Are We Nearly There Yet?

A Perspective on Data Sharing in (Chemical) Crystallography

by Simon Coles

When contemplating the subject of sharing knowledge in academia, I am particularly drawn to the following statements in a monologue on the "The Social Function of Science": The present method of publication...has the following deficiencies:

- Research results cannot be reported promptly.
- Research results cannot be reported completely, with all necessary data, illustrations, discussion, background, and other pertinent details.
- Waste is widespread, in that only a small number of subscribers are interested in any particular research report as now published.
- Increasing volume of the scientific literature is increasing the financial burden upon individual scientific workers and upon institutions.
- The multiplication of journals...Failure of libraries to subscribe to journals due to measures of economy will result in serious hindrances to the availability of requisite scientific literature.
- Editing and business management of journals is usually undertaken on an unpaid volunteer basis by scientists.
- A scientific worker can only read a small fraction of the papers.

I am not stating anything here that most readers don't already know—these issues have been raised by various arms of academia for at least the last decade and are the subject of much concern and debate. What is particularly striking, however, is that "The Social Function of Science" [1] was written by J.D. Bernal in 1939!

Visionary Beginnings...

Bernal was a crystallographer and an academic visionary, and he set out to address these (and many more interdependent) matters. Accordingly, the crystallographic community has gone on from Bernal's basis to lead the way in terms of the organisation of science. However, in some respects we actually find ourselves back in the same position despite moving forwards radically in the last 75 years. This article will reflect on how the crystallographic community, and in particular the chemical arm of the discipline, has structured its communication of results, leading up to the issues of the current times and how they are being addressed.

Back in 1939, Bernal also made two other remarks that are of particular significance to the topic of this article. Firstly, that "Science has moved from direct observation of nature and is more and more dependent on the previous observations of other workers and on their methods of interpretation." It became apparent over the following decades that crystallography is fundamentally a subject that is about the data, *i.e.*, a crystal structure, and that discussions, interpretations, and scientific insights are all secondary derived material based on that data. Secondly, he discussed at length the need for the reorganisation of science: that research should be carried out for the benefit of society as a whole. This requires the effective coordination of individuals: "The general problem of the organisation of scientific research can be simplified by dividing it into two problems, those of the inner and outer organisation. The first is an internal problem of how a laboratory should be run and the second of how the work of the different laboratories should be co-ordinated into a coherent structure of scientific research."

Many of Bernal's 1939 observations were based on the distribution problem for scientific publications and the lack of coordination of research, however it was during the next 30 years in collaboration with his *protégé*, Olga Kennard, that a route to addressing these problems was realised. In 1965, Kennard founded the Cambridge Structural Database (CSD). [2] What this did was:

- make the data the primary unit of scientific endeavour
- instigate a centralised model, whereby all data were contributed to, and held in, a single place
- provide a single resource for data sharing
- ensure that all data were held in the same form

These achievements, in a single approach, clearly address the majority of concerns raised many years earlier. However, their thinking had subsequently moved beyond these concerns, and in fact the CSD was also conceived in order to generate a collection of data that could facilitate new science in itself. This goal is captured in a comment by Kennard, "We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments." [3]

Figure 1. From top: John Bernal, FRS ("The Sage") and Olga Kennard, FRS: the pioneers of data sharing in crystallography.

In 1965, there were only around 1500 published structures, which were collated manually into printed volumes, named 'Molecular Structures and Dimensions'. However, this number began to rise quite rapidly and it soon became clear that a database approach would be required in the longer term. The CSD evolved as one of the world's first scientific databases, essentially containing bibliographic information and associated crystallographic numerical data.

The success of the CSD grew over time, but the largest changes occurred through the implementation of tools to search the database, which in turn engaged an audience that went far beyond the core crystallographic community. Software that was developed to enable the processing of structural data for incorporation into the CSD quickly evolved into applications for searching the data. The real key to success was developing the software beyond providing search functionality for crystallographers (essentially recall based on the data fields entered). Chemical searching was enabled by indexing the data in new ways. Ultimately, this enabled 2D structure searching and analysis of 3D structural data, both molecular conformation and, more recently, interactions between molecules. This functionality clearly went beyond simply questioning whether a structure had been done before to pioneer new areas of structure-based chemistry research. It drew in the chemistry community, not only by providing usable, intuitive tools that could help them with their research, but also by providing new insights into chemical phenomena. It is worth noting here that other, related crystallography databases have followed suit—the Protein Data Bank (PDB) and the Inorganic Crystal Structure Database (ICSD) are now equally successful in serving their respective communities, albeit with slightly different organisational structures.

