ICSTI Insights Series

The Living Publication

Synopsis: In crystallography the living publication, which we describe in our first article, is already here. In our second article we anticipate a continuous improvement of macromolecular crystal structures and dynamics descriptions arising from the greater access to the data supporting such living publications; this has obvious ramifications for the original publication. Finally, in our third article, we describe how archiving of raw crystallographic diffraction data might be carried out and the implications for associated publications; and analyze the costs and benefits of routine archiving of raw experimental data.

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

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

All three authors are Members of the IUCr's Diffraction Data Deposition Working Group (see http://forums.iucr.org/)

Overview

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. 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 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, neutrons or electron diffraction studies. In our current online age, a reader has available not only the printed word, but also the chance to display and explore the results with molecular graphics software of exceptional quality, even after just 30 years of developments. Furthermore, the routine availability of processed diffraction data allows readers to perform direct calculations of the electron (X-ray and electrons as probes) or nuclear density (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/). 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'.

Article 1

The Living Publication has existed for quite some years for crystallographers

John R Helliwell and Brian McMahon

1. Setting the scene

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 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, neutrons or electron diffraction studies. In our current online age, a reader has available not only the printed word, but also the chance to display and explore the results with molecular graphics software of exceptional quality, even after just 30 years of developments. Furthermore, the routine availability of processed diffraction data allows readers to perform direct calculations of the electron density (X-ray and electrons as probes) or nuclear density (neutrons as probe) on which the molecular models are directly based. This current community practice is described in this, our first, article.

Here we focus on biological crystallography; we could equally well have focused on chemical crystallography results but our illustration of what is possible can be served well via just one research area. 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 some 70,000 structures, largely derived from crystallography (approx 90%) the remaining percentage from NMR and electron microscopy. These structures allow readers of a publication to see directly a 'living version' of the printed page described results. The processed diffraction data allow detailed checks by diffraction specialists as may sometimes be felt necessary by a reader.

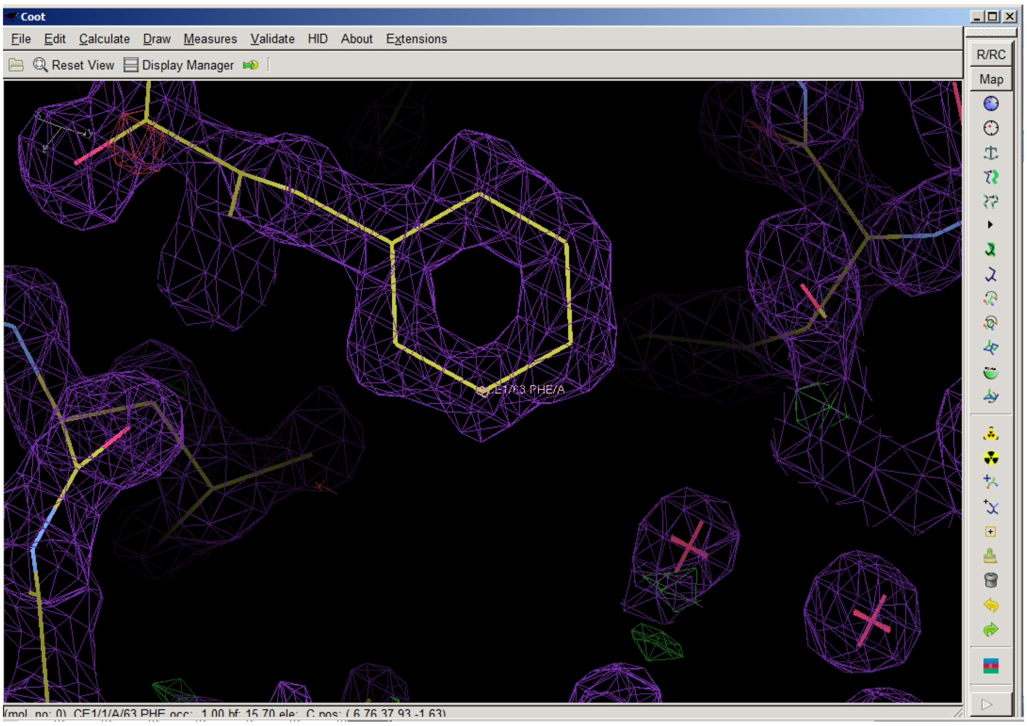
2. Connecting the printed word to the derived 'molecular model' and the processed diffraction data

