ICSTI Insights

THE LIVING PUBLICATION

MAY 2012

EXECUTIVE SUMMARY

Within the *ICSTI Insights* Series we offer three articles on the 'living publication' that is already available to practitioners in the important field of crystal structure determination and analysis. While the specific examples are drawn from this particular field, we invite readers to draw parallels in their own fields of interest. **The first article** describes the present state of the crystallographic living publication, already recognized by an ALPSP (Association of Learned and Professional Society Publishers) Award for Publishing Innovation in 2006. **The second article** describes the potential impact on the record of science as greater post-publication analysis becomes more common within currently accepted data deposition practices, using processed diffraction data as the starting point. **The third article** outlines a vision for the further improvement of crystallographic structure reports within potentially achievable enhanced data deposition practices, based upon raw (unprocessed) diffraction data.

The IUCr in its Commissions and Journals has for many years emphasized the importance of publications being accompanied by data and the interpretation of the data in terms of atomic models. This has been followed as policy by numerous other journals in the field and its cognate disciplines. This practice has been well served by databases and archiving institutions such as the Protein Data Bank (PDB), the Cambridge Crystallographic Data Centre (CCDC), and the Inorganic Crystal Structure Database (ICSD). Normally the models that are archived are interpretations of the data, consisting of atomic coordinates with their displacement parameters, along with processed diffraction data from X-ray, neutron or electron diffraction studies. In our current online age, a reader can not only consult the printed word, but can display and explore the results with molecular graphics software of exceptional quality. Furthermore, the routine availability of processed diffraction data allows readers to perform direct calculations of the electron density (using X-rays and electrons as probes) or the nuclear density (using neutrons as probe) on which the molecular models are directly based. This current community practice is described in our first article.

There are various ways that these data and tools can be used to further analyze the molecules that have been crystallized. Notably, once a set of results is announced via the publication, the research community can start to interact directly with the data and models. This gives the community the opportunity not only to read about the structure, but to examine it in detail, and even generate subsequent improved models. These improved models could, in principle, be archived along with the original interpretation of the data and can represent a continuously

improving set of interpretations of a set of diffraction data. The models could improve both by correction of errors in the original interpretation and by the use of new representations of molecules in crystal structures that more accurately represent the contents of a crystal. These possible developments are described in our second article.

A current, significant, thrust for the IUCr is whether it would be advantageous for the crystallographic community to require, rather than only encourage, the archiving of the raw (unprocessed) diffraction data images measured from a crystal, a fibre or a solution. This issue is being evaluated in detail by an IUCr Working Group (see http://forums.iucr.org). Such archived raw data would be linked to and from any associated publications. The archiving of raw diffraction data could allow as yet undeveloped processing methods to have access to the originally measured data. The debate within the community about this much larger proposed archiving effort revolves around the issue of 'cost versus benefit'. Costs can be minimized by preserving the raw data in local repositories, either at centralized synchrotron and neutron research institutes, or at research universities.

Archiving raw data is also perceived as being more effective than just archiving processed data in countering scientific fraud, which exists in our field, albeit at a tiny level of occurrences. In parallel developments, sensitivities to avoiding research malpractice are encouraging Universities to establish their own data repositories for research and academic staff.

These various 'raw data archives', would complement the existing processed data archives. These archives could however have gaps in their coverage arising from a lack of resources. Nevertheless we believe that a sufficiently large raw data archive, with reasonable global coverage, could be encouraged and have major benefits. These possible developments, costs and benefits, are described in our third and final article on 'The living publication'.

John R Helliwell[1], Thomas C. Terwilliger[2], Brian McMahon[3]

1. School of Chemistry, University of Manchester, UK; IUCr Representative to ICSTI.
2. Los Alamos National Laboratory, USA; Member of the IUCr Commission on Biological Macromolecules.
3. International Union of Crystallography, Chester; IUCr Representative to CODATA.

All three authors are Members of the IUCr Diffraction Data Deposition Working Group.
IUCr is the International Union of Crystallography (see http://www.iucr.org/).

John Helliwell [John.helliwell@manchester.ac.uk]
Tom Terwilliger [terwilliger@lanl.gov]
Brian McMahon [bm@iucr.org]

Article 1

The Living Publication has existed for many years for crystallographers

John R. Helliwell and Brian McMahon

1. Setting the scene

The phrase ‘living publication’ is beginning to be used in the scholarly publishing world to describe the various facets of research articles published online that differentiate them from the traditional paradigm of the printed paper as the authoritative – but static – record of scientific research. These include: rich linking to related publications that facilitate location and retrieval of cited articles to read alongside the current publication; reformatting of the article online in ways that allow new navigation through text, figures, tables and supporting materials; change in content of an article owing to its publication on preprint servers or on platforms permitting open review and subsequent revision; and the ability to interact with and interrogate associated experimental data in support (or refutation!) of the author’s arguments.

The International Union of Crystallography (IUCr) has published primary research journals since 1948, and currently has a stable of eight titles covering the many fields of research involved in the discipline of crystallography. Two of these are online only; one is fully open-access while the others offer a hybrid open-access/subscription model; and all are committed to using electronic publishing technologies to add maximum value to the published article, and to make them vibrant and living publications.

The IUCr, for example, is one of the first publishers to implement the ‘CrossMark’ service (CrossRef, 2012) that stamps the online version of record and provides information about updated content, associated supplementary material, access policies etc. Its journals implement linking from the online publication to cited articles and web resources, and also to associated data sets. And it has from the outset of online publishing provided open access to supporting data sets and subsequently to data validation reports. It has also implemented enhanced figures that allow the three-dimensional visualization and analysis of structural models within the article itself.

Crystal and molecular structure determination forms a major part of modern crystallographic research, and is an exercise that involves systematic collection, processing and analysis of well-defined data sets. As such, it affords particularly rich opportunities to add value to the scholarly article through linking and interaction with data; and it also demonstrates the additional opportunities and challenges that arise when the publication takes on a life of its own. In these *ICSTI Insights* notes, we describe some of the successes we have enjoyed in treating structure report articles as living publications, and we highlight some of the ways in which the entire scholarly publishing paradigm may evolve in consequence. While we focus necessarily on our own field of crystallography, where data management procedures already have an established place in the publication lifecycle, we invite our readers to look for parallels in their own disciplines and areas of expertise.

2. Crystal structure reports: data and publication

We begin by describing the type of experiment that gives rise to crystal structure reports, and identify the different sets of ‘data’ that are relevant in the process of publishing the results of such experiments. Again, while we describe (we hope in not too much detail) specifics of the crystallographic experiment, we hope that readers will readily draw parallels in other contexts.

Figure 1 is a schematic of an X-ray diffraction experiment to determine a molecular structure. Other probes (such as neutrons or electrons) may be used, but X-ray diffraction is the most widely used technique for structure determination and so it is highlighted here.

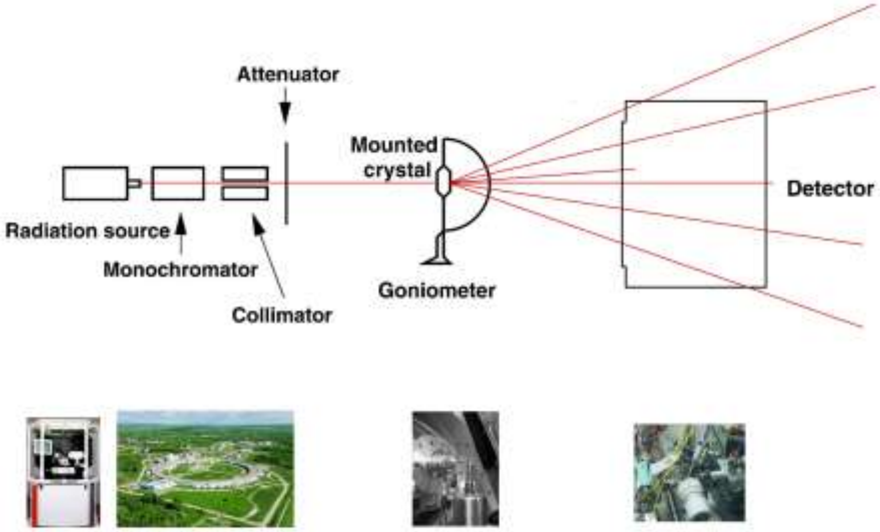


Figure 1. *Schematic of a crystal structure determination experiment*.

Upstream of the crystal is a radiation source – sometimes a benchtop-scale X-ray generator; sometimes a synchrotron facility occupying hectares of countryside, that directs a highly-collimated beam of radiation at the mounted crystal sample. Downstream of the crystal is a detector that records the positions and intensities of the many beams diffracted from the scattering planes of packed atoms in the crystal. By analyzing the positions and intensities of these beams, the locations of the scattering planes of atoms can be deduced, and thus the molecular structure itself determined.

In the course of publishing an article describing the molecular structure obtained from such an experiment, we might distinguish five distinct types or categories of data:

(1) *Primary experimental data.* These are the data generated directly by the experimental apparatus. For many years these were photon counts from a point detector, but in more recent times image detectors have captured real-time images of the diffraction spots (that is, the cross-sections of the diffracted beams where they intercept the detector plate), such as the example in Figure 2. In much of our discussion we shall refer to these for brevity as ‘raw’ data, although in many scientific experiments some early-stage processing of truly raw data may be performed within the detector electronics. [This is certainly the case, for example, in the field of large particle physics experiments such as the Large Hadron Collider, and is likely to be a strategy adopted in many modern experiments that have the potential to generate greater volumes of truly raw data than can be handled by the capacity of available computer systems. The future project in radio astronomy of the square kilometre array is expected to be the largest generator of raw data on the planet; see article 3.]

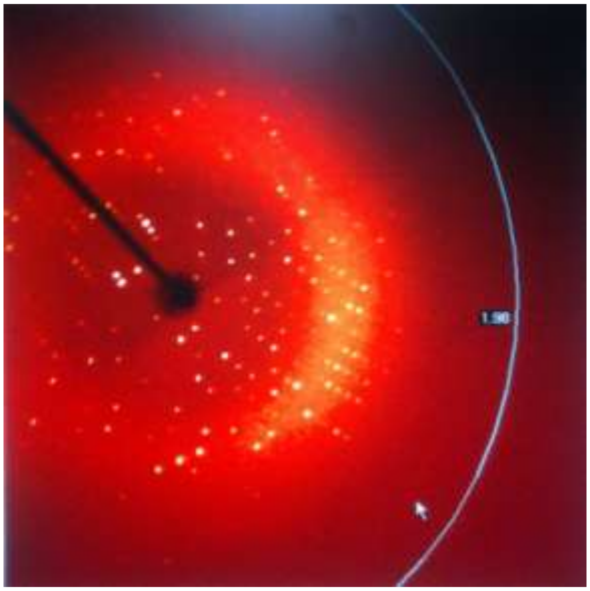


Figure 2. *An example protein crystal diffraction image; hundreds or even thousands of these are measured from each crystal.*

(2) *Processed numerical observations arising from some more or less standardized data reduction procedures*. These may simply refer to the application of calibration or scaling factors to otherwise raw data. In crystallography, the most common such reduced data are structure factors, simple tabulations of the positions and intensities of the diffraction spots. The positions are actually related to the location of scattering planes in the crystal, and hence imply some early interpretation of the data. They should certainly not be characterized as ‘raw’ data.

Nevertheless, in most cases the indexing and other corrections applied are considered so standard and so reliable that, for most experiments, a structure-factor listing would satisfy another researcher's desire to examine the experimental results.

(3) *The derived structural information*. In our example, these are the three-dimensional atomic coordinates, atomic displacement parameters and lists of bond distances, angles and torsions that at one time were tabulated in full as the 'meat' of the publication. They represent the result of the experiment; but are themselves numerical data that can be further analysed in ongoing research. Crystallography has a long and honourable tradition of curated databases of such information, many founded in the 1970s: the Inorganic Crystal Structure Database (ICSD), Cambridge Structural Database (CSD) and Protein Data Bank (PDB), now managed as a WorldWide (wwPDB) consortium.