Realising the Need for a Standard

By the late 1980's, it was clear that this data-based discipline was moving fast; the sheer scale at which the science was operating clearly required the community to be well organised in order to manage the situation. At this time the CSD had reached somewhere in the region of 100000 records, each of which had to be re-encoded from journal articles and their associated hard copy supplementary information. With this rate rapidly increasing, the manual approach was clearly unsustainable—an obvious fact when we look back from todays 'digital era' at how the CSD has grown. Figure 2 on the following page shows that the 1980's are the lead up to the inflection point.

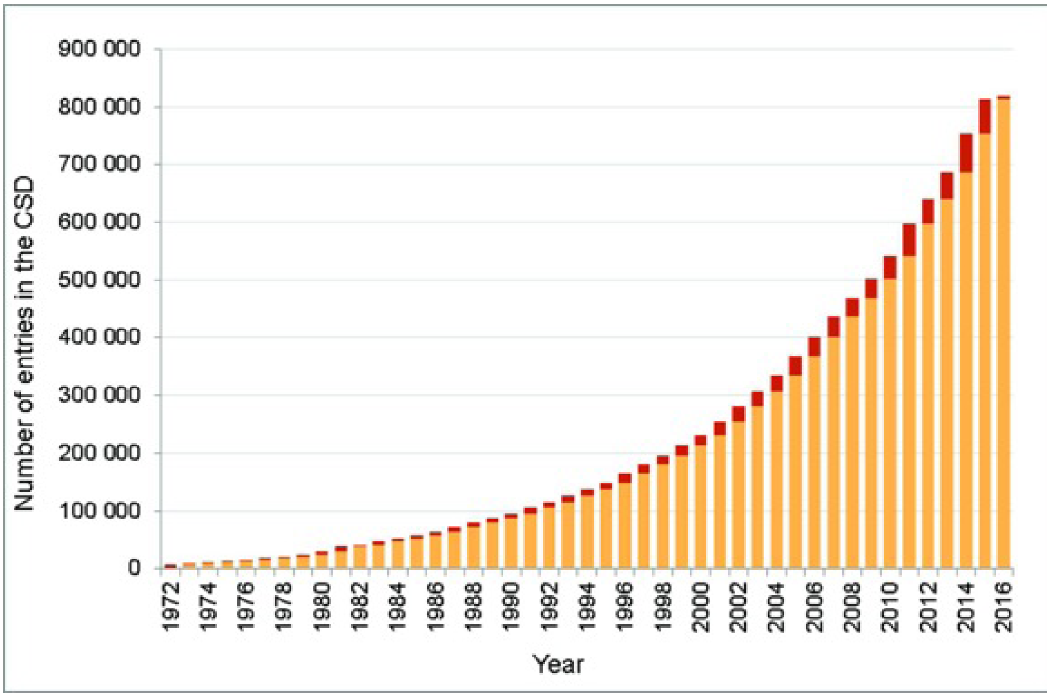


Figure 2. The growth of the CSD over the last 5 decades

With computing becoming more ubiquitous in the laboratory, as well as being available for all to use on demand and in a timely manner, it was not long before the community turned to finding a way in which crystal structures could be more 'computable', *i.e.*, processed by a computer. Accordingly, the Crystallographic Information File (CIF) [4] was adopted as a standard in 1991 by the International Union of Crystallography (IUCr). Although CIF pre-dates the prevalent use of the world-wide web, its ethos and design maps onto it very well and could easily be considered as 'semantic web compliant' today. Critically, it continues to be compatible with evolving standards.

CIF is a file structure developed for the archiving and interchange of crystallographic data and is now ubiquitously used in all aspects of the discipline, whilst enabling integration with other disciplines. CIF can capture the results of a diffraction experiment and take them all the way through to the publication of results, alongside journal articles and/or in data repositories. In fact, the IUCr use CIF as the basis for the submission of crystal structures to their journals and it can therefore be considered as a publishing format in its own right. The dictionary structure of CIF empowers much of this multidisciplinary work, as there is an overarching 'core' into which different dictionaries can be added, such that the same framework can be adopted for macromolecular crystallography as for chemical crystallography or for powder diffraction, for example. Its adoption success is due to its strengths as a general, flexible, and easily extensible free-format file that is both human and machine readable and can be edited by a simple text editor.