2.1 An example from Acta Crystallographica D Biological Crystallography

As examples of these additional 'reader opportunities' Figure 1 shows what the reader can do via visualisation of coordinates and calculation of the electron density on which they are based from the deposited processed diffraction data for a publication in *Acta Crystallographica D.*

Figure 1 Zoom in view of a portion of 1h91 [M. Cianci, P.J. Rizkallah, A. Olczak, J. Raftery, N.E. Chayen, P.F. Zagalsky and J.R. Helliwell "Structure of apocrustacyanin A1 using softer X-rays" (2001) Acta Crystallographica Section D-Biological Crystallography D57, 1219-1229.] displayed within the molecular graphics program COOT (P. Emsley, B. Lohkamp, W.G. Scott and K. Cowtan (2010) *Features and Development of Coot*. Acta Crystallographica D66, 486-501) 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 a phenylalanine ('PHE') and the number '63' is the 63rd amino acid in the protein's amino acid sequence, which in this case totals 181 'amino acid residues'; the entry '1h91', coordinates and processed diffraction structure factors, is publicly accessible at

http://www.rcsb.org/pdb/explore.do?structureId=1h91


Coot
File Edit Calculate Draw Measures Validate HID About Extensions
Reset View
Display Manager
R/RC
Map
CE1/1/3 PHE/A

(mol. no: 0) CE1/1/A/63 PHE occ: 1.00 bf: 15.70 ele: C pos: (6.76 37.93 -1.63)

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.

3. Now adding 'animation via live molecules in Acta Crystallographica F Structural Biology and Protein Crystallisation Communications

When the journal Acta Crystallographica F was launched in 2005 (Acta Crystallographica *D* was launched in 1993) it was in recognition of the need for a fast and streamline communications journal for structural biology and crystallization results.

In 2008 the Editors introduced as a standard feature a full animation of the protein, or nucleic acid, structure. The journal being all electronic was free of the print-on-paper limitations. The Editors remarked:-

" *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 (and not simply as supplementary files).*"

" *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*."

[H. Einspahr and M. Guss A new service for preparing enhanced figures in IUCr journals *Acta Cryst.* (2008). F64**, 154-155 [doi:10.1107/S1744309108005617].**

Brian to add a movie clip of a rotating molecule from Acta F here. I suggest the one of van Meervelt, which shows nucleic acid ie instead of a protein. ie *Acta Cryst.* (2010). F**66**, 1028-1031 [doi:10.1107/S1744309110031696]:-

Comparison between the orthorhombic and tetragonal forms of the heptamer sequence d[GCG(xT)GCG]/d(CGCACGC)

K. Robeyns, P. Herdewijn and L. Van Meervelt

PDB reference: 3lln

4. Limitations and the need to educate some readers

What are the limitations for readers to 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 of an article's words should approach any model on which those words are based with a critical appraisal. How is that achieved? Well we have shown how this is achieved via access to the processed diffraction data.

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? Actually, no, unfortunately. Thus effective communication, training and education assume a very important role; the Protein Data Bank and the IUCr take these aspects very seriously, providing active programs to train and/or for outreach.

5. Summary and next steps

So, overall, through being available with the printed journal, in the first example above of Acta Crystallographica D, the associated protein model can be brought alive by the reader via appropriate molecular graphics software. Furthermore the actual electron density is available by direct calculation via the processed diffraction data, again deposited with the Protein Data Bank linked to the publication in the, quite separate, journal. Then, as we further illustrated with Acta Crystallographica F a routine provision of an animation of the protein structure is provided 'up front'.

As next steps, in stimulating further discussion how we could improve and progress further, this leads us onto 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 improving. Article 2 describes then the ramifications of this along with the opportunities for the continuous improvement of macromolecular crystal structures that are presented to our wider research community by 'access to data'.

Article 2

Continuous improvement of macromolecular crystal structures

Thomas C. Terwilliger

Los Alamos National Laboratory, USA; Member of the IUCr Commission on Biological Macromolecules

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 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. 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 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.

In most circumstances the three-dimensional model of a macromolecule is the key product of a crystal structure determination. 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 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. Due to the limited amount of radiation the crystals can withstand, there are significant uncertainties in measurement. Further, during each of the steps in determining structures of macromolecules 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 models obtained from this technique typically do not fully explain the X-ray diffraction data, presumably because the features that are included in the models do not represent everything that is present in the crystals.

