Information Services & Use 30 (2010) 31–37
DOI 10.3233/ISU-2010-0609
IOS Press

The record of experimental science: Archiving data with literature *

John R. Helliwell [a,**] and Brian McMahon [b]
[a] *University of Manchester, Manchester, UK*
[b] *International Union of Crystallography, Chester, UK*

Abstract. Crystallography is presented as a case study of a scientific discipline where the experimental data that underpin research results can be integrated into the scientific record. Among other advantages, this maximises the degree of trust in science, since published results can thereby always be validated independently.

Keywords: Data validation, digital archiving, electronic publishing

1. Introduction

Contributors to the International Council on Scientific and Technical Information (ICSTI) Workshop on Interactive Publications and the Record of Science [14] were invited to consider a number of questions relevant to the digital archiving of interactive content. Since many interactive features of online research publications involve manipulation or visualization of data sets, the questions are equally relevant to the general challenge of digital archiving of research data associated with publications. Crystallography is a field that has long experience of associating data and publications, and leading journals have developed robust practices for publishing derived and reduced experimental data sets in support of research publications. There is growing pressure to extend these practices to accommodate primary diffraction images (reflecting practice in certain other disciplines regarding primary rather than processed data), and so we suggest that crystallography is a valuable discipline for highlighting and investigating some of the remaining issues.

2. The importance of data for publication

Many recent studies at national and regional level, e.g., [11,17] are raising the level of consciousness amongst legislators and policy makers of the importance of retaining data sets within the record of science. Among the reasons that we would identify as particularly important for science are: to enhance the reproducibility of a scientific experiment; to verify or support the validity of deductions from an experiment; to safeguard against error and against fraud; to allow other scholars to conduct further research based on experiments already conducted; to allow reanalysis at a later date, especially to extract 'new'

* Paper presented at the ICSTI Interactive Publications Conference 2010.

** Corresponding author: John R. Helliwell, Department of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK. Tel.: +44 161 275 4970; Fax: +44 161 275 4598; E-mail: john.helliwell@manchester.ac.uk.

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; and to permit systematic collection for comparative studies. In summary, and perhaps most simply, the purpose of providing access to the data is to enhance the reader's experience in understanding a piece of research work.

The increased interactivity of the scientific literature that was discussed in the ICSTI Workshop is yet another corollary of the integration of the underlying scientific data into the publication lifecycle. The availability of such data allows crystallographers, for example, to develop interactive three-dimensional molecular models for structural chemistry and biology; these models now form live enhanced figures in some journals (e.g., the IUCr's *Acta Crystallographica Sections E* and *F*). Such data may also be used by journal publishers to overlay additional semantic content on the research article, through links from the article directly to database entries, online dictionaries, product catalogues and other relevant resources.

In IUCr journals dealing with the structures of inorganic materials, small organic or metal-organic compounds, the structural data models discussed in a research article are always deposited with the journal, and are made openly available as supporting materials. Here, 'openness' means that they are freely available, even to non-subscribers. That is, for journals published by the IUCr under a subscription model, access to the text of research articles must be paid for; but supporting materials including the structural models can be downloaded free of charge. That has always been a principle of the IUCr's publication strategy, and such data used to be provided as photocopies of computer printouts of the data. Nowadays, the online medium allows immediate access to the numerical data in machine-readable form, so that they can be used directly for further research. Indeed, the IUCr has an active interest in open-access publication as a means for maximising access to scientific publications generally, and recently has made its short structural report journal *Acta Crystallographica Section E* fully open access (authors pay a publication fee to fund such open access). Its other titles allow some open access under a 'hybrid' model of subscription with individual open-access articles funded by the authors. Further progress in these directions will depend upon the development of sustainable economic models, and is not dictated by a particular publishing model doctrine.

However, such 'openness' also implies access to the data for the purposes of validation, and this has also been an important driver in the development of IUCr publishing strategy. Structural data sets deposited in support of a publication are mechanically analysed for logical consistency and chemical reasonableness, and the results of such analyses are provided to journal referees as a routine part of the peer review process [15]. These procedures have significantly reduced the number of erroneous structures published in IUCr journals over the course of the past several years. Unfortunately, recent events have brought to light the publication of some incorrect structures where circumstances other than human error have been involved [9]. Detection of these involved post-publication analysis of the associated 'structure factors' – processed experimental data files that the journal also requires authors to deposit. In consequence, such analysis of the experimental data alongside the derived structural data has now become routine for these journals. It is a sobering thought that most other journals reporting crystal structures do not currently require authors to deposit their structure factors.