However, its success is not just down to providing a great technical solution; community acceptance and coherence is crucial. The IUCr proposed the format in 1991: it was well on the way to being universally adopted within five years and rapidly thereafter became ubiquitous in all aspects of the field. This is also due to exceptional and rigorous governance—the Committee for the Maintenance of the CIF Standard (COMCIFS) [5] was established in 1993 under the auspices of the IUCr and comprises a range of working groups overseen by a management group.

CIF, therefore, has evolved in many ways and provides the underpinning for the whole discipline, transcending being merely a file format. CIF is now known as the Crystallographic Information Framework, and the breadth of support that the CIF provides for the discipline is illustrated in Figure 3. It has been adopted by the CSD (and other crystallographic databases) as the vehicle for rapid and automated processing of crystal structures, and has fuelled their accelerating growth. There are now also many graphics packages that can import and render a CIF, making it a format that supports a whole range of visualisation techniques.

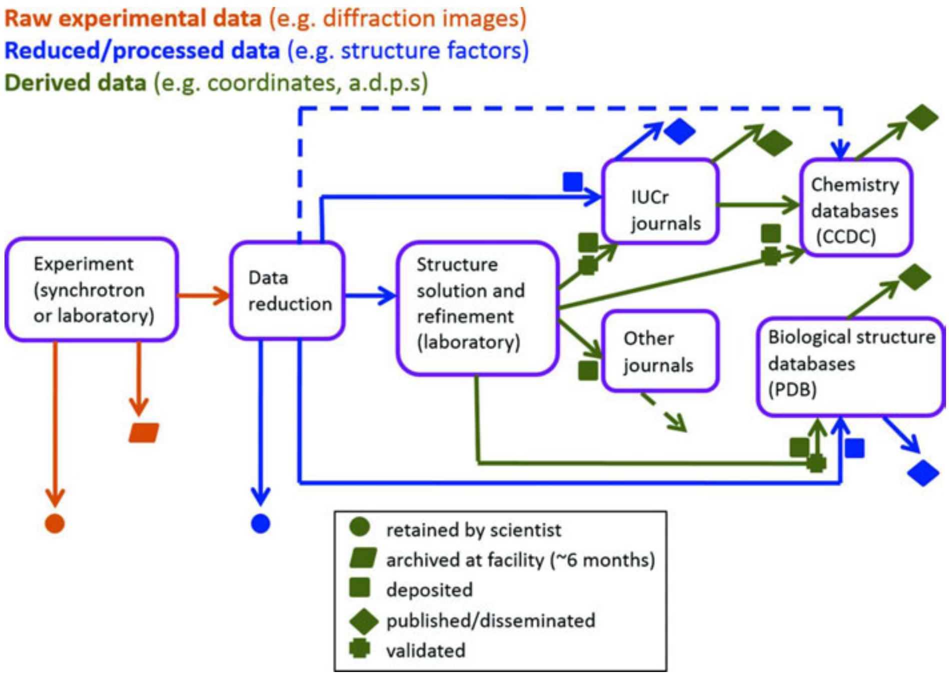

Raw experimental data (e.g. diffraction images)
Reduced/processed data (e.g. structure factors)
Derived data (e.g. coordinates, a.d.p.s)
Experiment (synchrotron or laboratory)
Data reduction
Structure solution and refinement (laboratory)
IUCr journals
Other journals
Chemistry databases (CCDC)
Biological structure databases (PDB)
retained by scientist
archived at facility (~6 months)
deposited
published/disseminated
validated

Figure 3. The coherent information flow provided by CIF for the whole crystallographic process. Image reproduced with permission from Kroon-Batenburg et al. *[7]*

Perhaps the greatest example of its power is the fact that software can automatically check and validate it, not just for syntactic correctness, but also against a set of rules using algorithms. This is embodied in the checkCIF service, [6] which checks for completeness, quality, and consistency in chemical structures and reports back to the user. This unique capability enables all practitioners, irrespective of their level of expertise, to rapidly assess a result and is a key factor in maintaining the rate of data growth in the field, not just according to sheer numbers, but also to quality.

