biology community and the opportunity costs to the biomedical enterprise. The experience of IRRMC and other resources that have appeared in recent years (Meyer *et al.*, 2016) indicates that preserving diffraction data on a large scale is feasible. Additionally, efforts to curate the data provide a context to the data, regardless of the eventual physical storage location of the data.

An important application of our resource would be preventing the loss of data from extinct large-scale projects and individual investigator laboratories that are about to be closed. This provision would be particularly timely in light of the termination of the Protein Structure Initiative and the contraction or conclusion of other large-scale, federally funded projects in the US. As these projects come to a halt, large amounts of valuable crystallographic data may well be abandoned or lost. Other data sets, each of which costs significant time and money to produce *via* a lengthy experimental pipeline, may linger unsolved on storage media owing to a loss of funding or limited manpower. Analyzing the TargetTrack repository of worldwide structural genomics targets (Gabanyi *et al.*, 2011), we found almost 900 targets for which diffraction data had been recorded but which had not progressed to the deposition of a crystal structure in the PDB. Of these, ∼60% were targets of the NIH-sponsored programs; subsequently, ∼5% of them were solved as a different target by other SG programs. Another ∼5% were solved independently by researchers from outside the SG community, but the majority of structure determinations in progress remain uncompleted. (Note: a target was considered to be 'solved' if the structure of the protein or of a homolog with greater than 98% sequence identity was subsequently deposited in the PDB.)

Preserving diffraction data may have additional benefits in the crystal physics arena *via* the analysis of diffuse scattering. When X-ray diffraction spots for macromolecular crystals do not form well shaped peaks but rather are 'smeared' or display other artefacts, this is often considered to be a nuisance that needs to be overcome. All information recorded between the diffraction spots is lost when the spots are integrated; generally, the parameters controlling the area integrated for each peak are narrowly specified, and all of the background diffraction data are ignored. However, these diffuse scattering effects may contain useful information about the properties of the macromolecular crystal (Glover *et al.*, 1991; Jovine *et al.*, 2008). Both static and dynamic disorder in crystals contributes to diffuse scattering. Moreover, analysis of these data could provide insights into diffraction resolution limitations and crystal anisotropy. Analysis and confirmation of crystal twinning and radiation damage will also benefit from the presence of the original data. After all, the preparation of diffraction-quality crystals remains the rate-limiting step in many structure-determination campaigns.

Preserving diffraction data may provide valuable input for synchrotron beamlines around the world for the analysis of performance and improvement of throughput. The importance of optimizing experimental protocols can be illustrated by analyzing the productivity of synchrotron beamlines, as measured by the number of PDB depositions during the last three years (Zheng, Hou *et al.*, 2014). It is surprising that even the best-performing beamlines in the world still average less than one deposited structure per day (yearly total/365). Annual beamline productivity metrics vary widely, ranging from one to 337 PDB depositions (as of 2014), with an average of 74. While a calculation of averages is not without flaws (*e.g.* synchrotron sources typically operate for only ∼200 days per year), only 1–5 min of data-collection time are usually needed to accumulate data sufficient for structure determination (Walsh *et al.*, 1999; Joachimiak, 2009). Detailed inspection reveals that there is no detectable correlation between beamline productivity and any aspect of the physical setup of the data-collection hardware. Assuming that the average sample quality is similar for projects at different beamlines, the 'productivity gap' of two orders of magnitude between the highest and lowest performing beamlines can be most likely attributed to variations in diffraction data-collection protocols (Zheng, Hou *et al.*, 2014), differences in the integration of software and hardware, and the proclivity of 'successful' beamlines to act upon feedback of visiting experimenters. Long-term preservation of diffraction data and associated metadata would provide benchmarks with which to determine best practices and thereby increase synchrotron beamline productivity worldwide.

The IRRMC is not the only ongoing effort to collect diffraction images (ESRF, 2016; Meyer *et al.*, 2016) with the intention of preserving and making the data publicly available for generations to come. This initial period of development should not be viewed as a competition between rival systems, but rather as a fertile testing ground from which innovation, collaboration and new functionality can arise. While it is

possible that several disparate, incompatible systems will emerge, or that one of the systems currently under development will so surpass all the others to become the standard, it is more likely that an international meeting will lead to the formation of a single entity overseeing the management of a diffraction-image data/metadata archive, with the IRRMC and similar efforts serving as data portals. The Protein Data Bank archive is currently operated this way by the Worldwide Protein Data Bank organization (http://wwpdb.org), which encompasses three regional data centers located in the United States (RCSB Protein Data Bank or RCSB PDB; http://rcsb.org), the United Kingdom (Protein Data Bank in Europe or PDBe; http://pdbe.org) and Asia (Protein Data Bank Japan or PDBj; http://pdbj.org), plus a specialized NMR data repository (BioMagResBank or BMRB; http://bmrb.org) that operates out of both Madison, Wisconsin, USA and Osaka, Japan.