(4) *Variable parameters in the experimental set-up or numerical modeling and information.* These are often vital in the correct analysis of the data, but are often the most difficult to record because they are buried in executable computer code, or may be changed heuristically or in a somewhat *ad hoc* manner depending on the behaviour of numeric computational procedures. We shall not discuss these in any great detail within this series of articles, but it is worth bearing in mind that they may be of the greatest interest in cases where a researcher carries out a redetermination of a structure and finds differences from that originally published.

(5) *The bibliographic and linking information* that would at one time have been considered the only 'data' in which publishers had an active interest – author names and affiliations, article title, publication reference etc. Again, we do not discuss these in detail in this series, but we do acknowledge that the IUCr as publisher makes every effort to handle these according to best current industry practice and standards.

A significant factor in the IUCr's transition to electronic publishing was the adoption within the crystallographic community of a data exchange standard – the Crystallographic Information Framework (CIF; Hall, Allen & Brown 1991) – that made no fundamental distinction between these different types of data; or, indeed, between 'data' and 'metadata'. In the philosophy of CIF, metadata are simply data that are of secondary interest to the current focus of attention. As a consequence of this, CIF was designed in part to be a suitable transmission medium for structural papers, and is today the only article format acceptable for submission to two of the IUCr journals. For an account of the role of CIF in IUCr journal publishing, see Strickland & McMahon (2008).

3. Crystal structure reports as distributed multi-component publications

The IUCr, through its Commissions and Journals, has for many years taken an active lead in emphasizing the importance of publications being accompanied by data and by the detailed interpretation of the data in terms of atomic models. Such a lead has been followed as policy by numerous other journals in the field and its cognate disciplines. As mentioned above, even before the development of electronic publishing platforms, crystallographers had a natural home for their derived data in the structural databases and archiving institutions such as the PDB, CSD and ICSD. Nowadays the databases play a complementary role to many journals, archiving the models that represent the interpretations of the data (consisting of atomic coordinates with their displacement parameters), along with (in some cases) processed diffraction data from X-ray, neutrons or electron diffraction studies. For small-molecule structures as determined in the field of chemical crystallography, IUCr journals require both the model data and the processed experimental data to be submitted with the article. These are validated as part of the peer-review procedures of the journal, and they are also made available to readers upon publication. Not all journals in the field follow this practice, but a large proportion do make use of the validation procedures developed by the IUCr in their own workflows for handling crystal structure reports.

In Figure 3 we illustrate the flow of numerical data in crystallographic publishing, from the experiment through the subsequent analysis, publication and deposition in databases. It is a simplified representation, but we wish to make a general point with this Figure. Different workflows exist within different fields of crystallography and amongst different journals. This is a potential problem within our field, and even more so for generic publishing platforms that aim to provide services across a wide range of disciplines. However, by showing the different workflows acting in parallel amongst different stakeholders, we suggest that it is possible to delegate some aspects of data handling to external agents, so that journals can focus on what they do best, and leave the specialized aspects of research data management to experts in the individual scientific fields. If general information interchange procedures and protocols are developed between databases, validation services, data archives and other software services, then it will be easier for journals to manage the publication of more complex multi-component documents.

Because of the existence of curated data archives, there are usually clear linkages between publication and deposited data through the assignment of managed identifiers. Historically the databases assigned identification codes (CCDC ‘refcode’, PDB ‘deposition code’) that could be cited in publications. More recently, publications themselves are uniquely identified through digital object identifiers (DOIs), allowing linking back to the publication from database records; and, even more recently, the databases have begun to assign DOIs to their data depositions, a move that will facilitate bidirectional linking and can easily be extended outside the specific practices of the crystallographic community. We see already that in a growing number of scientific fields, separate data publications are emerging, which take advantage of persistent identifiers assigned by organizations such as DataCite to allow data citation and attribution, and to link to primary research literature (Brase *et al*., 2009).

It is then clear that the DOI may be used as a simple vehicle for linking publications to *raw* experimental data, so long as the raw data are stored reliably and have a registered DOI.

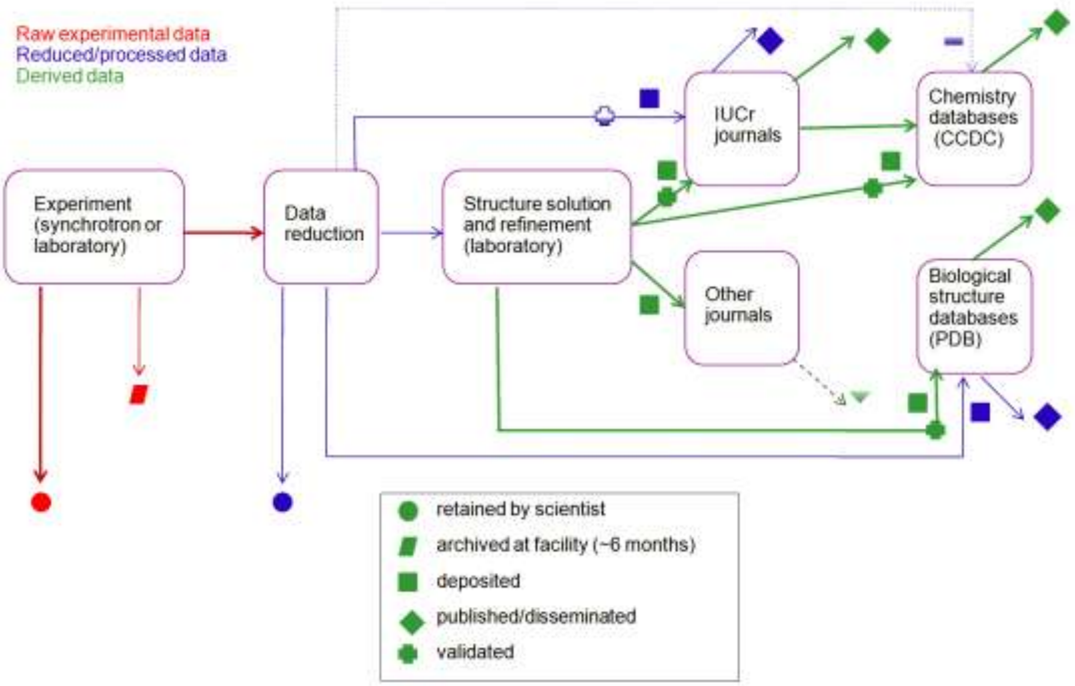

Raw experimental data
Reduced/processed data
Derived data
Experiment (synchrotron or laboratory)
Data reduction
Structure solution and refinement (laboratory)
IUCr journals
Other journals
Chemistry databases (CCDC)
Biological structure databases (PDB)
retained by scientist
archived at facility (~6 months)
deposited
published/disseminated
validated

Figure 3. *A schematic diagram showing the flow of data from experiment through analysis, publication and deposition.*

In the current online age, a reader of a crystallographic structure report has available not only the printed word that records the author's account of a work of scholarly research, but also associated numerical data that provide opportunities to display and explore the results with molecular graphics software of exceptional quality. In some areas of crystallography, the routine availability of processed diffraction data allows readers to perform their own direct calculations of the electron density (X-ray and electrons as probes) or nuclear density (neutrons as probe) on which the published molecular models are directly based. This current community practice is described further in this article.

In the current examples we focus on biological crystallography; similar interactions between publication and deposited data are possible with chemical crystallography, although the practical details vary somewhat between these different research areas. We should note though that in chemical crystallography, where the diffraction data resolution is nearly always at atomic resolution, the chemical 3D models are nearly always what one could perhaps call 'fully mature'.

Accurate crystal structures of macromolecules are of very high importance in biological fields. The current Protein Data Bank (PDB) comprises over 80,000 structures, 90% of which are derived from crystallography (the remaining 10% are derived mostly from NMR, and also electron microscopy). Access to these structures allows readers of a publication to see directly a 'living version' of the results described on the printed page. The processed diffraction data allow detailed checks by diffraction specialists if this is felt necessary by a reader.

4. Connecting the printed word to the derived 'molecular model' and the processed diffraction data

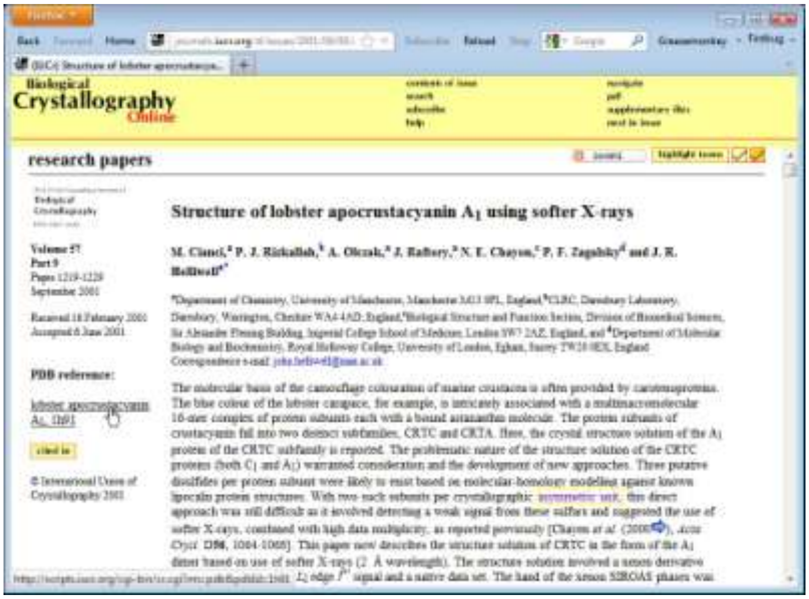
The ICSTI Winter Workshop 2010, on the theme of ‘Interactive publications and the record of science’ (McMahon, 2010; Helliwell & McMahon, 2010), showcased a number of ways in which journal articles could take on a life of their own through allowing the reader to interact (for example through visualization software) with the data associated with a research article. The National Library of Medicine, in collaboration with the Optical Society of America, created a mechanism for archiving data sets that could be retrieved via links embedded in an article and visualized or otherwise analysed on the reader’s desktop alongside the article (Ackerman *et al.*, 2010). In Section 4.1 we illustrate how similar procedures can be performed in crystallography by providing links from articles directly to the data sets deposited in curated databases. In Section 4.2 we illustrate how the connection between article and data is made even more intimate through the publication of interactive diagrams as an integral part of the published article.

4.1 Linking from published article to deposited data

We shall demonstrate how an interested reader of an article describing a protein structure of biological interest (Cianci *et al.*, 2001) can visualize the atomic coordinates and calculate the electron density on which they are based from the deposited processed diffraction.

Figure 4 shows the start of the published article on the online platform of the IUCr. Alongside the abstract is a direct link to the deposited data in the PDB. Upon following this link, the reader is taken to a page on the PDB site (Figure 5) that allows access to the model and experimental data sets in a variety of formats (as well as linking back to the publication, and onwards to related entries in other biological structure and sequence databases).

From the PDB it is straightforward to download the data and perform a separate analysis of some or all of it. Figure 6 shows a zoomed-in view of a portion of this structure (1h91) displayed within the molecular graphics program *COOT* (Emsley *et al.*, 2010), showing the molecular model and the electron density calculated from the processed diffraction data on which the model is based. The amino acid shown is chosen for its benzene-like amino acid side chain with which many ICSTI readers will be familiar; it is called phenylalanine ('PHE') and the number '63' indicates that it is the 63rd amino acid in the protein's amino-acid sequence, which in this case totals 181 amino-acid residues.