The Rate of Change—New Models and New Science

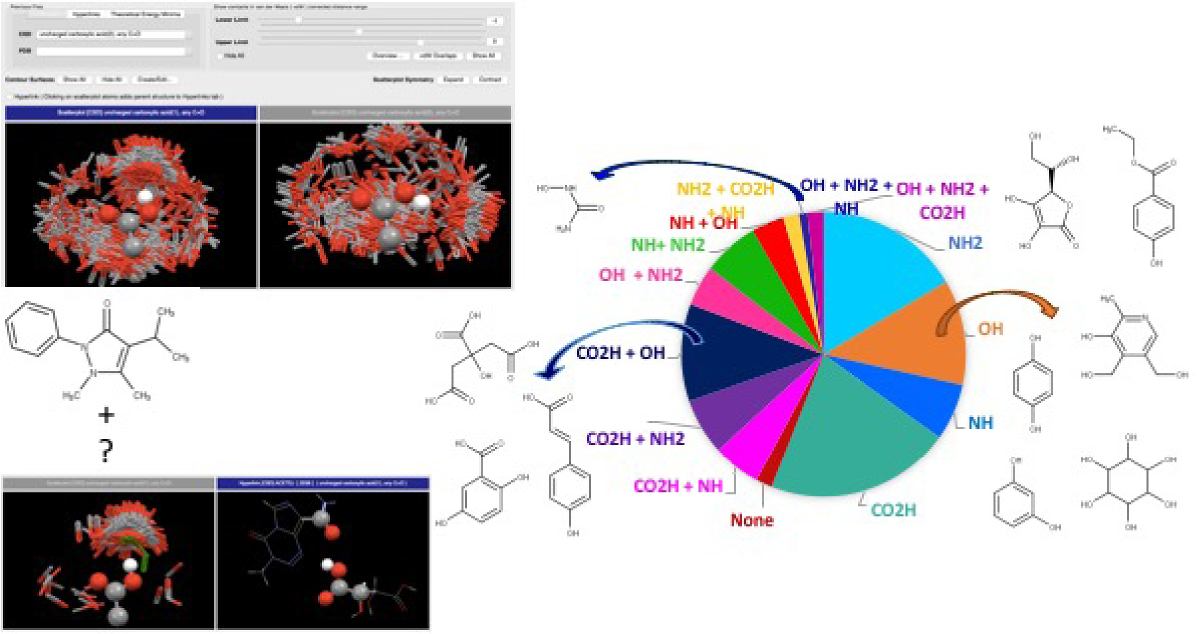
A unified file format and a centralised model don't in themselves mean that a database will grow on its own—there are other stakeholders involved and incentives necessary. In 2015, 90 % of structures released into the public domain as part of an academic publication were deposited into the CSD prior to acceptance by the journal. This is a remarkable statistic, due mainly to coordination in the community and engagement with formal publication routes, and this new model of 'push' rather than 'pull' has certainly eased the path for data into the database.

This is the generally accepted route for the publication of crystallographic data and thus is coupled with, and often governed by, the underlying science and the peer review process. In some senses this can be considered a curse as much as a blessing, for a crystal structure will reach the database only if it is associated with a formal publication. Straw polls indicate that only around 20 % of all crystal structure analyses performed are "published", largely a result of being tied to this process. Moreover, while electronic publication is now the route of choice for the dissemination and discovery of scientific works, this remains simply a mechanism for the process, and the structure and content of an electronic article is largely the same as that of a paper version. The data contained within articles, which is often as important to other scientists as the commentary, remains a second-class citizen in this model. The current rate at which data may be generated and captured, therefore, far outweighs the rate of dissemination.

In the last 15 years the data repository concept has been much explored and, as ever, the crystallographic community has been at the forefront. Institutional, or research group, projects have been undertaken [8] and the Crystallography Open Database [9] has gained significant traction. These approaches enable orphaned data to be readily deposited and disseminated without the need for a parent publication. The CSD has had a similar, 'Private Communication' submission route in place for around 25 years, which has recently been rebranded as 'CSD Communications' and has grown in popularity. It is very interesting to consider that in 2016 this route had a greater number of submissions than any one journal. These innovations raise interesting opportunities for a distributed, somehow federated, network of repositories that could open up the flood gates from the laboratory into the public domain. It is important to note, however, that this doesn't mean the 'centralised CSD' model should be considered outmoded. Rather, it takes on the role of the authoritative resource, which a diverse set of systems feed into. These new models therefore also require the consideration of a range of factors that have hitherto not been of primary concern, such as socio-cultural issues, sustainability, advocacy, organisational factors (both academic/research/labs and library/IT support), the context of the digital scholar, the digital research cycle and data curation, different research practices and workflows, technical interoperability, and open standards.