Due 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 conformation 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 lower). 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 sequence of a protein or 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 70,000 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. 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.

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. 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. 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 is 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 through the PDB 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 are 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 of a publication would then have to refer to the exact structure that the authors used at the time if they wished 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 may be invested in its correctness and completeness because they have done all the work necessary to determine the structure and have deposited and published it. They 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 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 has 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) 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 used and not to actual differences in the crystal structures.

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 that are 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, then there are likely to be many small differences between the resulting structures that are simply due 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 would be 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 advancements in structure determination methodology, carrying out the production of protein, crystallization, and 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 phi-psi-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 A- 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 is used (e.g., TLS, solvent representation)?
3. What general approaches were used to determine the model (molecular replacement, SAD 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 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 to John Helliwell and Brian McMahon of the IUCr Diffraction Data Deposition Working Group for extensive discussion of the issues presented in this work.

Article 3

The Living Publication 3: Should the crystallographic community require, rather than only encourage, the archiving of the raw (unprocessed) diffraction data measured from a crystal, a fibre or a solution?

John R Helliwell and Brian McMahon

1. Introduction

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/). 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.

The debate within our 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.

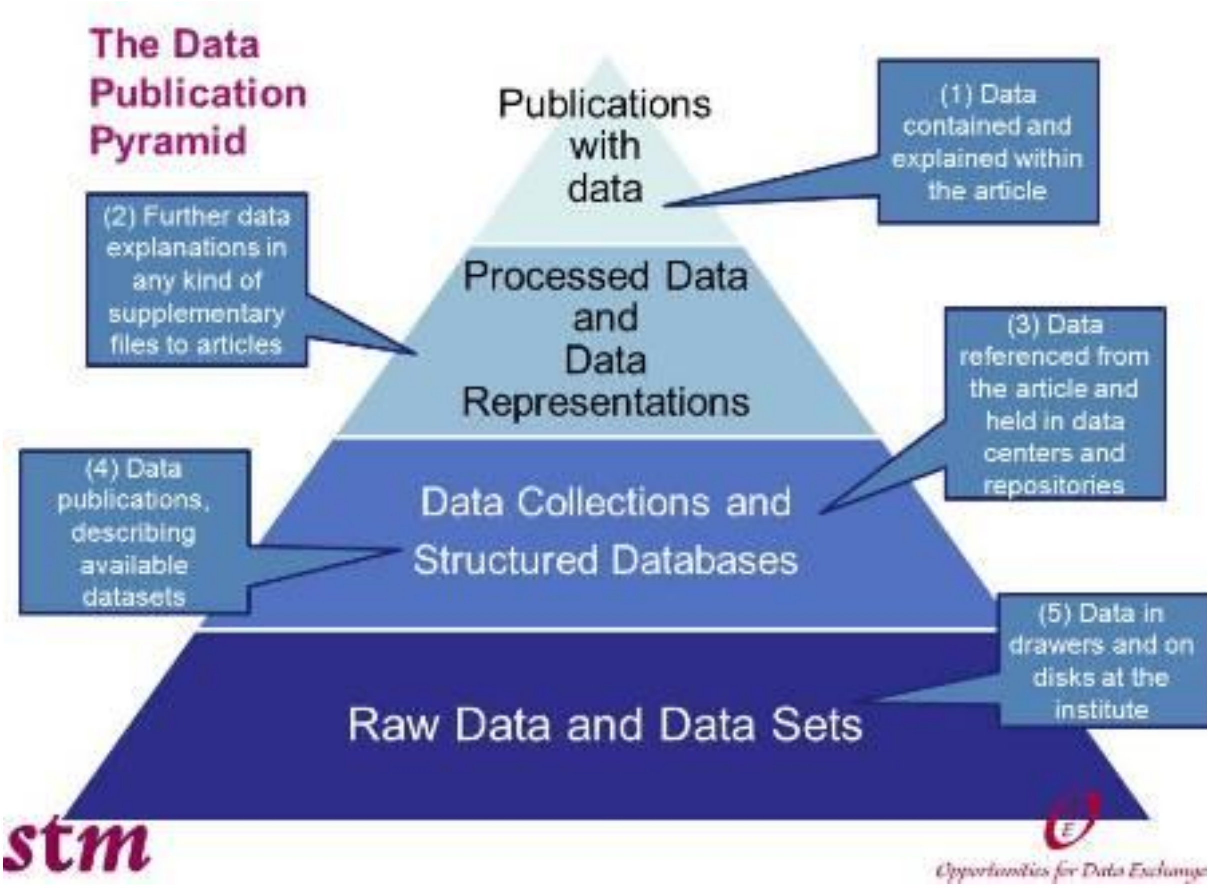
In parallel developments, sensitivities to avoiding research malpractice are encouraging Universities to establish their own data repositories for research and academic staff. These various 'raw data archives', would complement the existing processed data archives. These archives could however most likely have gaps in their global 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 here in our third and final article on 'The Living Publication'.