Biological Crystallography Online

research papers

Biological Crystallography

Volume 57
Part 9
Pages 1219-1229
September 2001

Received 18 February 2001
Accepted 6 June 2001

PDB reference:

lobster apocrustacyanin A1, 1H91

cited in

© International Union of Crystallography 2001

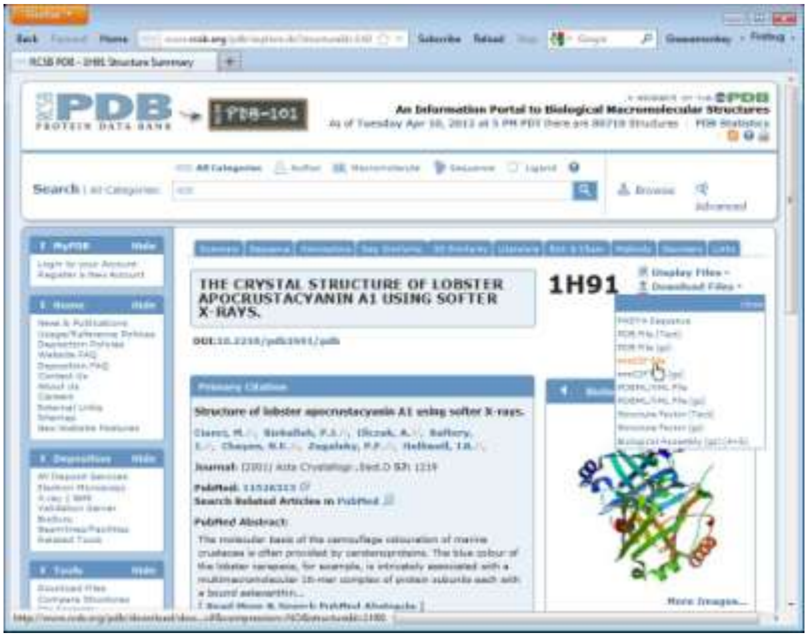
Structure of lobster apocrustacyanin A_1 using softer X-rays

M. Cianci,[a] P. J. Rizkallah,[b] A. Olczak,[a] J. Raftery,[a] N. E. Chayen,[c] P. F. Zagalsky[d] and J. R. Helliwell[a*]

[a]Department of Chemistry, University of Manchester, Manchester M13 9PL, England, [b]CLRC, Daresbury Laboratory, Daresbury, Warrington, Cheshire WA4 4AD, England, [c]Biological Structure and Function Section, Division of Biomedical Sciences, Sir Alexander Fleming Building, Imperial College School of Medicine, London SW7 2AZ, England, and [d]Department of Molecular Biology and Biochemistry, Royal Holloway College, University of London, Egham, Surrey TW20 0EX, England
Correspondence e-mail: john.helliwell@man.ac.uk

The molecular basis of the camouflage colouration of marine crustacea is often provided by carotenoproteins. The blue colour of the lobster carapace, for example, is intricately associated with a multimacromolecular 16-mer complex of protein subunits each with a bound astaxanthin molecule. The protein subunits of crustacyanin fall into two distinct subfamilies, CRTC and CRTA. Here, the crystal structure solution of the A_1 protein of the CRTC subfamily is reported. The problematic nature of the structure solution of the CRTC proteins (both C_1 and A_1) warranted consideration and the development of new approaches. Three putative disulfides per protein subunit were likely to exist based on molecular-homology modelling against known lipocalin protein structures. With two such subunits per crystallographic asymmetric unit, this direct approach was still difficult as it involved detecting a weak signal from these sulfurs and suggested the use of softer X-rays, combined with high data multiplicity, as reported previously [Chayen et al. (2000), Acta Cryst. D56, 1064-1066]. This paper now describes the structure solution of CRTC in the form of the A_1 dimer based on use of softer X-rays (2 Å wavelength). The structure solution involved a xenon derivative L_1 edge f'' signal and a native data set. The hand of the xenon SIROAS phases was

Figure 4. The online article carries an easily found link to the associated data (both three-dimensional molecular model and processed diffraction intensities) in the Protein Data Bank. The link is immediately below the cursor in the left-hand column, alongside the article abstract.



RCSB PDB - 1H91 Structure Summary

An Information Portal to Biological Macromolecular Structures

THE CRYSTAL STRUCTURE OF LOBSTER APOCRUSTACYANIN A1 USING SOFTER X-RAYS.

1H91

DOI:10.2210/pdb1h91/pdb

Primary Citation

Structure of lobster apocrustacyanin A1 using softer X-rays.

Cianci, M., Rizkallah, P.J., Olczak, A., Raftery, J., Chayen, N.E., Zagalsky, P.F., Helliwell, J.R.

Journal: (2001) Acta Crystallogr.,Sect.D 57: 1219

PubMed: 11526313

Search Related Articles in PubMed

PubMed Abstract:

The molecular basis of the camouflage colouration of marine crustacea is often provided by carotenoproteins. The blue colour of the lobster carapace, for example, is intricately associated with a multimacromolecular 16-mer complex of protein subunits each with a bound astaxanthin...

[Read More & Search PubMed Abstracts]

Figure 5. The 'entry page' for the deposited structure at the Protein Data Bank includes selected images of the molecular structure, the ability to visualize the complete model in three dimensions, links to the model and experimental data files (including processed diffraction data), and links to related entries in other databases and publications.

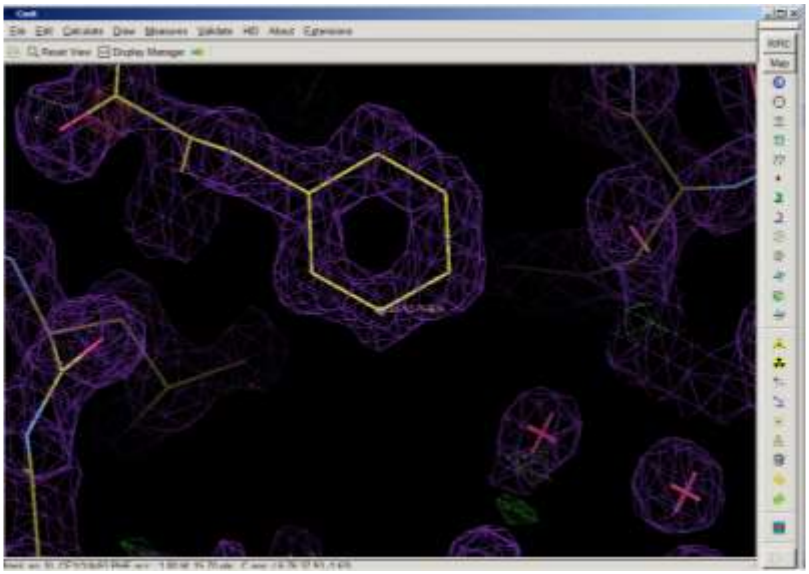

Coot
File Edit Calculate Draw Measures Validate HID About Extensions
Reset View
Display Manager
RMC
Map

Figure 6. *Representation using COOT of the molecular model. The yellow lines connect the positions of carbon atoms to form the three-dimensional skeletal representation of the structure; other colours identify other bonded atoms. The purple mesh is a contour map of the local electron density derived from the structure factors, and indicates the three-dimensional space-filling extent of the atoms.*

Indeed, the community has put in place not simply archives of publication and data, but web-based services for carrying out a variety of analyses. If one does not have access to a program such as *COOT* (or is not skilled in its use), it is possible to visualize the molecule and computed electron density directly (Figure 7) using a web viewer hosted at the Electron Density Server of the University of Uppsala (Kleywegt *et al.*, 2004).

This example then illustrates the detailed way in which the reader of the article can select at will many different details of the primary results, the protein coordinates, and the primary, processed, diffraction data.

4.2 Interacting with molecules

When the journal *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* was launched in 2005, it was in recognition of the need for a fast and streamlined communications journal for structural biology and crystallization results. In 2008 the Editors introduced as a standard feature the ability to render the protein or nucleic-acid structure within a three-dimensional visualization and data analysis applet as an integral part of the publication. The journal, being all electronic, was free of print-on-paper limitations. The Editors remarked:

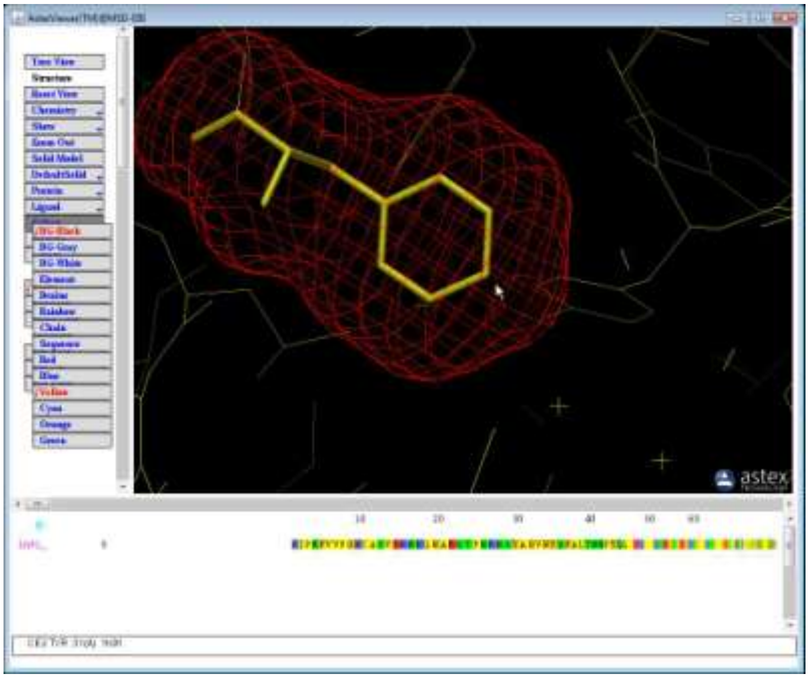


Figure 7. *A web service providing the same ability to visualize data associated with a publication. Compare with Figure 6; the PHE 63 residue is located at the cursor. In this view we have suppressed the electron density present in other regions to help us to focus on the residue of interest.*

'*Visualization of data is one of the most powerful tools available to a scientist. In the biological structural sciences, the visual representation of three-dimensional molecular models often provides insights into biological function. Indeed, different representations (space-filling, trace or cartoon abstractions; colouration by atom species, amino-acid group or secondary structure, etc.) help the understanding of different aspects of the structure, its interactions and biological function ... We are pleased to introduce with this issue a new service, unique to IUCr journals, that allows authors to create enhanced illustrations of molecular structures that will be published as intrinsic components of their articles ... We look forward to the appearance in the journal of an increasing number of enhanced illustrations that will not only be works of beauty in their own right, but that will add immensely to the reader's understanding and enjoyment of the science being presented*.' (Einspahr & Guss, 2008)

While there have been numerous prototypes of such interactive figures, the IUCr journals wanted to integrate them within the standard journal production workflow, and thereby open them to the peer review process, and attempt to preserve their value as illustrations of record despite the volatility of any software implementation.

A very important element of this strategy was to identify a software application that was cross-platform, essentially independent of operating system, built on a sound open-source architecture and supported by an active and committed software development community. The application of choice was *Jmol* (Hanson, 2010), a Java program that could run both as a standalone application and as an applet embedded within a web page.

The next step was to create a tool to allow authors to create rich and complex visual representations, with the option of creating different preferred views of the structure that the reader could select (McMahon & Hanson, 2008). Such a tool had to cater for authors who had no knowledge of *Jmol*, or of the JavaScript or HTML that would be needed to set up a richly interactive online figure. It also had to be fully integrated into the journals' submission and review system, interacting directly with the data files submitted to the journals by the author (in the case of small-molecule structures) or deposited in the PDB (biological macromolecules). Further, it was designed also to create a static counterpart of the initial view (in TIFF or some other standard graphic file format) that would be shown on browsers that did not support Java (or JavaScript), and that could act as a fallback record of the illustration if the rendering application itself could not be supported (or seamlessly replaced) at some point in the future.

Figure 8 is an example of such an enhanced figure. Although at first glance it may appear little (if at all) different from a conventional figure, it is in fact a live rendering of the molecular structure, performed within the reader's browser.