Moving to these data centric approaches is not just about efficiency and maximising the size of databases—most importantly, this move has enabled new science. Data mining in crystallography in order to provide new chemical insight stems from the work of pioneers, such as Dunitz in the 70's and 80's. However, the first time the power of systematically analysing such data collections struck me was in the heroic work that resulted in reams of tables of bond lengths for particular chemical environments in organic and organometallic compounds, [10,11] which became not only the go-to reference for every crystallographer, but also the basis for a wealth of other work. It is the inspiration for a knowledge base of molecular geometry, [12] essentially a database of these values, but one which has been extended to include angles and torsions and can therefore be used to assess both geometry and conformation in an automated manner. Such data is also the basis for many approaches in molecular mechanics, where knowledge of all the observed geometries for a particular environment is encoded into a force field that is used for predictive purposes—most theoretical studies these days begin with a quick geometry optimisation of the proposed molecule/system that is based on such data. This approach has been extended to intermolecular interactions, but was in fact initiated before the molecular geometry knowledgebase. The wealth of information in a crystal structure goes beyond the boundaries of an individual molecule! This in turn has resulted in a knowledge base of molecular interactions [13] that encodes hydrogen bonding and other intermolecular interaction information, which opens up a whole new area. One can not only query, then make use of, the propensity of a particular interaction, as in crystal engineering, but also evaluate the range of possible interactions between two entities (*e.g.* see Figure 4). This function has extensive applications in fields such as crystal/materials engineering

Figure 4. The use of a molecular interactions knowledge base in the design of pharmaceutically acceptable co-crystals.



and drug docking. The complementarity of chemical and macromolecular crystallography in using these approaches in the field of drug discovery is particularly well covered in a recent article by Colin Groom, the current Director of the Cambridge Crystallographic Data Centre. [14]

What Now?

It is clear that every crystal structure is important—it contributes to a collection which can then be exploited in a range of new scientific areas. As alluded to above, the challenges that persist are the requirement for quality assurance, custodianship, and organisation of the data and also the need to enable routes that maximise the volume of data. When considering all the possible, hypothetical 'chemical space' that has or could be explored and comparing that with coverage in the CSD, there are large, significant differences—many classes of compounds don't have any crystal structures present. There are several significant likely reasons for this. While some of these are perhaps insurmountable, others are not. Examples of the latter include: for one reason or another structures were never published in the scientific literature; crystals were never made, as they were not deemed necessary; compounds were too difficult to crystallise and no significant efforts were made to overcome this; a compound simply won't crystallise; only a small number of structures were deemed necessary to characterise a whole class of compounds; or synthetic chemists have not tried to make a particular class of compounds (and may never do so if there are not research 'drivers' to do so!). A significant culture change is still required to ensure that the considerable secondary benefits of databases continue to evolve and grow in scope.

However, we now live in a data-driven world. The application of algorithms and data to address chemical problems, Cheminformatics, is very much on the rise. Crystallographic data is now being used in ever different ways, which weren't envisaged, either in Bernal's time or more recently! A single example that illustrates this point is the notion of crystallisability, which is of considerable importance in manufacturing many products in the chemical industries. Efforts are being made by mining databases of reported syntheses and comparing them to those of crystal structures in order to understand whether a compound is likely to crystallise. There is also a need to be able to optimise crystallisation based on this work. There are obvious big pitfalls in these studies and, at the very least, the utility of these approaches would be much improved if more laboratory data were available. For this kind of derivative research to be really useful, we would need to know the outcome of *every* crystallisation trial, not just the '1 in a 100' that work.

This leads into my final point—it is primarily the results of science that get published. Often, very little about the research journey that leads to these results is made known. Yet surely science can move forward at a much greater rate if we can run computer algorithms on bodies of data that include information on what a researcher intended to do (and why) and what they actually did, tied to what resulted—preferably for everything that they do! This is still lacking in crystallography—the databases are full of results, not raw data. 'Data' can generally be considered to be raw data, processed data, and derived data. In the crystallographic context, these are diffraction images, structure factors, and crystal structures, respectively. Recently, some progress has been made, in that software will include derived data (structure factors) in the CIF result: validation processes and the CSD will make use of and curate this data. But we can go further still—not only would raw data improve validation processes and provide valuable training sets for software developers to improve algorithms, *etc.*, but there is a more interesting issue— a diffraction experiment records the average signal from the whole sample, which includes defects, impurities, *etc.*, yet often only the data that gets a perfect result is extracted. For materials engineering, it can be crucial to be able to understand these additional effects, yet it is never made public that they have been observed!