2. Let's now define types of data more formally

In article 1 we described the current access to and use of processed and derived crystallographic data. We glossed over the fact that the means by which the protein structure was solved, involving several diffraction data sets eg measured at several X-ray wavelengths at a synchrotron, do not require deposition of those processed diffraction data. Rather the deposited data is that from the final protein model refinement step. At this point the reader of this article is encountering the 'data pyramid' (http://www.stm-assoc.org/integration-of-data-and-publications) Figure 1. This crystallographic example actually does not quite fit the labels in the generic figure 1 as this particular group of processed data sets (from several X-ray wavelength) are not deposited but 'reside on CDs, DVDs or external disk drives in a researcher's desk draw' and are not attached to the publication.

Figure 1 The data pyramid; this applies to all scientific fields that work in their research with 'data'. From http://www.stm-assoc.org/integration-of-data-and-publications


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

So, at this point we had better define the different types of our data! To try to be definitive we follow community terms and nomenclature whereby we turn to the IUCr Journals Notes for Authors for biological macromolecular structures, and thus where we have our ie macromolecular crystallography's version of the 'data pyramid'. Thus in crystallography we have:-

2. 1 Derived data

• Atomic coordinates, anisotropic or isotropic displacement parameters, space group information, secondary structure and information about biological functionality must be deposited with 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 approximately hundreds of kbytes in filesize.

• Relevant experimental parameters, unit-cell dimensions are required as an integral part of article submission and are published within the article.

2.2 Processed experimental data

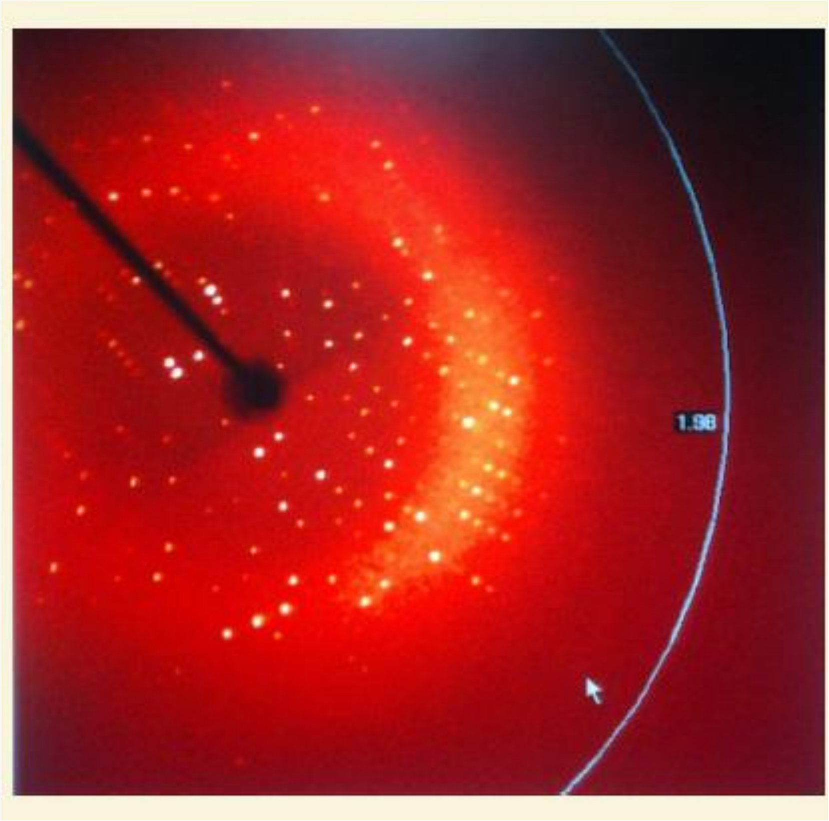
• Structure factors, which must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code. These total typically approximately several Mbyes in filesize.