In the case of such 'small molecule' structures, each deposited data set is assigned its own digital object identifier (DOI), and links between the data sets and the parent publication are made explicit in the article metadata deposited with the CrossRef DOI registration agency. In these metadata depositions, the data sets are described as components of a composite article.

For biological macromolecular structures, the crystallography community has different, though not so dissimilar, procedures. Protein and nucleic acid structures (from crystallography and other techniques such as NMR and electron microscopy) are deposited with an autonomous databank, the Protein Data

Bank (PDB). Many journals insist that a structure be deposited with the Protein Data Bank before they will accept an article based on that structure. The PDB also routinely collects structure factors, mandatory since 2008 (this is not currently the case for established databases of small molecule structures) and (since 2009) runs validation procedures analogous to those for small molecule structures of the IUCr journals.

Having been established for several decades the PDB recently undertook a remediation of existing PDB entries [3], which improved the uniformity of standards expected. Joosten et al. [12] also undertook a study of automated re-refinement of existing structure models in the PDB using grid computing techniques. Their large-scale benchmark of 16,807 PDB entries showed improvements of fit to the deposited experimental X-ray data and of geometric quality. This emphasises the scientific benefits of access to the processed diffraction data. The resulting structure models have also been made available independently through the PDB_REDO databank (http://www.cmbi.ru.nl/pdb_redo/). From this, it can be argued that macromolecule model refinement checks should become part of the process of acceptance for publication (in the spirit of 'prevention of errors rather than cure'). A key difference now between chemical and biological article submissions is that referees of biological structures may not have access to coordinate or diffraction data, although they may have access to the PDB validation report. A publishing model that performed macromolecule model refinement checks would parallel that for small-molecule crystal structure publications in IUCr journals but could entail substantial extra work for Editors and referees.

Each structure deposited in the PDB has its own identifier code, but it also has a DOI registered with CrossRef. In principle, therefore, this allows component data sets to be associated with parent articles through DOI linking. In practice, CrossRef records will not explicitly reveal that these can be considered parts of a composite article (because the metadata for article and data set in this case have been supplied independently by separate publishers). Indirect methods of course exist for tracing such relationships: the PDB file 'header' carries bibliographic details of the relevant publication; but the use of a standard such as DOI would be easily extensible to other disciplines.

Examples such as these provide an encouraging model for linking data sets with publications through DOIs, which is also a goal of the new DataCite organisation founded in 2010 to manage the citation and archiving of research data (http://www.datacite.org). We suggest that further such developments would benefit from the adoption of a standard schema for describing composite documents (including research article, component figures and tables, associated data sets and other supplementary materials) that allows individual components to be hosted on different platforms. This would of course entail publishers developing the idea of a distributed composite article in concert with data storage managers.

3. Scaling up in crystallography

In the existing model of crystallographic data publishing described in the preceding section, the data sets themselves are rarely more than a few megabytes in size (sometimes only a few kilobytes) – minuscule by the standards of current large-scale storage systems. However, there are strong grounds for considering the archiving of the raw experimental data sets (for crystallography these will take the form of a collection of diffraction images often measured at a synchrotron X-ray beamline or neutron source instrument). These are much larger in size. For a protein crystal structure, the final structural information, in the form of tabulated atomic positional coordinates, may be 250 kB in size. This may be derived from a structure-factor file of about a megabyte; but a typical set of diffraction images contributing to this result can easily be a thousand times larger – a gigabyte or more. The routine management of such