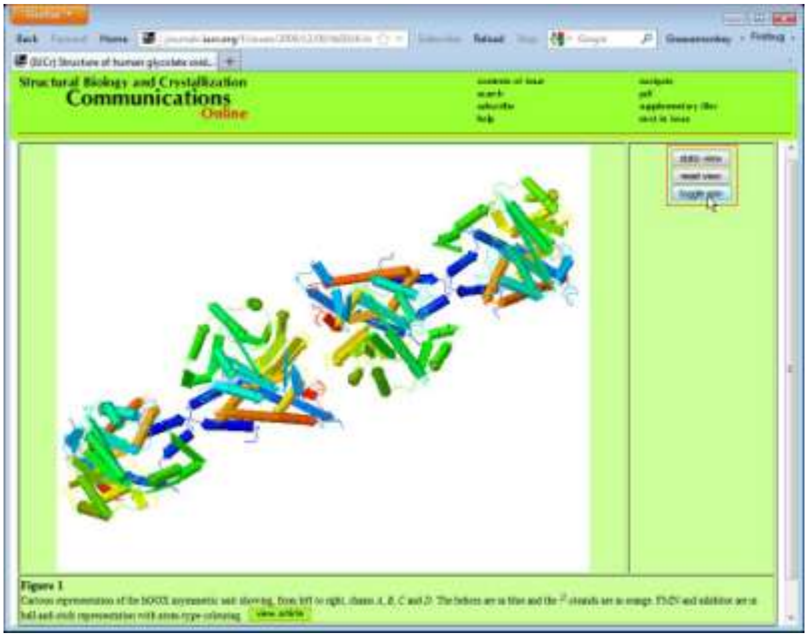


Figure 1
Cartoon representation of the hGOX asymmetric unit showing, from left to right, chains *A*, *B*, *C* and *D*. The helices are in blue and the β strands are in orange. FMN and inhibitor are in ball-and-stick representation with atom-type colouring

Figure 8. *An enhanced figure in the article of Bourhis et al. (2010) allows the reader not only to rotate the molecular representation in three dimensions and zoom in and out to any level of detail, but also to change the style of representation, show distance and angle measurements between any atoms within the molecule, explore how the molecules pack in the crystallographic lattice, generate Ramachandran plots (see article 2 in this series), and display symmetry relationships within the unit cell.*

Notice that enhanced figures in IUCr journals currently handle only the derived structural data (the molecular model), despite the fact that *Jmol* is capable of rendering electron density [see, for example, the interactive Figure 19 in Hanson (2010)]. This arises because only the structural data are stored by the journal (for protein structures, they are copied from the PDB at the time the enhanced figure is created; for small-molecules, they are archived as a matter of policy by the journal). For full integration of all available data in a single visualization, applications must be developed that can dynamically integrate data sets from distributed sources (web security policies currently prevent *Jmol* from doing this); or journals must be able to act as proxies that can copy and locally archive data sets that are referenced by such applications. While this is not very difficult technically, it illustrates the administrative and policy issues that arise when the components of a complex 'living' publication are distributed across multiple sites.

5. Limitations and the need to educate some readers

What are the limitations for readers of our version of the 'living publication'? The coordinates of the atoms in a protein, or nucleic acid, model are just that, a model. The reader should approach any model on which the words in an article are based with a critical appraisal. How is that to be achieved? We have shown examples via the access to the processed diffraction data that the journals and crystallographic databases can now routinely provide.

But is *every* reader competent to make the additional calculations that are now possible via the derived model and processed diffraction data attached to a publication? Unfortunately not.. While we can provide tools that are easy to use, the proper use of those tools does require a deep understanding of the objects that they are manipulating. Thus effective communication, training and education assume a very important role; the Protein Data Bank and the IUCr take these aspects very seriously, and provide active programmes to train and/or for outreach.

6. Summary and next steps

In this article we have demonstrated how a protein structural model associated with the printed journal can be 'brought to life' by the reader via appropriate molecular graphics software. Furthermore the actual electron density can be obtained by direct calculation via the processed diffraction data – in these examples, deposited with the Protein Data Bank but linked to the publication in the quite separate journal.

As readers increasingly subject the printed article to active scrutiny and re-evaluation against the experimental data, so we might come to expect that greater community feedback will occur to prompt routine revisiting, and occasional revision, of published structures. This leads us on to article 2. Most existing models of crystal structures in the Protein Data Bank (PDB) in fact have some correctible errors, and methods for modeling protein structures and for determination of structures are continually improving. Article 2 discusses the ramifications of this along with the opportunities for the continuous improvement of macromolecular crystal structures that are presented to our wider research community through access to data. Once more we shall see how the published structure is no longer a scientific result set in stone, but an interpretation that continues to live and develop.

References

Ackerman, M. J., Siegel, E. & Wood, F. (2010). *Interactive Science Publishing: A joint OSA-NLM project. Information Services and Use*, **30**, 39–50

Bourhis, J.-M., Vignaud, C., Pietrancosta, N., Guéritte, F., Guénard, D., Lederer, F. & Lindqvist, Y. (2009). *Structure of human glycolate oxidase in complex with the inhibitor 4-carboxy-5-[(4-chlorophenyl)sulfanyl]-1,2,3-thiadiazole. Acta Cryst.* (2009). F**65**, 1246–1253

Brase, J., Farquhar, A., Gasti, A., Gruttemeier, H., Heijne, M., Heller, A., Piguet, A., Rombouts, J., Sandfaer, M. & Sens, I. (2009). *Approach for a joint global registration agency for research data. Information Services and Use*, **29**, 13–27

Cianci, M., Rizkallah, P. J., Olczak, A., Raftery, J., Chayen, N. E., Zagalsky, P. F. & Helliwell, J. R. (2001). *Structure of apocrustacyanin A1 using softer X-rays. Acta Cryst.* D**57**, 1219–1229

CrossRef (2012). *CrossMark*. http://www.crossref.org/crossmark

Einspahr, H. & Guss, M. (2008). *A new service for preparing enhanced figures in IUCr journals. Acta Cryst.* (2008). F**64**, 154–155

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Features and Development of Coot*. *Acta Cryst.* D**66**, 486–501

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography*. *Acta Cryst*. A**47**, 655–685

Hanson, R. M. (2010). *Jmol – a paradigm shift in crystallographic visualization*. *J. Appl. Cryst.*, **43**, 1250–1260

Helliwell, J. R. & McMahon, B. (2010). *The record of experimental science: Archiving data with literature. Information Services and Use*, 30, 31–37

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *The Uppsala Electron-Density Server*. *Acta Cryst.* D**60**, 2240–2249

McMahon, B. (2010). *Interactive publications and the record of science. Information Services and Use*, **30**, 1–16

McMahon, B. & Hanson, R. M. (2008). *A toolkit for publishing enhanced figures*. *J. Appl. Cryst.* **41**, 811–814

Strickland, P. R. & McMahon, B. (2008). *Crystallographic publishing in the electronic age. Acta Cryst.* A**64**, 38–51

ICSTI Insights

Article 2

Continuous improvement of macromolecular crystal structures

Thomas C. Terwilliger

Summary

Accurate crystal structures of macromolecules are of high importance in biological and biomedical fields. Models of crystal structures in the Protein Data Bank (PDB) are in general of very high quality, but methods for modeling protein structures and for determination of structures are still improving. We suggest that it is both desirable and feasible to carry out small and large-scale efforts to continuously further improve the models deposited in the PDB. Small-scale efforts could focus on optimizing structures that are of interest to specific investigators. Large-scale efforts could focus on systematic optimization of all structures in the PDB, on redetermination of groups of related structures, or on redetermination of groups of structures focusing on specific questions. All the resulting structures could be made generally available, with various views of the structures available depending on the types of questions that users are interested in answering.

1. Introduction

1.1 Crystal structures of macromolecules

The three-dimensional structures of biological macromolecules such as proteins, DNA and RNA are of high importance in many areas of biology and biotechnology. Structures of proteins and of complexes between proteins, between proteins and small molecules, and between proteins and nucleic acids are all crucial for understanding how these molecules function to catalyze chemical reactions and to control metabolism, growth and development. Structures of proteins bound to candidate drug molecules are highly useful in the development of new pharmaceuticals. Structures of natural and engineered proteins are crucial for rational engineering of these molecules to give them new functions or altered properties.

One of the most important methods for determination of the three-dimensional structures of macromolecules is X-ray crystallography. The essence of this technique is creating crystals of a macromolecule and obtaining a diffraction pattern by hitting the crystals with an X-ray beam (see Figure 1 of article 1). The intensities of the diffraction spots can then be combined with phase information (obtained from parallel experiments or from related crystal structures) to create a three-dimensional picture (an 'electron-density map') of the macromolecule. This picture is then interpreted to obtain a three-dimensional model of the macromolecule, typically including positions of most of the non-hydrogen atoms in the molecule (in many cases the hydrogen atoms are not included however). This procedure has been used to determine many thousands of structures of macromolecules. The hydrogen atoms for ionizable amino acids are sometimes placed based on neutron macromolecular crystallography. This is because the electron of a hydrogen atom is shared in a bond and the already weak X-ray scattering signal of hydrogen is made weaker (indeed a naked proton is only visible to a neutron scattering experiment). The best signal for practical purposes is a deuterium atom substituted for the hydrogen, which, for neutrons, scatters as well as a carbon atom.

In most circumstances the three-dimensional model of a macromolecule is the key product of a crystal structure determination, and often merits an individual publication in a research journal. Most biological or biophysical interpretation of a molecule is done using a model as a representation of what is in the crystal (as opposed to using the electron density map). This means that the details of the model are of great importance, and that the uncertainties and limitations in the model are crucial.

1.2 Errors and uncertainties in three-dimensional models of macromolecules

In general, the structures of macromolecules in the PDB are of very high quality and most features of these structures are well determined. Nevertheless there is always some (small) level of uncertainty in the coordinates of atoms in models representing macromolecular structures. Additionally there may be (usually small) correctible errors in interpretation. Finally, the conceptual framework used to represent a macromolecular structure is itself limited, preventing a complete description of what is in the crystal.

The diffraction data from crystals of macromolecules are typically measured using position-sensitive digital imaging systems (Figure 2 of article 1 shows the type of data that are collected). Owing to the limited amount of X-radiation the crystals can withstand, there are significant uncertainties in measurement. Further, during each of the steps in determining structures of macromolecular, decisions are made about how to treat the data, what outside information to include, and what features to include in the modeling process. These factors complicate the interpretation of the electron-density maps and introduce uncertainty in some details of the final models; these are particularly the mobile parts or 'outer loops' on the surface of biological macromolecules. The three-dimensional structure models obtained from this technique typically do not fully explain the diffraction data, presumably because the features that are included in the models do not represent everything that is present in the crystals.

Owing to the complexity of the analysis, some errors are typically made in interpretation of crystallographic data, and, in addition, alternative interpretations of the data are often possible. Most crystallographic models contain some features that, given a thorough inspection, would generally be thought of as incorrect interpretations. For example, errors could include side-chains in proteins that are placed in physically implausible conformations when the electron density map clearly shows another conformation. The identification of small-molecule ligands bound to macromolecules and their precise conformations and locations can be challenging and lead to errors in interpretation. Additionally, crystallographic models typically do not fully describe the range of structures actually contained in a crystal. For example, parts of a molecule might be represented in one conformation when the data are more compatible with several conformations, and it might not be clear from the data exactly what those conformations are. Normally these errors and limitations decrease markedly if the X-ray data extend to high resolution (resolution is essentially how close features in a structure can be and remain well resolved in the electron density maps; high resolution is typically finer than 2 Å), while they can be very severe for crystal structures determined with X-ray data extending to only low resolution (*e.g.* 3.5 Å or poorer). The errors and limitations in representation of models of macromolecules can limit the utility of these models in interpretation of their biological roles, how drugs bind to the molecules, and what effects changes in the chemical amino-acid sequence of a protein or base sequence in RNA have had on their structures and functions.

1.3 The Protein Data Bank (PDB)

For the past 40 years, most of the models of macromolecules determined by crystallography have been deposited in the Protein Data Bank (PDB), an enormously important resource that includes macromolecular structures determined by nuclear magnetic resonance and electron microscopy techniques as well. The PDB contains models representing over 80,000 crystal structures, with several thousand added yearly. For most of the crystal structures in the PDB, the intensities (or amplitudes) of the diffraction data are deposited as well. These are the 'processed numerical observations' that we introduced in our taxonomy of data categories in article 1. This makes it possible, at least in principle, both to evaluate the models and to improve them.