2.3 Primary experimental data

* The raw diffraction images (see eg figure 2 where one such 'diffraction image' is one of

hundreds or even thousands that are measured from each crystal; these in total would be approximately 1 Gbyte in size.)

Figure 2 An example protein crystal 'diffraction image', and which is one of hundreds or even thousands that are measured from each crystal:-



3. Current IUCr Journal policies re raw data

For small-unit-cell crystal/molecular structures and macromolecular structures IUCr journals have no current, binding, policy regarding publication of diffraction images or similar raw data entities. However, the journals welcome efforts made to preserve and provide primary experimental data sets. Authors ***are indeed 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.***

In more specialised 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.

4. Important principles and standards of data deposition

The IUCr is very enthusiastic to help define standards of data deposition. The reasons are quite general ie apply to all science fields:-

- To enhance the reproducibility of a scientific experiment
- To verify or support the validity of deductions from an experiment
- To safeguard against error
- To better safeguard against fraud than is apparently the case at present
- To allow other scholars to conduct further research based on experiments already conducted
- To allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
- To provide example materials for teaching and learning
- To provide long-term preservation of experimental results and future access to them
- To permit systematic collection for comparative studies

5. Complying with Funding Agencies

Also, it is worth restating, that 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 e.g. the Research Councils UK Common Principles on Data Policy (http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx) 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). See also http://forums.iucr.org/viewtopic.php?f=21&t=58 for discussion on policies relevant to crystallography in other countries. Nb these policies extend over derived, processed and raw data, ie without really an adequate clarity of policy from one to the other stages of the 'data pyramid' ((see above namely:- http://www.stm-assoc.org/integration-of-data-and-publications).

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

We now come to several linked, central, questions:-
When might derived and/or processed diffraction data become inadequate? ie When might raw data become valuable?How often might this be the case? What is the cost and benefit of retaining and having access to raw diffraction data?

Firstly the processed diffraction data are the 'diffraction structure amplitudes' from the diffraction spots in Figure 2. So, what do we perhaps ignore between the spots? Figure 3 from [*J. Appl. Cryst.* (2008). **41**, 659 [doi:10.1107/S0021889808008832]

Of crystals, structure factors and diffraction images

L. Jovine, E. Morgunova and R. Ladenstein] shows an example of what lies between the spots.

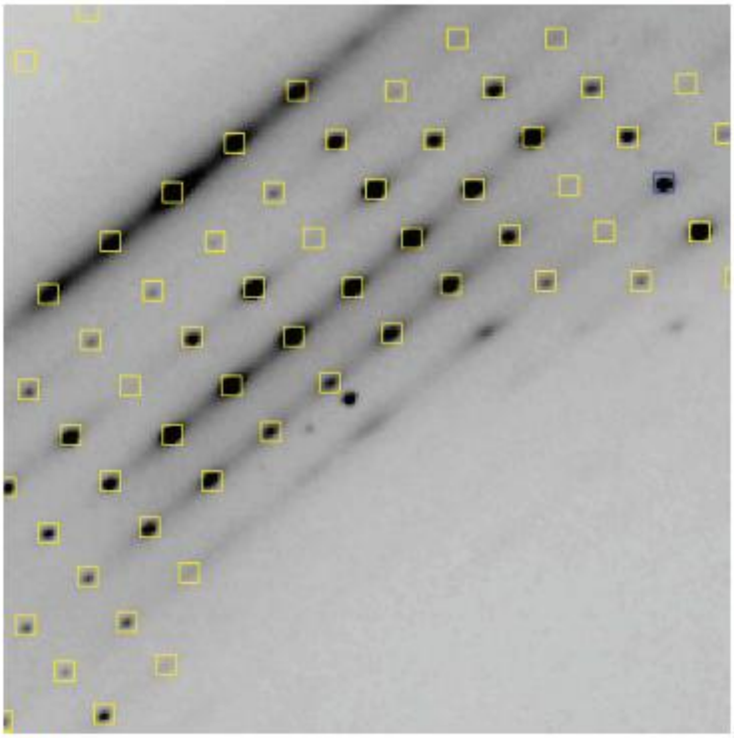


Figure 3 Example of strong diffuse scattering from an RNA crystal [Jovine et al 2008].