Raw data availability, therefore, can be very important, but there are often counter arguments related to costs. The diffraction experiment is relatively quick and cheap, so why not just do it again? The real cost of doing a structure again was assessed by the UK National Crystallography Service as part of the 'Keeping Research Data Safe Project'. [15] There are many nuances to such a cost calculation, but if one has to factor in that the research expertise/group/laboratory that originally generated the material may no longer exist or still be set up (people, apparatus, *etc.*) to make such materials, then the cost rapidly escalates. The replacement cost of the CSD is therefore almost immeasurably large!

Data transfer and storage are problems that are now being overcome. For around 15 years there has been an 'extension' to CIF (imgCIF) that can cater for raw data, although its uptake has been very slow. So why aren't we amassing more of our valuable raw data for the community to widely exploit? For the last five years, a group known as the Diffraction Data Deposition Working Group has been looking into the issues surrounding this topic. The outcome from the activity of this group is that the International Union of Crystallography has recently convened a Committee on Data, 'CommDat', [16] as an advisory committee to the Executive. CommDat will work across the organisation, subsuming data-related interests of Journals, the now discontinued Committee on Crystallographic Databases, and of the Committee on Electronic, Publishing, Dissemination and Storage of Information, and will also have a formal relationship with COMCIFS.

To conclude, the answer to the question in the title, "Are we nearly there yet?", is that I doubt we ever will be! The scientific endeavour will continue to stretch us, and data is the foundation of our research. However, the lessons we have learned along the route so far are applicable to many disciplines outside of crystallography and provide us with a very strong basis for the future. As a coherent community that considers data matters very seriously indeed, we continue to look forward and encourage others to follow.

References

1. J.D. Bernal, *The Social Function of Science,* 1939, London: George Routledge & Sons Ltd.
2. C.R. Groom, I.J. Bruno, M.P. Lightfoot & S.C. Ward, The Cambridge Structural Database, *Acta Cryst.* **B72**:171-179, 2016.
3. O. Kennard, *The Impact of Electronic Publishing on the Academic Community*, 1997, London: Portland Press Ltd.
4. S.R. Hall, F.H. Allen & I.D. Brown, The Crystallographic Information File (CIF): a new standard archive file for crystallography, *Acta Cryst.* **A47**:655–685, 1991.
5. www.iucr.org/resources/cif/comcifs
6. A.L. Spek, Structure Validation in Chemical Crystallography, *Acta Cryst*. **D65**:148-155, 2009.
7. L.M.J. Kroon-Batenburg, J.R. Helliwell, B. McMahon & T. Terwilliger, Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements, *IUCrJ* **4**:87-99, 2017.
8. http://ecrystals.chem.soton.ac.uk
9. www.crystallography.net/cod/
10. F.H. Allen, O. Kennard, D.G. Watson, L. Brammer, A.G. Orpen & R. Taylor, Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds, *J. Chem. Soc., Perkin Trans*. **2**:S1-S19, 1987.
11. A.G. Orpen, L. Brammer, F.H. Allen, O. Kennard, D.G. Watson & R. Taylor, Tables of bond lengths determined by X-ray and neutron diffraction. Part 2. Organometallic compounds and co-ordination complexes of the d- and f-block metals, *J. Chem. Soc., Dalton Trans.* S1-S83, 1989.
12. I.J. Bruno, J.C. Cole, M. Kessler, J. Luo, W.D.S. Motherwell, L.H. Purkis, B.R. Smith, R. Taylor, R.I. Cooper, S.E. Harris & A.G. Orpen, Retrieval of Crystallographically-Derived Molecular Geometry Information, *J. Chem. Inf. Comput. Sci.* **44**(6):2133–2144, 2004.
13. I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor & M.L. Verdonk, IsoStar: A library of information about nonbonded interactions, *J Comput Aided Mol Des*, **11**:525, 1997.
14. C.R. Groom & J.C. Cole, The use of small-molecule structures to complement protein–ligand crystal structures in drug discovery, *Acta Cryst.* **D73**:240-245, 2017.
15. Keeping Research Data Safe (KRDS1) Final Project Report, 2008, www.beagrie.com/krds/.
16. IUCr Committee on Data (CommDat) www.iucr.org/iucr/governance/advisory-committees/committee-on-data.

Simon Coles <S.J.Coles@soton.ac.uk> is Professor of Structural Chemistry at the University of Southampton and Director of the UK National Crystallography Service. ORCID.org/0000-0001-8414-9272