The IRRMC platform was designed with an alternative view in mind. Our system architecture allows data to be hosted on many different servers, which can in turn be federated to establish a distributed global archive system, with various nodes supporting shared access to all archived data/metadata using a common data-exchange format/protocol. Such a federated system would allow regional support of individual resources. Individual federation members could benefit their user communities by incorporating additional functionalities, such as storing or linking to related experimental data (*e.g.* detailed expression and purification workflows, or the outcomes of biological assays and ligand-screening campaigns).

In conclusion, the IRRMC resource provides open access to the original, unprocessed data together with associated, curated metadata. This resource not only adds another layer of transparency and opportunity for validation for structures in the PDB, but also allows contributing crystallographers to participate in the process of improving the science and technology of X-ray crystallography. Structures with available primary data will be likely to contribute to the continued improvement in structure quality as new techniques, procedures and algorithms become available: a step towards the vision of the structural biology universe as a 'dynamic body of continuously improving results in symbiosis with continuously improving methods and software' (Terwilliger, 2014; Terwilliger & Bricogne, 2014). Within this vision, metadata should allow one to traverse the complete structure-determination trajectory, from processing raw diffraction images, through integrating reflections to give intensities and structure factors, and culminating in a refined atomic level structural model of a macromolecule. The IRRMC and similar resources serve as a proof of concept and foster the first stage of this trajectory by archiving raw diffraction data and associated metadata from X-ray crystallographic studies of biological macromolecules.

Acknowledgements

The authors thank all depositors of diffraction experiments, in particular the researchers and PIs of the CSGID, SSGCID, JCSG, MCSG, NYSGRC and SGC, as well as George Phillips and other depositors from individual laboratories. We also wish to thank James Spencer for reading and commenting on the manuscript, Ivan Shabalin, Katarzyna Handing and Tomek Osinski for helpful discussions, and Monika Grabowska for helping to implement the search functionalities. This work was supported by NIH Grant U01 HG008424 and by funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under Contract No. HHSN272201200026C. M-AE was supported by a NIGMS Protein Structure Initiative (PSI) grant U54-GM094586 to the Joint Center for Structural Genomics (JCSG).

References

Androulakis, S. *et al.* (2008). *Acta Cryst.* D**64**, 810–814.
Bagaria, A., Jaravine, V. & Güntert, P. (2013). *Comput. Biol. Chem.* **46**, 8–15.
Baker, E. N., Dauter, Z., Guss, M. & Einspahr, H. (2008). *Acta Cryst.* F**64**, 231–232.
Berman, H. M. *et al.* (2006). *Structure*, **14**, 1211–1217.
Berman, H., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.
Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. & Westrip, S. P. (2016). *J. Appl. Cryst.* **49**, 277–284.
Brown, E. N. & Ramaswamy, S. (2007). *Acta Cryst.* D**63**, 941–950.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Chang, G., Roth, C. B., Reyes, C. L., Pornillos, O., Chen, Y.-J. & Chen, A. P. (2006). *Science*, **314**, 1875.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Choi, J., Chon, J. K., Kim, S. & Shin, W. (2008). *Proteins*, **70**, 1023–1032.
Collins, F. S. & Tabak, L. A. (2014). *Nature (London)*, **505**, 612–613.
Cooper, D. R., Porebski, P. J., Chruszcz, M. & Minor, W. (2011). *Exp. Opin. Drug Discov.* **6**, 771–782.
Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, J. S. (2007). *Nucleic Acids Res.* **35**, W375–W383.
Diederichs, K. & Karplus, P. A. (2013). *Acta Cryst.* D**69**, 1215–1222.
Domagalski, M. J., Zheng, H., Zimmerman, M. D., Dauter, Z., Wlodawer, A. & Minor, W. (2014). *Methods Mol. Biol.* **1091**, 297–314.
ESRF (2016). *Data Policy*. http://www.esrf.eu/files/live/sites/www/files/about/organisation/ESRF%20data%20policy-web.pdf.
Frazer, J. G. (1921). Translator. *Apollodorus, The Library*. *Loeb Classical Library Volumes 121 and 122*. Cambridge, USA: Harvard University Press.
Gabanyi, M. J. *et al.* (2011). *J. Struct. Funct. Genomics*, **12**, 45–54.
Glover, I. D., Harris, G. W., Helliwell, J. R. & Moss, D. S. (1991). *Acta Cryst.* B**47**, 960–968.
Grabowski, M., Chruszcz, M., Zimmerman, M. D., Kirillova, O. & Minor, W. (2009). *Infect. Disord. Drug Targets*, **9**, 459–474.
Guss, J. M. & McMahon, B. (2014). *Acta Cryst.* D**70**, 2520–2532.
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.
Holton, J. M. (2012). Personal communication.
International DOI Foundation (2016). *The DOI System*. https://www.doi.org/.
Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D. & Ioannidis, J. P. (2016). *PLoS Biol.* **14**, e1002333.
Joachimiak, A. (2009). *Curr. Opin. Struct. Biol.* **19**, 573–584.