The PDB is more than a repository of structural information for macromolecules. It is broadly viewed as the definitive repository of this information. This distinction has several consequences. One is that worldwide users of the PDB, many of whom do not have in-depth knowledge about structure determination and its limitations, may use the models from the PDB as if they were unique representations of the structures of the corresponding macromolecules. Another is that any secondary repositories of structural models are not likely to reach a broad audience of users unless they add a great deal of value beyond that available in the structures from the PDB. A third is that the deposited structure in some measure represents a publication in its own right. While not currently recognized as a bibliographic citation, its identity (via unique deposition code, and also a registered DOI), the validation procedures through which it has progressed (and often been improved), and its provenance through recorded, named depositors give it a substantial weight in the relevant academic community.

2. Validation of structures

The limitations of crystal structures of macromolecules have been recognized for a long time, and there has been great effort in the macromolecular crystallography community to develop criteria for evaluating the resulting models. Very recently a task force of structural biologists, in conjunction with the PDB, developed a comprehensive set of criteria for evaluation of crystallographic structures (Read *et al.*, 2011). These criteria will allow the PDB to make structures available in parallel with systematic measures of their quality.

2.1 The current paradigm: one-time interpretation of the data

In the structural biology community, the usual procedure in structure determination is for a single person or group to collect X-ray diffraction data, obtain information on phases, create electron density maps, interpret these maps in terms of an atomic model, and refine that model to optimize its agreement with the diffraction data and with geometrical expectations. Once this procedure is carried out, the resulting model and X-ray diffraction intensities are deposited as an 'entry' in the PDB and become available to anyone who wishes to use them. As mentioned above, it is almost always the models that are used at this stage. It is unusual for the diffraction intensities to be considered by the end users of information from the PDB.

In most cases, the interpretation of the crystallographic data made by the group that carried out the structure determination is the only one that exists today in the PDB. There is a mechanism for the depositor to update their structure, removing the existing entry and replacing it with a new one, but this is done relatively infrequently. There is also a mechanism for anyone at all to use the deposited data, create a new model and deposit it as a new PDB 'entry', however this is rarely done.

3. Automation of macromolecular structure determination and analysis

In the past decade the process of determining the structure of a macromolecule by X-ray diffraction has become increasingly automated. It is now possible in most cases to carry out all the steps from integration of diffraction intensities to interpretation of the data in terms of a nearly-final atomic model in an automated fashion. The final steps of checking the structure, fixing small errors, and interpretation of regions in the electron-density map that involve multiple conformations are normally still done manually, however.

Recently the ability to automate many aspects of structure determination has been applied systematically to a periodic reanalysis of entries in the PDB that contain X-ray data (Joosten *et al.*, 2012). The automated PDB_REDO system carries out validation, model-improvement, and error checking on PDB entries and provides updated models that are often improved over the original PDB entries as judged by agreement with crystallographic data and with expected geometry. Procedures for automated crystal structure interpretation continue to improve, and it seems likely that in the near future fully automated procedures for structure determination of macromolecules may be applied in many cases.

4. The current focus of structural biology community is on models rather than data

For the first 20 years of the PDB (~1970–1990), most structural biologists deposited only the three-dimensional models of the structures they had determined and not the crystallographic data. There are many reasons why this was done. Probably the main reason was that the models are what can be used to interpret the functions and properties of the macromolecules, and the crystallographic data are just a means to obtain a model. Once the model was obtained, the crystallographic data seemed almost unnecessary. More recently it became widely accepted that making some form of crystallographic data available was essential for validation of structural information, and currently nearly all deposits of crystal structures in the PDB are accompanied by crystallographic data. Nevertheless the focus of the worldwide community of users of data from the PDB remains on the models rather than on the crystallographic data. Correspondingly, the access of information in the PDB is focused at the level of a PDB entry, which for crystallographic data normally consists of a single model and any supporting data and metadata.

4.1 Why crystallographic data are rarely reinterpreted and redeposited in the PDB today

It might seem surprising that the models in the PDB are *not* updated systematically and made available as new and improved methods for crystal structure analysis are developed. It is well known that some degree of uncertainty and levels of error exist in crystallographic models, and increasingly automated methods for structure determination are becoming available. There are both practical and sociological reasons why this is infrequently done.

One practical reason models in the PDB are infrequently updated comes about because users of the PDB often do not have detailed knowledge about how to choose which model is the most appropriate one for their uses. This means that if many models were available, there would be confusion about which one to use. Another reason models are not updated is that if a series of models representing a structure were all to be deposited in the PDB and a set of papers was published describing features of the structure, then there could easily be confusion about the description of the model in a publication and which model in the PDB it is associated with. All the coordinates described in the publication would change slightly even upon simple re-refinement of a structure. A reader would then have to refer to the exact structure that the authors used at the time in order to compare with the published information. A third practical problem is that updated versions of structures could have different nomenclature or different numbers of atoms in the model (if some structures were incomplete). These simple changes would make comparisons between publications and any updated structures more difficult. A fourth practical reason is that it requires a great deal of work to deposit a structure in the PDB. The structure and all data and metadata that go with it must be deposited, validated and checked for accuracy. To do this for a large number of structures would be a huge undertaking.

A key sociological reason why models in the PDB often remain static is that structural biologists typically regard a structure as their personal scientific contribution. This view of a structure has consequences both for the scientist or group that determines a structure and for all others. The scientist who determines a structure has invested in its correctness and completeness because he or she has done all the work necessary to determine the structure and has deposited and published it; and may also have published other papers based on this interpretation of the structure. There is therefore substantial motivation not to update the structure unless it is seriously deficient. This view of a structure also has implications for other scientists. If another scientist updates a structure and deposits the updated structure, this could easily be taken as a criticism of the work of the original depositor, even if the intent were solely to build on and possibly add to the work of the original depositor.

5. Continuous improvement of macromolecular crystal structures

We suggest that the structural biology community now can and should systematically improve the tens of thousands of models in the PDB that represent macromolecular crystal structures. A change of focus from a fixed interpretation of a crystal structure to an ever-improving modeling of that structure is technically feasible and is highly desirable as this will improve the quality, utility, and consistency of the structures in the PDB.

5.1 Reinterpretation of the data is feasible

Automation of structure determination algorithms and the availability of crystallographic data for most of the macromolecular structures in the PDB have made it feasible to systematically reinterpret these structures. The full-scale validation of crystal structures in the PDB [*e.g.* the Uppsala electron density server (Kleywegt *et al.*, 2004)] shows that automated procedures can reproduce many of the validation analyses needed to reinterpret structures, including the comparison of models with crystallographic data. The re-refinement and model correction carried out by PDB_REDO further shows that improvement of models can be systematically carried out. These developments, along with the continuous and dramatic improvements in automation of macromolecular structure determination, make it feasible to systematically re-interpret macromolecular crystal structures.

5.2 Reinterpretation of the data is desirable

There are many reasons why it is highly desirable to reinterpret crystallographic data. At a basic level, reinterpretation with modern approaches can easily correct small but clear errors in existing structures. Certainly if two interpretations of a structure are identical except that one has fixed clearly incorrect features in the other, then it would be advantageous to use the corrected structure in any analyses involving that structure.

Also at a basic level, if a consistent set of procedures were to be applied to the structure determination of all structures in the PDB, then the resulting models would have a higher degree of consistency than is currently present. This would reduce the number of differences between models in the PDB that are due only to the procedures and not to actual differences in the crystal structures. A good analysis of how the exact methods used can affect a crystal structure (in this case the bond lengths involving the copper in this structure) was described some 20 years ago (Guss *et al*., 1992).

At a second level, a reinterpretation of a structure with new algorithms or new outside information might yield structural information that was not present in an initial structure. This could include structures of less well-ordered regions ('floppy bits') that could not be modeled in the initial structure or small-molecule ligands that were not interpreted in the initial structure.

At a more sophisticated level, the entire formalism of how crystal structures are described is likely to change over time. At present a structure is typically described by a single model, occasionally containing a few regions represented by multiple conformations. It is likely that in the future most macromolecular crystal structures will be represented by ensembles of models representing the diversity of structures among all the copies of a molecule in a crystal.

Additionally, at present there is too little information on the uncertainties in the models representing macromolecular structures. It will be useful to have a measure of these uncertainties as part of a crystallographic model. It is possible that these uncertainties may also be represented as ensembles of models that are compatible with the data. It is even possible that these uncertainties will best be represented as a group of ensembles, where the group of ensembles represents the range of ensembles that are compatible with the data.

At a very sophisticated level, the most useful model for a particular analysis may depend on what the analysis is intended to achieve. For example, suppose the goal is to determine the structural differences between a pair of proteins that are crystallized in the same crystal form in the presence and absence of a small-molecule ligand. If these two structures are determined and refined against the crystallographic data independently, there are likely to be many small differences between the resulting structures, simply owing to minor differences in procedure. In contrast, if the two structures were refined together, and only differences that are reflected in differences in the crystallographic data were allowed, then the structures would be much more similar, and the differences much more meaningful. Although such a pair of jointly-determined structures may have the most accurate differences in structure, they may or may not have the most accurate individual structures. This example suggests that it may be desirable to have custom sets of structures where all the structures in a group are modeled together so as to have the most accurate set of comparisons of these structures.

Also at a sophisticated level, crystallographic models currently in the PDB may have been based on structural information from earlier structures, but never from later ones. If the entire PDB is reinterpreted, this no longer has to be the case. An approach related to joint refinement of structures is the increasingly important method of using a high-resolution structure as a reference model in refinement of a low-resolution model. This approach essentially uses the expectation that the low-resolution structure is generally similar to the high-resolution structure and that it only differs in places where the low-resolution crystallographic data requires it to be different. Such an approach can now be applied retrospectively to structures in the PDB.

5.3 Reinterpretation is desirable even though the PDB is growing rapidly

It might be argued that because the PDB is growing so rapidly, there is little point in worrying about the structures that are already deposited. It is indeed very likely that soon today's structures will be a fraction of the total in the PDB. On the other hand, the structures that have already been determined represent a tremendously important set of structures, as most of these structures were chosen based on their biological importance. Despite advances in structure determination methodology, carrying out the gene cloning, expression and purification of proteins, crystallization, and X-ray data collection on these tens of thousands of structures all over again will remain prohibitively expensive for a very long time (to redetermine them all today from the beginning might cost in the range of $1-10 billion even using current high-throughput approaches such as those used in the field of structural genomics). Consequently it is indeed important to have the best representation of today's structures as well as of those that are determined in the future.

5.4 Validation and evaluation of reinterpretations of crystal structure data

One of the key reasons that it is appropriate to begin the continuous reinterpretation of macromolecular crystal structure data now is that comprehensive validation tools suitable for widespread application have become available. The validation suite developed for the PDB provides a way to evaluate a structure for geometrical plausibility and fit to the data and to compare these metrics with values for other structures in the PDB determined at similar resolution. This means that systematic criteria are available for evaluation of new models relative to existing ones.

It is important to note that the validation criteria currently used are not direct measures of the accuracy of the structure, if accuracy is defined in terms of the positional uncertainty in the coordinates of the atoms in the model. Rather, the validation criteria are indirect indicators of the accuracy. For example one validation criterion is the Ramachandran plot, the distribution of φ–ψ angles along a polypeptide chain of a protein, where this distribution is compared to those of thousands of well-determined protein structures. A structure with an unlikely Ramachandran distribution is unlikely to be accurate; but there is no simple correspondence between these measures.

Although metrics for structure quality are available, there is not any single metric that can be used effectively to rank structures. Rather, for most metrics there is a histogram of values of that metric for structures in the PDB. For many geometrical criteria there is also an underlying histogram of values from small-molecule structures. A particular structure may be in the most common range for some criteria and an outlier for others. Having unusual values for some metric does not necessarily mean that the structure is incorrect. That could be the case, or it could be the case that the structure has an unusual feature. However, structures with many unusual values for many criteria are generally found to have serious errors. Another type of metric is the Cruikshank–Blow Diffraction Precision Index which gives an overall estimate of uncertainties in atomic coordinates. While this is a useful measure of quality it does not differentiate between different types of errors (inadequacies in the model representation itself compared to coordinate errors for example). Overall existing validation metrics can be used to identify whether a structure is generally similar in quality to other structures in the PDB. It is likely that structures with better metrics overall are generally more accurate than structures with worse metrics, though this has not been demonstrated except for extreme cases.