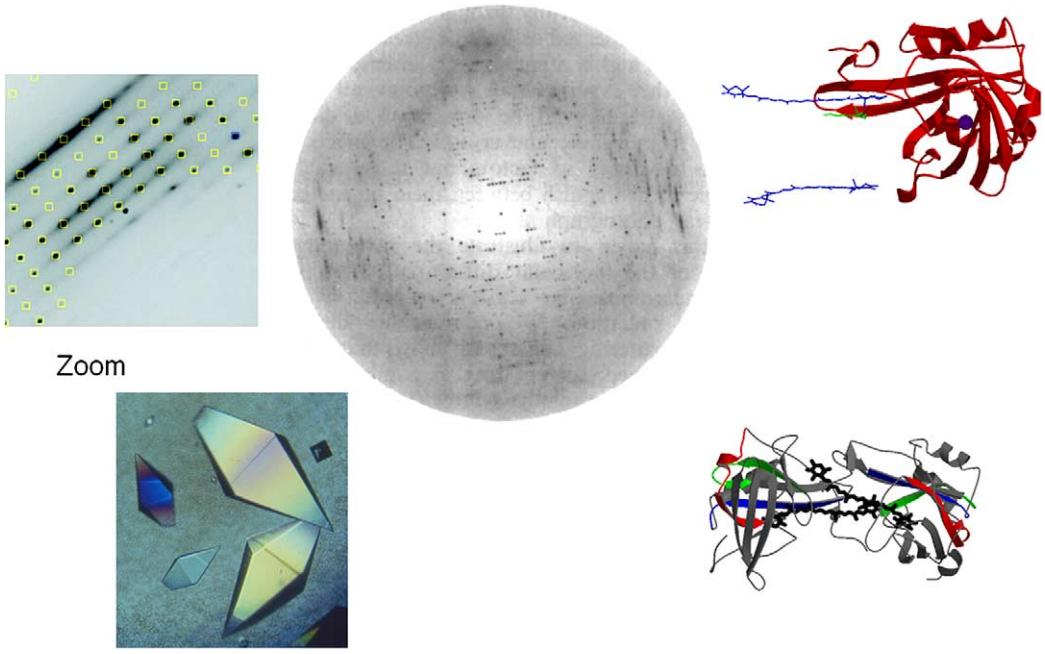

Fig. 1. A representative set of macromolecular crystallography crystals, diffraction patterns (including diffuse smears as well as well defined spots) and protein structures represented in ribbon form. Top left: Example of strong X-ray diffuse scattering from a trigonal crystal of 4.5S RNA domain IV (PDB ID code 1DUH). Although the significant intensity of the regions between Bragg spots (circled) contains information about crystal disorder, this is not taken into account by current standard data processing software (from [13] with permission of the authors and IUCr). Top middle: t-RNA(met) 1.5° oscillation *diffraction pattern image*, wavelength 1.54 Å, also showing diffuse scattering; 240 such images at successively rotated crystal orientations would comprise a, full revolution, diffraction data set (from [7] with permission of the authors and IUCr). Top right: A movie snapshot of β-crustacyanin (see Supplementary file deposited with [4]; with permission of the authors and IUCr). Bottom right: Ribbon representation of β-crustacyanin solved by protein X-ray crystallography (from [5]; copyright 2002 National Academy of Sciences, DC, USA). The molecular weight of β-crustacyanin is ∼40,000 Daltons, comprising approximately 3000 atoms (excluding hydrogens); within the protein dimer two astaxanthin carotenoid 'small molecule' ligands can be seen, which are bound to the protein, and whose structure has also been studied by X-ray crystallography in a non-protein bound crystal of astaxanthin. Bottom left: Crystals of thaumatin protein showing beautifully formed morphology grown by Prof. Naomi Chayen, Imperial College, London (with permission of the author).

large files in the context of the laboratory or publishing house is still quite challenging, and solutions for the practical problems that arise may need to be sought outside of the traditional researcher/author–publisher relationship. Some examples of the types of diffraction and molecular structure data that we are talking about are shown in Fig. 1.

For the integrity of the full scientific record, the archiving of whole diffraction images would be preferable, as no data-processing software interpretation is involved, and the way then remains open for subsequent reprocessing and reinterpretation if thought desirable. Despite ever-improving techniques within the science of crystallography, the chances of misinterpretation can be serious and can include mis-indexing of the diffraction spots, incorrect assignment of the crystal space-group symmetry (including the hand or chirality), not recognising crystal twinning and over-optimistic diffraction data resolution. Perhaps the probability of such errors will increase with larger-scale distributed scientific endeavours such as high-throughput structural genomics projects (http://www.isgo.org/), although so far these enterprises have an excellent track record in the quality of their structure determinations, and they have produced at least one pioneering archive of diffraction data sets (see

http://www.jcsg.org/prod/scripts/technologies1.html#downloads). As well as retaining the diffraction-spot intensities used for the structure analysis, archiving data images also preserves the record of diffuse scattering (i.e., information between the diffraction or Bragg spots). Currently most of this information is not susceptible of analysis; but in future we expect to be able to fully interpret such scattered intensity data between Bragg spots in terms of protein dynamics, a very important factor in any consideration of biological function based on structure.