These data are not always ignored; see Figure 4 [eg *Acta Cryst.* (2011). B**67**, 516-524 doi:10.1107/S0108768111037542 Diffuse scattering resulting from macromolecular frustration

T. R. Welberry, A. P. Heerdegen, D. C. Goldstone and I. A. Taylor]

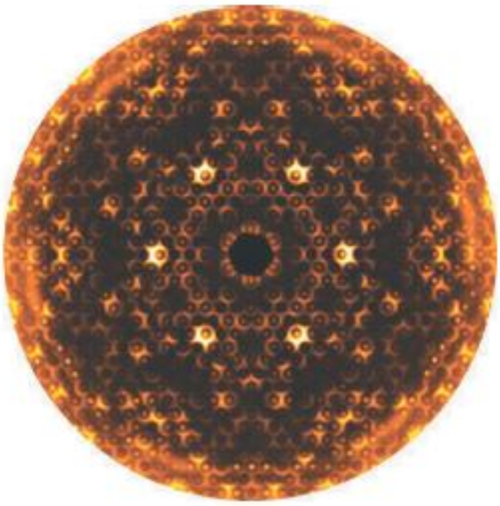


Figure 4 A predicted reciprocal lattice section (diffraction pattern in effect) normal to the unit cell c axis and from the crystal is asserted to suffer from 'molecular frustration' in its crystal packing layout [Welberry et al 2011].

However such diffuse features are not routinely straightforward to interpret at present ie the example above of Welberry et al 2011 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 I.D. Glover, G.W. Harris, J.R. Helliwell and D.S. Moss 'The variety of X-ray diffuse scattering from macromolecular crystals and its respective components' Acta Cryst. (1991) B47, 960-968.).

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 ie which 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 $I2_13$. [S.J. Harrop, J.R. Helliwell, T. Wan, A.J. Kalb (Gilboa), Liang Tong, and J. Yariv "Structure solution of a cubic crystal of concanavalin A complexed with methyl alpha–D–glucopyranoside" (1996) Acta Cryst. D52, 143–155]. Our own thoroughness avoided a calamity of incorrect structure determination. Various validation checks, these days, seek to help avoid such calamities in an ever expanding perhaps less experienced community of researchers. The availability of processed ***but unmerged*** processed diffraction data would allow more checking by readers than at present. Indeed a recommendation made by a recently PDB convened and reporting task force on validation at the PDB has recommended preserving unmerged processed diffraction data (See Read et al Structure 19, 1395–1412, October 12, 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', lets say the predominant one, and not the other(s); preservation of just that one crystal lattice processed diffraction data obviously does not include the others, and which are lost 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 Lab. 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 the category a bit like 'negative results' and which could 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 cut 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 (sigmas) (ie <I/sig(I)>) decrease below 2.0. The community is keen though to not 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 of course occur well beyond the obvious pattern edge. These occasional spots are known about but are deemed rare and thereby 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 then why the raw diffraction data images are of interest to be archived, with a doi registration, so as to be linked to the relevant publication and, in most cases, the primary PDB deposition files highlighted in the analyses of the publication.

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 prove achievable to fabricate raw diffraction data too, ultimately.

So there are certainly five, perhaps 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? These lead naturally to a cost to benefit analysis:-

7. Cost benefit analyses

Costs could be minimised by preserving each raw diffraction data set at the local centre (synchrotron facility, neutron research centre or university) thus avoiding quite large network file-transfer costs. Also costs could be cut by preserving only a proportion of each of the raw data sets or by making some form of data set compression (lossy or lossless; James Holton pers comm) or both.

The benefits could be maximised 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 eg because the diffraction pattern showed extensive diffuse scattering, currently ignored, or showed multiple crystal lattices, of which detailed on just one was made 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 still under active debate within the crystallographic community and so we draw your attention to the IUCr Forum on such matters at:-
http://forums.iucr.org/

Within this Forum you can find for example the ICSU convened Strategic Coordinating Committee on Information and Data fairly recent report; within this we learn of many other areas of science efforts on data archiving and eg that the radio astronomy square kilometre array will pose the biggest raw data archiving challenge on the planet.[Our needs as crystallographers are thereby relatively modest.]

We hope you have been stimulated by, and even enjoyed, this trilogy of articles of what we do in crystallography in managing our literature along with our data within which we have coined the term 'The Living Publication' for the purposes of this ICSTI Insight group of three articles.

Acknowledgements

We are grateful to the IUCr Diffraction Data Deposition Working Group (DDDWG) Members and associated consultants, for their conscientious work on these matters and all 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 held in Madrid.