In addition to quality metrics, it may be important in some cases to evaluate model quality by considering the information that is used in crystal structure determination. As a simple example, some piece of experimental information (*e.g.* anomalous diffraction data) might be used in refinement of one model but not in another. Although this might not change the overall metrics substantially, the structure obtained using the greater amount of experimental information might generally have smaller coordinate errors (provided that the additional experimental data are accurate and not from a crystal with serious radiation damage). Similarly, if two structures are determined using nominally the same data, but one structure is refined using only a subset of the data and the other using all the data, the one obtained with all the data is likely to be the more accurate of the two.

It is also important to further develop the metrics for structure quality. Particularly important will be development of an understanding of the relationship between quality metrics and coordinate uncertainties. Also critically important will be development of metrics that identify the uncertainties in features in electron density maps. Such metrics would greatly strengthen the ability to distinguish features of models that must be present to be consistent with the data from those that simply can be present and are consistent with the data. For structures at low resolution, the latter situation can lead to models that contain features that are not actually present in the crystal.

5.5 Which structure or group of structures should be used in an analysis?

As there is no single measure of the quality or accuracy of a structure, but there are metrics that collectively indicate something about that quality, it is not simple to decide which model is the best representation of a particular structure if several are available. Also, as mentioned above, the structure or set of structures that is most informative may even depend on the question that is being asked.

A useful approach to addressing what structure to choose may be to start with the question that is to be addressed. Some questions could be enumerated in advance and grouped according to the kind of information that is needed to answer them. Others might require a custom analysis of what structures are available to identify the best structure. Still others might require a custom redetermination of structures to best be answered.

Questions that do not depend on the fine details of a model might include, 'What is the overall fold of this protein?' and 'Are these two molecules similar in conformation?' These questions can be answered for a protein molecule without even knowing the conformations of its side chains, and with knowledge of the main-chain atomic coordinates even being rather approximate, as differences of less than about 1.5–2 Å would not change the answers to these questions very much. To answer these questions, any model that is not grossly inaccurate would suffice.

Another set of questions, perhaps the most common set, would depend more on overall correctness of a model. For example, 'What is the buried contact area between the proteins in this complex?' would depend on the positioning of main- and side-chains in the contact region of the two proteins. If two models for this complex based on the same data were available, it is likely that the model that is more generally correct would be more useful. Similarly, if two models that are nearly identical are available and one had clearly incorrect features and the other did not, the one without clear errors would be most likely to be most useful. A related approach would be to start with the original model for a given structure. Then if another model for the structure was available that had some clearly better quality metrics and similar or better quality for all other metrics, and the new model was as complete as the original, that new model might be most likely to be useful.

Other questions might depend on the details of a model. For example, 'What is the coordination of this iron atom?' depends on interatomic distances and correct placement of the iron atom and side-chains coordinating the iron. The model that best answers this question probably will have had a careful consideration of the positions of the iron and coordinating side-chains and their agreement with both the crystallographic data and plausible geometry. If the oxidation state of the iron is known, then the refinement would be expected to include appropriate geometry and distances for that state. Another question depending on the details of a model is, 'What is the distance between this arginine side chain and this glutamate side chain?' Answering this question requires knowing whether these two side chains are largely in single conformations, and if so, what those conformations are. A structure where these two particular side chains agree closely with the electron-density map is more likely to be useful in answering this question than one where they do not.

Still other questions might depend on the relationship between one or more models and require a custom or grouped analysis. 'What is the variability in side-chain conformations depending on temperature?' requires a comparison of several structures. Most likely a useful comparison would involve an analysis of the several structures done using the same refinement and modeling techniques for all the structures.

Another, completely different, approach to choosing which model to analyze will be to use all of them. Nearly all structures will have some useful information. By analyzing all the models and all of their agreements with geometrical considerations and with the crystallographic data it might be possible to identify what is known and what is not known in this structure. A more general approach would be (as mentioned above) to deliberately create many models representing what is in the crystal and to use the variation among these models (or ensembles) as an estimate of uncertainty in the models.

5.6 How will a user find the right model or models to analyze?

If there are many models for each crystal structure, then users will need an easy way to find the model or models that suit their needs. Based on the discussion above, one way to do this would be to have different views of the PDB depending on the question that is being asked. For a user that doesn't have any question in mind or does not share their question, there might be standard views. One of these might be similar to the current view of the PDB, with all original structures or structures revised by their authors shown. Another, as discussed above, might be a view of the original model or the model most clearly improved over the original. Other views might be to include groups of structures that were all redetermined together, or groups of structures redetermined with particular questions in mind.

6. Generating and storing interpretations of crystal structure data

The generation of new interpretations of crystal structure data could be carried out in a variety of ways. Individuals could continue to reinterpret their own data and could reinterpret the data of others, particularly structures in which they have specific interest and expertise. Additionally, however, large-scale efforts (such as PDB_REDO) could systematically reinterpret crystal structure data using standardized procedures. Some efforts might focus on individual structure redeterminations, while others might focus on joint refinement of groups of structures. An important feature of such large-scale efforts would be that the procedures would be essentially identical for all structures, lending an increased consistency to that set of structures as a whole. Both small and large-scale efforts might create multiple reinterpretations of any given structure.

A key outcome of this process is that re-interpretation of crystallographic data would no longer be considered to be a statement that the original model is in error. Rather it would be seen as a process of continuous improvement of models in general.

An important aspect of generating new interpretations of crystal structures is the checking and storage of the data, models, and metadata associated with the new interpretations. As mentioned above, PDB depositions currently require a substantial investment of effort for an individual depositor. This will likely remain the case in the future. For large-scale efforts, however, the corresponding process might be highly automated, perhaps with only a component of manual checking to identify situations that were not handled properly by automated procedures. The availability of existing models that can be used as a comparison with any new models for a particular structure could facilitate the development of a highly effective process for identifying any errors or omissions in new models. This could in turn allow a fully automated process for continuous improvement of models for a structure.

The storage of several or even many models for each structure represented in the PDB presents a significant challenge. Storage at the PDB would be optimal, as this would ensure long-term stability of the models. The PDB may not have sufficient resources to analyze and store such a large number of models, however, and other alternatives could be followed. At present models created by PDB_REDO are stored locally, for example. Such a system would be able to make models available only as long as the local servers were supported. That would mean that some data could be available for a limited period of time only. Though not optimal, this could still be useful. A particular model might have a limited lifetime during which it is an important source of information (and after which some other, better, model serves the same purpose). The significant disadvantage of any system that is not centralized is that it may not be possible to reproduce a particular analysis of the entire PDB at a later date. The counter argument is that it is not always necessary to be able to reproduce an analysis exactly, only to reproduce the process, which would generally give a similar overall result.

6.1 Data and metadata needed to facilitate reinterpretation

The PDB already accepts essentially most of the information that would be important in facilitating reinterpretation of macromolecular crystal structures. Information that the PDB accepts includes crystallographic data, model information, and metadata on the procedures used. As discussed in the third of these articles, the IUCr is considering the utility of storing raw crystallographic images as well.

6.2 Overall metadata

There are several types of metadata that are very helpful in understanding what was done in a structure determination and that can be crucial for carrying out a new structure determination based on the original data. These include:

1. What information was used to obtain the final model (crystallographic data, other structures, restraints libraries)?
2. What type of model was used (*e.g.* TLS, solvent representation)?
3. What general approaches were used to determine the model (molecular replacement, SAD or MAD phasing)?
4. What are the values of all the validation metrics?

By systematizing the whole deposition process and incorporating many error checks, the PDB is already answering the question:

5. Was this model checked to make sure that errors that are not considered in validation did not occur?

In addition to this metadata, the model and the raw (or a processed form of the) data themselves can be collected:

6. What are all the values of all the parameters in the model and their uncertainties?
7. What are the values of all the crystallographic data used to determine the model?

As mentioned above, the raw crystallographic data are currently not normally archived by the PDB. However data that have been subjected to minimal processing (*e.g.* where measurements that may or may not be duplicates of each other depending on the space group of the crystal are not averaged) *can* be deposited and are themselves substantially more useful than fully processed crystallographic data.

All these metadata can be recorded along with any other specialized information about the structure, such as:

8. What are all the components in the crystallization droplet, including any chemical connectivities and modifications, and stoichiometries?
9. What existing structures were used as templates in structure determination and how were they modified?

6.3 Crystallographic data and metadata that is not consistently deposited in the PDB at present

To facilitate systematic reinterpretation of crystal structures, the structural biology community would need to consistently deposit all the information listed above. At present, most of this information is required for PDB deposition. Items that are not required but that would make full reinterpretation feasible would include raw diffraction images and all diffraction data, including data collected at multiple X-ray wavelengths and data from heavy-atom derivatives.

Raw diffraction images, whether exactly as collected or processed to conform to standardized image formats, are an important source of information about a crystal structure because they contain information about disorder in the crystal that is discarded during integration and calculation of diffraction intensities. They also may contain information about multiple crystals that may have been in the X-ray beam. Most importantly, they contain the diffraction data in a form before it has been processed based on a very large number of decisions about space group, crystal shape, absorption, decay, and diffraction physics. It is very likely that methods for interpretation of raw diffraction images will improve in the future, allowing more accurate interpretations of crystal structures. Consequently the preservation of this information will make an important contribution to the future improvement of models of crystal structures.

A second type of data that is not consistently preserved consists of multiple crystallographic datasets that were used in structure determination. In many cases only the crystallographic data corresponding to the final model that is deposited are preserved, and multiple wavelengths or heavy-atom derivatives used to obtain phase information are not deposited. As these crystallographic data contain information about the same or very closely related structures, preservation of these data will very likely be helpful in obtaining improved models of these structures.

7. Conclusions

The continuous improvement and updating of models of macromolecular structures is now becoming feasible. Having systematically-analyzed models available could improve the overall quality and consistency of models, allowing better biological and engineering conclusions to be drawn from these models. There remain some challenging aspects to continuous updating of models, including choosing views of these models for the diverse users of macromolecular structures, developing procedures for storage and checking of models, and providing resources to make these models available. The prospects nevertheless appear highly favorable for some implementation of continuous improvement and updating to be carried out. This will further enrich the possibilities for crystallographic science results to be available within 'the living publication'.

Acknowledgements

The author would like to thank the IUCr Commission on Biological Macromolecules and John Helliwell and Brian McMahon of the IUCr Diffraction Data Deposition Working Group for extensive discussion of the issues presented in this work.

References

Guss, J. M., Bartunik, H. D. & Freeman, H. C. (1992). *Accuracy and precision in protein structure analysis: restrained least-squares refinement of the structure of poplar plastocyanin at 1.33 Å resolution. Acta Cryst.* B**48**, 790–811

Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *PDB_REDO*: *constructive validation, more than just looking for errors. Acta Cryst.* D**68**, 484–496

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004), *The Uppsala Electron-Density Server*, *Acta Cryst.* D**60**, 2240–2249

Read, R. J., Adams, P. D., Arendall, W. B. III, Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G. & Zwart, P. H. (2011). *A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure* **19**, 1395–1412

Article 3

Should the crystallographic community require the archiving of raw diffraction data from a crystal, a fibre or a solution?

John R. Helliwell and Brian McMahon

1. Introduction

Our previous articles have described how immediate access to the numerical data describing a molecular model, and to the processed experimental data forming the basis for such a model interpretation, transforms the scholarly article from a static account of record to a dynamic, living publication. A significant question under examination by the International Union of Crystallography (IUCr) is whether it would be advantageous for the crystallographic community to require, rather than only encourage, the archiving of the raw (unprocessed) experimental data – typically in the form of diffraction images – measured from a crystal, a fibre or a solution. This issue is currently being evaluated in some detail by an IUCr Working Group (see http://forums.iucr.org/). The archiving of raw diffraction data could allow as yet undeveloped processing methods to have access to the originally measured data. Archiving raw data is also perceived as being more effective than just archiving processed data in countering scientific fraud, which exists in our field, albeit at a tiny level of occurrences.

On the other hand, such data sets are orders of magnitudes larger than the structure factors and molecular coordinates that we have previously discussed. The debate within our community about this much larger proposed archiving effort revolves around the issue of 'cost versus benefit'. Costs can be reduced by preserving the raw data in local repositories, either at centralized synchrotron and neutron research institutes, or at research universities.