There have also been recent well-publicised discoveries of malpractice in the reporting of some ten or so protein structures [6]. Analysis of the diffraction structure factors and coordinates revealed unusual patterns; i.e., in these cases the diffraction data images were not necessary to do this. For some of the protein structures in this case, structure factors were not deposited with the PDB. However, even structure factors can be fabricated, and the archiving of diffraction images would provide a higher level of security against such fraudulent practices. These cases should not be taken as indicative of widespread fraud in structural science – the actual number of structures involved is minute in comparison with the half million organic structures curated in the Cambridge Structural Database and the 64,000 macromolecular structures in the Protein Data Bank. Nonetheless, the routine availability of primary experimental data would help to encourage the highest levels of trust in the scientific method, as well as contributing in all the other ways described above.

Within the crystallographic community, reliance on deposited diffraction images as the primary raw data archive has been championed, for example, by Jovine et al. [13] and the Editors of *Acta Crystallographica Sections D* and *F* [2], who recently stated: 'we believe that, sooner rather than later, deposition of (diffraction) images will be treated as naturally and routinely as deposition of coordinates is treated now. To us, it is a question then of not whether, but when'. So, in principle, researchers in our community (in the current research paradigm, these are the primary *risk holders*) would prefer to archive primary data for some or all of these reasons; but are their Funders concerned enough to fund this? Properly archived data sets need to be held in multiple redundancy, further increasing storage and management costs. The projected costs are currently greater than existing individual service providers such as the Protein Data Bank are able to shoulder, and there is little commitment (especially of resources) from national research councils or other sponsoring bodies. Researchers currently look after their own data for a few (typically about 5) years, but there is no immediate prospect of a centralised long-term curated archive for diffraction images. A pragmatic "next step" compromise would be to store unmerged, but still processed, diffraction intensities rather than fully merged structure factors as is currently the practice.

Could the answer lie in developing federated data archives at the laboratory level, or at experimental synchrotron X-ray and neutron central facilities, which heavily serve crystallography researchers, and where many (but by no means all) of the experiments are conducted? A pilot federation has been developed within a number of Australian research institutions [1], utilizing open-source software and a metadata model consistent with one under development by the UK Science and Technology Facilities Council (STFC). This effort is progressing and now is moving from Fedora towards TARDIS and MyTARDIS protocols so as to download multiple diffraction data images at a single click (Buckle pers. comm.). Data sets and associated metadata from federated repositories are given a unique and persistent handle, providing a simple mechanism for search and retrieval *via* web interfaces. The new UK national synchrotron facility Diamond (operational since late 2007) was asked to state its policy on retaining diffraction images, and responded [16] that no data collected had yet been deleted; that the aim was to retain data sets for as long as practical; that archiving would be done at the Atlas Data Centre of the Rutherford Appleton Laboratory (a UK facility designed to operate as a CERN Tier 1 storage node); and that the current policy (under review) is to guarantee data preservation for a minimum of 6 months. The

eCrystals Project, led by the UK National Crystallography Service at the University of Southampton [10], also uses institutional repository software to manage collections of solved structures, with primary diffraction images stored at the Atlas Centre. eCrystals is planning to acquire additional federated partners. While not a publishing platform, it also uses DOIs to provide persistent identifiers for structural data sets and to allow interlinking with derived research publications.

If this does prove to be the way forward for data archiving, an information infrastructure would need to be built to link federated repositories. Components of such an infrastructure would include distributed search and discovery tools, distributed redundant copying, preservation protocols for online data, provenance tracking and integrity checking, all of which could usefully be modelled on information structures developed for scholarly publishing.

4. The vision of archiving data with literature

We have tried to demonstrate that the journals published by the IUCr have consistently linked scientific literature with pertinent data sets. Not all journals publishing crystallographic results have developed such procedures to the same extent; and it is clear that even the IUCr journals cannot include all the data underlying a research article in a future where everything can be stored and accessed digitally. Yet we strongly believe that science will benefit from incorporating all the richness of the full experimental output within the scientific record. We therefore identify within our own field the following essential stages in working towards the vision of including data with literature, both properly validated and reviewed, within the authoritative record of scientific discovery and invention:

- increased willingness by all journals to deposit structural data and structure factors;
- increased willingness by structural databases to store and curate structural data, structure factors and perhaps primary data;
- increased collaboration between journals, structural databases and other institutions that archive data sets;
- assignment of persistent identifiers according to standard protocols (e.g., DOIs) to unpublished data held in:
  - domain-specific repositories (e.g., eCrystals Southampton);
  - institutional repositories;
  - synchrotron, neutron and other experimental facilities (e.g., new and upcoming X-ray laser sources);
  - image data stores (e.g., Atlas, Rutherford Appleton Labs);
- community codes of practice for storing, managing, archiving and accessing research data sets;
- standard ‘compound document’ descriptions to link data and publications.