Jones, T. A., Kleywegt, G. J. & Brünger, A. T. (1996). *Nature (London)*, **383**, 18–19.
Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Cryst.* D**68**, 484–496.
Joosten, R. P. *et al.* (2009). *J. Appl. Cryst.* **42**, 376–384.
Jovine, L., Morgunova, E. & Ladenstein, R. (2008). *J. Appl. Cryst.* **41**, 659.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). *Acta Cryst.* D**70**, 2502–2509.
Luo, Z., Rajashankar, K. & Dauter, Z. (2014). *Acta Cryst.* D**70**, 253–260.
Matthews, B. W. (2007). *Protein Sci.* **16**, 1013–1016.
McGuire, P. (2008). *Getting Started with Pyparsing.* Sebastopol: O'Reilly Media.
Meyer, G. R., Aragão, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). *Acta Cryst.* D**70**, 2510–2519.
Meyer, P. A. *et al.* (2016). *Nature Commun.* **7**, 10882.
Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.
Minor, W., Dauter, Z., Helliwell, J. R., Jaskolski, M. & Wlodawer, A. (2016). *Structure*, **24**, 216–220.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Protein Data Bank (1971). *Nature New Biol.* **233**, 223.
Raaijmakers, H. C. & Romão, M. J. (2006). *J. Biol. Inorg. Chem.* **11**, 849–854.
Ramachandraiah, G., Chandra, N. R., Surolia, A. & Vijayan, M. (2002). *Acta Cryst.* D**58**, 414–420.
Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.
Read, R. J. & Kleywegt, G. J. (2009). *Acta Cryst.* D**65**, 140–147.
Rupp, B. (2012). *Acta Cryst.* F**68**, 366–376.
Sato, Y., Mitomi, K., Sunami, T., Kondo, J. & Takénaka, A. (2006). *J. Biochem.* **140**, 759–762.
Shabalin, I., Dauter, Z., Jaskolski, M., Minor, W. & Wlodawer, A. (2015). *Acta Cryst.* D**71**, 1965–1979.
Tanley, S. W. M., Schreurs, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). *J. Appl. Cryst.* **46**, 108–119.
Terwilliger, T. C. (2014). *Acta Cryst.* D**70**, 2500–2501.
Terwilliger, T. C. & Bricogne, G. (2014). *Acta Cryst.* D**70**, 2533–2543.
Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.
Touw, W. G., Baakman, C., Black, J., te Beek, T. A., Krieger, E., Joosten, R. P. & Vriend, G. (2015). *Nucleic Acids Res.* **43**, D364–D368.
Van Benschoten, A. H., Liu, L., Gonzalez, A., Brewster, A. S., Sauter, N. K., Fraser, J. S. & Wall, M. E. (2016). *Proc. Natl Acad. Sci. USA*, **113**, 4069–4074.
Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.
Weichenberger, C. X., Pozharski, E. & Rupp, B. (2013). *Acta Cryst.* F**69**, 195–200.
Westbrook, J. D. (2012). *Some Economic Considerations for Managing a Centralized Archive of Raw Diffraction Data.* http://www.iucr.org/__data/assets/pdf_file/0009/69597/08-bergen-raw-data.pdf.
Westbrook, J. D. & Bourne, P. E. (2000). *Bioinformatics*, **16**, 159–168.
Westbrook, J., Feng, Z., Burkhardt, K. & Berman, H. M. (2003). *Methods Enzymol.* **374**, 370–385.
Zaborsky, N., Brunner, M., Wallner, M., Himly, M., Karl, T., Schwarzenbacher, R., Ferreira, F. & Achatz, G. (2012). *Acta Cryst.* F**68**, 377.
Zheng, H., Chordia, M. D., Cooper, D. R., Chruszcz, M., Müller, P., Sheldrick, G. M. & Minor, W. (2014). *Nature Protoc.* **9**, 156–170.
Zheng, H., Hou, J., Zimmerman, M. D., Wlodawer, A. & Minor, W. (2014). *Exp. Opin. Drug Discov.* **9**, 125–137.
Zimmerman, M. D., Grabowski, M., Domagalski, M. J., Maclean, E. M., Chruszcz, M. & Minor, W. (2014). *Methods Mol. Biol.* **1140**, 1–25.