In parallel developments, sensitivities to avoiding research malpractice are encouraging Universities to establish their own data repositories for research and academic staff. These various raw data archives would complement the existing processed data archives. Such archives would, however, probably have gaps in their global coverage arising from a lack of resources.

Pioneering examples of such raw data archives already exist in the USA, Australia and the UK. In the USA, for example, the Joint Center for Structural Genomics (JCSG) state on their website: '*The Joint Center for Structural Genomics has created a unique repository of X-ray crystallographic datasets for the structures that it has solved and deposited in the Protein Data Bank. This archive contains the experimental data and analyses from the data collection, data reduction, phasing, density modification, model building and refinement of JCSG structures. It also includes full sets of diffraction images for each of our deposited structures, enabling complete reconstruction of the data processing. In most cases, phasing was carried out either* *by SeMet, MAD or Molecular Replacement. These datasets are freely available to the scientific community for developing and testing new algorithms and benchmarking and teaching.*' In Australia a federated repository for raw data is MyTARDIS (Androulakis *et al.*, 2008). In the UK the Diamond Light Source, operational for a few years, is retaining all raw data (Alun Ashton, personal communication). The IUCr Diffraction Data Deposition Working Group intends to carry out a survey of global synchrotron radiation facilities, led by the IUCr Commission on Synchrotron Radiation, to assess the willingness of such facilities to act as local raw data archives.

On the other hand, it is also possible that a sufficiently large raw data archive, with reasonable global coverage, could be developed for the benefit of our particular scientific community; such an initiative would have major benefits.

These possible developments, and their potential costs and benefits, are described here in our third and final article on 'The Living Publication'.

2. Types of data

Figure 1, from a recent report on integration of data and publications (Reilly *et al.*, 2011), illustrates a generic view among many publishers of the types of data associated with published scientific research. Readers will find it interesting to compare this with the taxonomy of data types identified in crystallographic research and publication (article 1 in this series). There are many differences of detail. For example, as we described in article 2, curated data sets at the Protein Data Bank (PDB) have rich provenance, validation and identification that confers on them a significant level of trust within the relevant research community. Nevertheless, the overall picture is similar. As one descends the pyramid, the volume of data grows rapidly, and the level of organization, long-term preservation and, to some extent, quality control all diminish rapidly. Nevertheless, it is the lower levels of the pyramid that underpin and support the published results and conclusions.

We remind the reader in the next few subsections of the different types of experimental data in crystallography that we characterized in article 1. It will be seen that, because of the distributed information model of crystallographic publication, these may be stored in various locations, and that there is consequently some overlap or merging of the different strata in the STM pyramid.

2. 1 Derived data

These are the atomic coordinates, anisotropic or isotropic displacement parameters, space group information and, for biological structures, secondary structure and information about biological functionality. For small-molecule structures these are archived as supplementary materials (in machine-readable format) by IUCr journals. For biological macromolecules they must be deposited with the Protein Data Bank before or in concert with article publication; the article links to the PDB deposition using the PDB reference code. These total approximately hundreds of kbytes in filesize. Relevant experimental parameters, unit-cell dimensions, and other pieces of information that characterize the individual structures are required as an integral part of article submission and are published within the article.

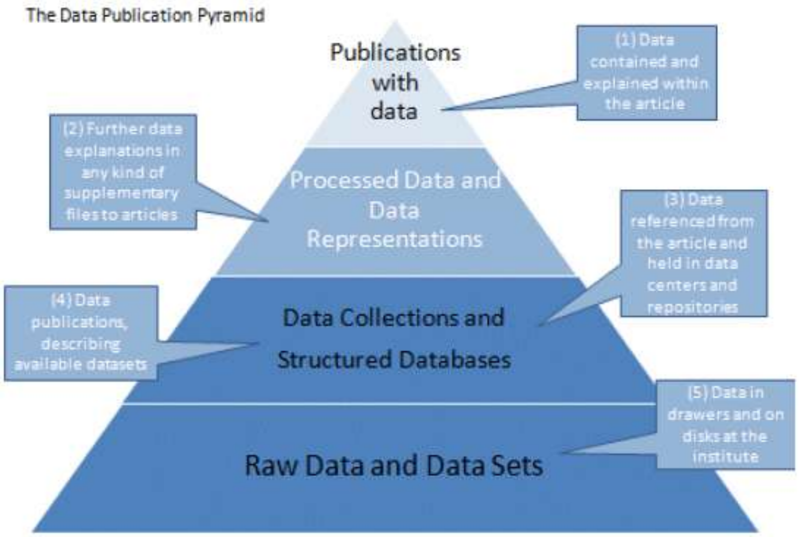

The Data Publication Pyramid
Publications with data
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
Processed Data and Data Representations
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
Data Collections and Structured Databases
(5) Data in drawers and on disks at the institute
Raw Data and Data Sets

Figure 1. *The data publications pyramid, used by the International Association of Scientific, Technical and Medical Publishers (STM) to characterise data types and their place in the publishing process.*

2.2 Processed experimental data

These are the structure factors, tabulations of the positions and intensities of beams diffracted from the different crystal scattering planes, which must be deposited with IUCr journals (for small molecules) or, for biological macromolecules, the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code. These total typically approximately several Mbytes in filesize.

2.3 Primary experimental data

These are usually raw diffraction images (see Figure 2 of article 1). Each image is of order one or a few megabytes in size. Hundreds or even thousands are measured from each crystal, so that the output from each experiment is of order 1 Gbyte in size. In common with many experimental sciences that use electronic detectors, the output data sets are expected to grow larger as each new generation of detector attains higher resolution and data throughput rates.

3. Current IUCr Journal policies regarding raw data

As suggested above, while there are existing requirements for publishing or depositing derived and processed data, IUCr journals have no current, binding, policy regarding publication of diffraction images or similar raw data entities. From Figure 3 of article 1, it may be seen that after initial experimental reduction procedures, the data subsequently flow via one or more parallel streams through the various validation, archiving and publication processes provided by journals and curated databases. The raw data, however, progress no further than the local storage facilities of the scientist's own laboratory, or at larger-scale facilities such as synchrotron or neutron labs. While the service facilities often provide some security of storage on behalf of the scientist, this can be for a strictly limited period of time, and there are no current community norms dictating best practice in this area. We do note the example cited in Section 1 of the Diamond Light Source, which so far has retained all its data; but this is not universal practice, and may not persist into the indefinite future even at Diamond, without specific policy directives.

However, crystallography journals in their Notes for Authors increasingly welcome efforts made to preserve and provide primary experimental data sets. In its 2012 Notes for Authors, *Acta Crystallographica Section D: Biological Crystallography* states that authors are '*encouraged to make arrangements for the diffraction data images for their structure to be archived and available on request*'; this is in likely compliance with research funding agency policy and employer research good practice requirements.

The same journal also encourages retention and deposition of experimental data in more specialized cases: '*For articles that present the results of protein powder diffraction profile fitting or refinement (Rietveld) methods, the primary diffraction data, i.e. the numerical intensity of each measured point on the profile as a function of scattering angle, should be deposited. Fibre [diffraction] data [such as from DNA] should contain appropriate information such as a photograph of the data. As primary diffraction data cannot be satisfactorily extracted from such figures, the basic digital diffraction data should be deposited.*'

Even here, however, it is likely that journal policies will need to adapt progressively to changing procedures in different fields; for two-dimensional powder diffraction data the raw image is considered by many practitioners to be as important as the integrated data.

4. Important principles and standards of data deposition

We offer a collection of reasons for depositing data and making them available alongside a scientific publication. This is not necessarily a complete list; but it provides a useful set of criteria that are relevant, to greater or lesser degree, across most scientific fields.

1. To enhance the reproducibility of a scientific experiment
2. To verify or support the validity of deductions from an experiment

3. To safeguard against error
4. To better safeguard against fraud than is apparently the case at present
5. To allow other scholars to conduct further research based on experiments already conducted
6. To allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
7. To provide example materials for teaching and learning
8. To provide long-term preservation of experimental results and future access to them
9. To permit systematic collection for comparative studies

In some cases these goals are adequately met by processed data. In other cases, they may be satisfied by processed data, but better results can be achieved using raw data. In a few cases, the raw data are essential (for example, in extracting new science from experiments that are not repeatable). In determining the usefulness of deposited raw data against each of these criteria, judgements must be made of the likely benefits versus the real costs that will be incurred.

We illustrate this by considering the suggestion that image data files be compressed prior to storage, in order to reduce the amount of disk space needed. There are various possible compression techniques, and the greatest savings in space will be achieved by using ‘lossy’ techniques (*i.e.* ones in which there may be some unrecoverable loss of information from the raw image). While it seems undesirable to discard any information (especially where collected from an experiment that would be extremely expensive or impossible to repeat), diffraction images archived under a lossy compression scheme would still be useful for prevention of fraud (item 4 in the list above), redoing steps in early decisions in structure determination (item 1), interpretation of diffuse scattering, multiple lattices and solving unsolved problems (aspects of items 5 and 6). However, lossless compression is needed for accurate structure comparisons and information on uncertainties in the models and establishing the effects of different procedures (see article 2 in this series), determining the resolution cut-off (an aspect of item 2), and for future improved data interpretation and analysis of *e.g.* the disorder description by ensembles (also discussed in article 2). To our knowledge, this type of analysis has not been widely discussed within the community.

5. Complying with funding agencies

Also, it is worth restating that, in a number of countries, publishing data with one's publication allows one to comply with one's funding agency's grant conditions. Increasingly, funding agencies are requesting or requiring data management policies (including provision for retention and access) to be taken into account when awarding grants. See, for example, the Common Principles on Data Policy (http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx) of Research Councils UK, and the Digital Curation Centre overview of funding policies in the UK (http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies).

It is worth noting, however, that these policies do not explicitly differentiate amongst derived, processed and raw data. We suggest that funding agencies might usefully develop greater clarity of policy from one to the other stages of the 'data pyramid'.

6. Central issues and examples for a possible future involving raw data archiving

We now come to several linked, central, questions: When might the derived and/or processed diffraction data that are currently deposited become inadequate (*i.e.* when might the raw data become valuable)? How often might this be the case? What are the costs and benefits of retaining and having access to raw diffraction data?

Firstly, the processed diffraction data (structure factors) describe the diffraction structure amplitudes associated with the discrete spots imaged in Figure 2 of article 1. So, what do we perhaps ignore *between* the spots? Figure 2 shows an example of what lies between the spots. It is apparent that significant amounts of the scattered X-radiation may be measured between the Bragg-diffracted beams, yet this information is routinely discarded in structure analysis.

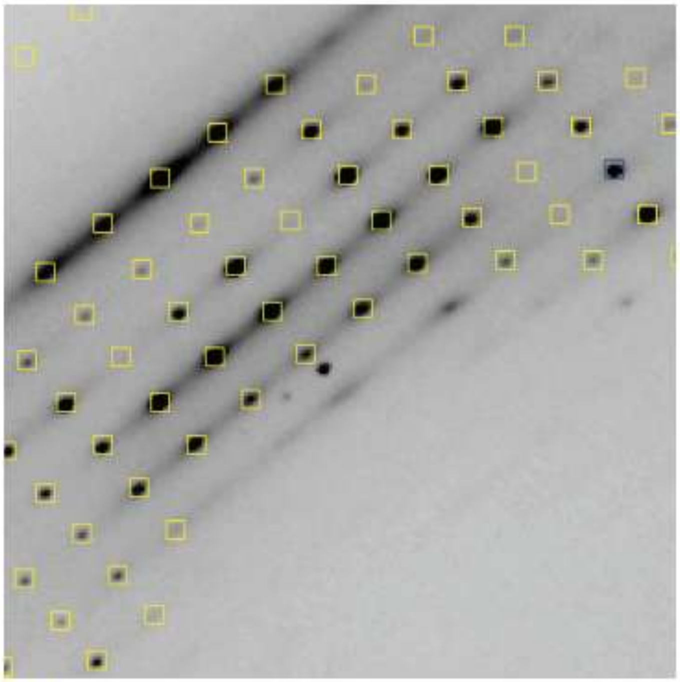


Figure 2. *Example of strong diffuse scattering from an RNA crystal (Jovine et al., 2008). The boxes identify the 'Bragg peaks' – the locations of diffracted beams scattered from regular packing planes of atoms in the crystal. The significant intensities between the peaks contain important information about disorder within the crystal, but this is rarely taken into account by standard data processing software.*

These data are not *always* ignored; Figure 3 is taken from a study of a protein crystal exhibiting ‘macromolecular frustration’, a strained crystal lattice conformation more usually found in inorganic materials with unusual electronic and magnetic properties.