In many respects crystallography is very fortunate in addressing these challenges because there is a high degree of homogeneity within the type of experiments and analysis for structure determination of different types of material and using different probes (X-rays, neutrons, electrons). A relatively small community ($\sim$11,000 people are listed in the *World Directory of Crystallographers*) shares knowledge, software and practical experience; and a widely accepted standard for information exchange and archive has been in place for almost two decades [8]. Nevertheless, we see that many other communities are addressing the same challenges, so that there is much to be gained by exchanging experience and best

practice across disciplinary boundaries. Sharing information handling approaches between the once separate worlds of literature and data management is beginning to bear fruit in the richly linked electronic publishing environment of the 21st century, by facilitating data visualization, search and access, and reuse, direct from the published literature.

References

[1] S. Androulakis, J. Schmidberger, M.A. Bate, R. DeGori, A. Beitz, C. Keong, B. Cameron, S. McGowan, C.J. Porter, A. Harrison, J. Hunter, J. L. Martin, B. Kobe, R.C.J. Dobson, M.W. Parker, J.C. Whisstock, J. Gray, A. Treloar, D. Groenewegen, N. Dickson and A.M. Buckle, Federated repositories of X-ray diffraction images, *Acta Crystallographica* **D64** (2008), 810–814.

[2] E.N. Baker, Z. Dauter, M. Guss and H. Einspahr, Deposition of diffraction images to be discussed at the Open Meeting of the Commission on Biological Macromolecules of the IUCr in Osaka, *Acta Crystallographica* **D64** (2008), 337–338.

[3] H.M. Berman, K. Henrick, H. Nakamura and J.L. Markley, The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data, *Nucleic Acids Research* **35** (2007), 301–333.

[4] N. Chayen, M. Cianci, J.G. Grossmann, J. Habash, J.R. Helliwell, G.A. Nneji, J. Raftery, P.J. Rizkallah and P.F. Zagalsky, Unravelling the structural chemistry of the colouration mechanism in lobster shell, *Acta Crystallographica* **D59** (2003), 2072–2082.

[5] M. Cianci, P.J. Rizkallah, A. Olczak, J. Raftery, N.E. Chayen, P.F. Zagalsky and J.R. Helliwell, The molecular basis of the coloration mechanism in lobster shell: β-crustacyanin at 3.2 A resolution, *Proceedings of the National Academy of Sciences* **99** (2002), 9795–9800.

[6] Z. Dauter and E.N. Baker, Black sheep among the flock of protein structures, *Acta Crystallographica* **D66** (2010), 1.

[7] 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 Crystallographica* **B47** (1991), 960–968.

[8] S.R. Hall, F.H. Allen and I.D. Brown, The Crystallographic Information File (CIF): a new standard archive file for crystallography, *Acta Crystallographica* **A47** (1991), 655–685.

[9] W.T.A. Harrison, J. Simpson and M. Weil, Editorial, *Acta Crystallographica* **E66** (2010), e1–e2.

[10] M.B. Hursthouse, S.J. Coles, J.G. Frey, L. Carr, C. Gutteridge, L. Lyon, R. Heery, M. Duke and M. Day, ECRYSTALS(CHEM.SOTON.AC.UK): open archive publication of crystal structure data, *Acta Crystallographica* **A61** (2005), c481–c482.

[11] e-IRG Data Management Task Force, Report on data management (2009), available at http://www.e-irg.

[12] R.P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A.-C. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A.L. Da Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I.J. Tickle and G. Vriend, PDB_REDO: automated re-refinement of X-ray structure models in the PDB, *Journal of Applied Crystallography* **42** (2009), 376–384.

[13] L. Jovine, E. Morgunova and R. Ladenstein, Of crystals, structure factors and diffraction images, *Journal of Applied Crystallography* **41** (2008), 659.

[14] B. McMahon, Interactive publications and the record of science, *Information Services and Use* (2010), DOI: 10.3233/ISU-2010-0607.

[15] P.R. Strickland, M.A. Hoyland and B. McMahon, Small-molecule crystal structure publication using CIF, *International Tables for Crystallography* **G** (2006), 557–569.

[16] D. Stuart, Personal communication on behalf of diamond light source, Division of Structural Biology and The Oxford Protein Production Facility, Oxford OX3 7BN, UK, 2010.

[17] UK Research Data Service, The data imperative: Managing the UK's research data for future use, 2009, available at: http://www.ukrds.ac.uk/resources/download/id/14.