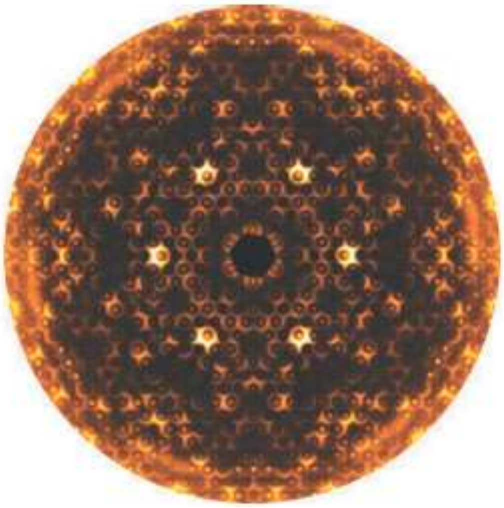


Figure 3. *A predicted reciprocal lattice section (diffraction pattern in effect) normal to the unit-cell* ***c*** *axis. Based on such 'diffuse scattering' predictions the crystal is asserted to suffer from 'molecular frustration' in its crystal packing layout (Welberry et al., 2011), which interrupts the 'simple' periodicity of the molecules packed in the crystal.*

Such diffuse features are not routinely straightforward to interpret at present – an example such as this still proves to be rare. This does not mean that in the future they need to be ignored. That said, the proportion of protein crystals showing such features is not 100% (see Glover *et al*., 1991).

Secondly, the processing of the diffraction spots themselves leads to an early decision by the crystallographer on just what the symmetry layout of a crystal actually is, namely its space-group symmetry. In Nature there are 230 possible space groups for all molecule types. For proteins comprising left-handed amino acids only, these 230 choices narrow down to 'just' 65 space groups, as mirror planes and inversion centres of symmetry cannot be present (otherwise they would generate right-handed amino acids, which do not exist). Errors in space group choice are possible; an example of an error 'nearly made' in JRH's lab experience was with a choice involving space group $I23$ versus $I2_13$ (Harrop *et al*.,1996). In that instance, our own thoroughness avoided the calamity of incorrect structure determination. Nowadays, various validation checks carried out by journals and databases help to avoid such gross erorrs in an ever expanding, but perhaps less experienced, community of researchers. Nevertheless, the possibility of such errors remains. The availability of processed *but unmerged* diffraction data would allow more checking by readers than at present. Indeed, a recommendation made by a recent task force convened by, and reporting to, the PDB has recommended preserving unmerged processed diffraction data (Read *et al.*, 2011).

Thirdly, there are situations of challenging cases where the sample is actually a composite of two or more crystals and more than one diffraction pattern is then obviously visible in the raw diffraction image. The crystallographer will, by likely current practice, choose one such 'crystal lattice', typically the predominant one, and not the other(s); preserving the processed diffraction data of just that one crystal lattice obviously does not include the others, which are lost completely upon deletion of the raw diffraction data.

Fourthly, sometimes the raw diffraction data do not lead to *any* final interpretation in the hands of one crystallographer or laboratory. Such data could be made available, if a researcher finally chooses, to the wider community to attempt structure determination if anyone wishes. This would be a category somewhat similar to the 'negative results' obtained in other scientific fields. The failed attempt could still usefully be described in a research article and could explain what methods had been attempted thus far and what might work in the future etc., all linked to the raw data.

Fifthly, and perhaps most importantly, is the question of the diffraction resolution limit of any crystallographic study and whether the processed diffraction data were in effect artificially truncated at an arbitrary resolution limit even though the diffraction raw images extend to higher scattering angles, as they very often do. This fifth reason, then, to preserve raw diffraction images is that the edge of the diffraction pattern (the 'diffraction resolution') is not so easy to set a community agreed standard for. The pattern fades basically due to the atomic mobilities along with possible static disorder (called atomic displacement parameter effects), and in the case of X-rays and electrons as probes the finite size of the electron charge cloud causes a further drop in scattering of each atom. With neutrons the scattering, being off the much smaller nucleus, does not add to the atomic displacement parameter effect. In special cases the diffraction resolution may be anisotropic due to the nature of a crystal's overall quality. In practice one often-used parameter-descriptor is to simply describe where the average diffraction spot intensities divided by their standard deviations, σ, (*i.e.* $<I/\sigma(I)>$) decrease below 2.0. The community is keen, though, not to artificially cut the data here, as this would falsely eliminate diffraction spots even further from the centre of the diffraction pattern. Indeed, it is hoped that protein model refinement programs should cope formally with the diffraction pattern fade out. This is not general practice, nor is it even championed by software writers if this is coded for in their mathematical algorithms. Indeed, as one watches the diffraction patterns as they are measured, occasional spot intensities do occur well beyond the obvious pattern edge. These occasional spots are known about but are deemed rare, and so small in number to be considered inconsequential; but they are surely – or should surely be – of interest and potential help to define better the molecular models. Deletion of raw diffraction data and/or their loss due to inadequate archiving means a loss to future possible revisions of molecular models using diffraction data beyond a given publication's actual analysis diffraction resolution.

These five situations illustrate sound scientific reasons why it could be useful to archive raw diffraction data, ideally with a DOI registration, so that they can be linked to the relevant publication and, in most cases, associated PDB deposition.

There is a sixth reason that is proposed for the utility of preserving raw diffraction data, namely the prevention of scientific fraud. Thus the raw data would present a much greater hurdle against fabrication. The crystallographic community is somewhat divided on the effectiveness of this, though, in that it may ultimately prove achievable to fabricate raw diffraction data too.

So there are certainly at least five or six reasons to preserve raw diffraction data. How often might such data be accessed by the wider community? What would be the cost of preserving and accessing them?

7. Cost benefit analyses

Costs could be reduced by a number of strategies. In the short term, distributing the burden of storage of raw data among the collection sites (synchrotron facility, neutron research centre or university laboratory), the users' sites (*e.g.* in university institutional repositories) and longer-term central repositories where appropriate (*e.g.* journal supplementary data repositories, university data archives, and, perhaps, commercial sites such as Google and Amazon) would help to avoid large network file-transfer costs. Such a strategy may in any case be forced upon the community in the short term, as there is not universal agreement among the various stakeholders as to their potential role in providing secure long-term storage of raw data. Perhaps we may elaborate on this, since it is an area that is ripe for growth. University researchers could archive their data in their own laboratory, on their laboratory shelves, or more formally, for small data sets in the university institutional repository and for large data sets in the nascent data archive centres that are beginning to be established. Alternatively they could leave their raw data with a central facility, if measured there. Successful implementation of such a policy would depend upon the use of DOIs or similar location-independent identifiers to track the data as it migrates over the various networks involved among collection sites, user sites and archives.

Costs could further be cut by preserving only a proportion of each of the raw data sets, or by using some form of data set compression (see the comments in Section 4 above about the merits of lossy versus lossless compression), or both. It may be useful for the community to consider protocols involving some combination of lossy and losslessly compressed data sets to achieve a balance between retention of forensic quality data and copies of sufficient scientific validity to facilitate continuous improvement of analyses.

The benefits could be maximized by authors, referees and/or editors flagging up cases where preservation of raw diffraction data is going to have a high chance of further utility, *e.g.* because the diffraction pattern showed extensive diffuse scattering, currently ignored, or showed multiple crystal lattices, of which details of just one were provided in the publication.

The weakness of any policy allowing for deletion of raw data sets, though, is that mistakes can be made, and the raw data are then lost forever upon deletion.

8. Summary

Overall, many IUCr Commissions are interested in the possibility of establishing community practices for the orderly retention and referencing (via a DOI) of raw data sets, and the IUCr would like to see such data sets become part of the routine record of scientific research in the future, to the extent that this proves feasible and cost-effective.

These matters are currently under active debate within the crystallographic community, and we draw your attention to the IUCr Forum on such matters at http://forums.iucr.org. While this forum is specifically for crystallographers, we do track developments in other areas. For example, we reference the recent ICSU report of the Strategic Coordinating Committee on Information and Data (ICSU SCCID, 2011). Within this we learn of many other scientific efforts in data archiving; for example, the forthcoming radio astronomy Square Kilometre Array that will pose the biggest raw data archiving challenge on the planet, reaching zettabyte levels of file storage needs. (One zettabyte equals 10^{24} bytes. According to Wikipedia, '*As of March 2012, no storage system has achieved one zettabyte of information*'.) This encourages us to think that our needs as crystallographers are relatively modest – and perhaps thereby amenable to orderly solution!

We hope you have been stimulated by, and even enjoyed, this trilogy of articles, illustrating what we do in crystallography in managing our literature along with our data, and in realizing the ideal of 'The Living Publication'.

Acknowledgements

We are grateful to the members and consultants of the IUCr Diffraction Data Deposition Working Group (DDDWG) for their conscientious work on these matters, and to other IUCr Officers (IUCr Executive Committee, Chairs of Commissions etc.) for their comments and interactions since formation of the DDDWG in August 2011 at the IUCr Congress in Madrid. We thank Loes Kroon-Batenburg for the discussion on the advantages and shortcomings of lossy compression in archiving image data, Herbert Bernstein for his insights into the *modus operandi* of a distributed storage strategy, and Pamela Whitfield for comments on the future needs of powder diffractionists. Colin Nave is thanked for suggesting a global synchrotron radiation Facilities survey, which should be extendable to global neutron facilities. Marian Szebenyi is thanked for requiring greater clarity in developing the metaphor of 'The Living Publication'.

References

Androulakis, S., Schmidberger, J., Bate, M. A., DeGori, R., Beitz, A., Keong, C., Cameron, B., McGowan, S., Porter, C. J., Harrison, A., Hunter, J., Martin, J. L., Kobe, B., Dobson, R. C. J., Parker, M. W., Whisstock, J. C., Gray, J., Treloar, A., Groenewegen, D., Dickson, N. & Buckle, A. M. (2008). *Federated repositories of X-ray diffraction images*. *Acta Cryst.* D**64**, 810–814

Glover, I. D., Harris, G. W., Helliwell, J. R & Moss, D. S. (1991). *The variety of X-ray diffuse scattering from macromolecular crystals and its respective components. Acta Cryst.* (1991) B**47**, 960–968

Harrop, S. J., Helliwell, J. R., Wan, T., Kalb (Gilboa), A. J., Tong, L. & Yariv, J. (1996). *Structure solution of a cubic crystal of concanavalin A complexed with methyl alpha–D–glucopyranoside. Acta Cryst.* D**52**, 143–155

ICSU SCCID (2011) Ad-hoc Strategic Coordinating Committee on Information and Data. *Interim Report to the ICSU Committee on Scientific Planning and Review*. Available from http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report

Jovine, L., Morgunova, E. & Ladenstein, R. (2008). *Of crystals, structure factors and diffraction images*. *J. Appl. Cryst.* (2008). **41**, 659

Read, R. J., Adams, P. D., Arendall, W. B. III, Brunger, A. T., Emsley, P., Joosten, R. P., Kleywegt, G. J., Krissinel, E. B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J. S., Sheffler, W. H., Smith, J. L., Tickle, I. J., Vriend, G. & Zwart, P. H. (2011). *A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure* **19**, 1395–1412.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E. & Wilkinson, W. (2011). *Report on Integration of data and Publications.* Available from http://www.stm-assoc.org/integration-of-data-and-publications/

Welberry, T. R., Heerdegen, A. P., Goldstone, D. C. & Taylor, I. A. (2011). *Diffuse scattering resulting from macromolecular frustration. Acta Cryst.* B**67**, 516–524

Prepared for ICSTI by John R. Helliwell, Thomas Terwilliger, Brian McMahon