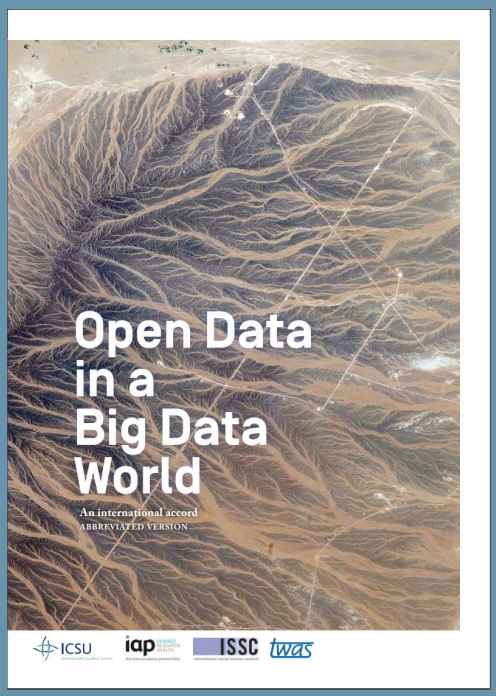
# Open Data in a Big Data World

A position paper for crystallography

# The 2015 Science International Accord

**The International Union of Crystallography (IUCr) notes the publication by the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC) of an international Accord on *the values of open data in the emerging scientific culture of big data*, following the 2015 Science International meeting. The IUCr *acknowledges the importance* of this Accord, and *endorses the analysis of the values of open data and the Principles of Open Data* set out in the document *Open Data in a Big Data World*, published in short and long forms on the ICSU website at http://www.icsu.org/science-international/accord.**

**The Accord is very general, and has applicability across the entire panorama of science, which it defines as embracing 'all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines'. Because the specific values, significance and implementation of Open Data principles will vary in detail between disciplines, the IUCr considers it useful to contribute this *detailed response* to the Accord, as a case study of best practice emerging in one particular field.**




This text was prepared for the IUCr by

**Marvin L. Hackert,** *IUCr President and IUCr Representative to ICSU*
*Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA*

**Luc Van Meervelt,** *IUCr General Secretary and Treasurer*
*Chemistry Department, Katholieke Universiteit Leuven, Celestijnenlaan 200F, BE-3001, Leuven, Belgium*

**John R. Helliwell,** *IUCr Representative to CODATA*
*School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK*

**Brian McMahon,** *IUCr Research and Development Officer*
*International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK*

August 2016

# Executive Summary

Science is best served when access barriers to data (and publications) are low. However, the maintenance of the highest levels of quality in collecting, evaluating, storing and curating data is a very expensive component of the scientific process. Crystallography has a diverse ecosystem of approaches to sustainability and quality assurance.

Technological advances in scientific instrumentation and computer technology have dramatically increased the quantities of data involved in scientific inquiry. This Accord expresses the dependence of scientific assertions on supporting data. Inasmuch as science is by its nature interrogative, the Accord asserts that 'openness and transparency have formed the bedrock on which the progress of science in the modern era has been based'. The IUCr supports this assertion, but notes that openness alone (by which we mean the ability to access and re-use scientific data with little or no restriction) is not sufficient. The data openly accessed must be subject to critical scrutiny, through peer review and automated validation where possible. There is also a case to be made for raw data to be retained for as long a time as feasible, to permit re-evaluation that takes account of novel analytic techniques or that periodically employs different methodologies to eliminate systematic procedural error.

All scientific data must be subject to rigorous first analysis to exclude or quantify systematic bias or error; all software implementations should employ open algorithmic procedures and their results should ideally be cross-checked by independent implementations. An overlooked challenge in handling ever-growing volumes of data is the need to apply the same level of critical evaluation as has been applied to historically smaller volumes.

***Openness alone is not sufficient. Data openly accessed must be subject to critical scrutiny, through peer review and automated validation where possible***

The Accord does not formally define 'Open Data', but implies certain properties throughout its careful discussions. We hold that the essential component of openness is that the data supporting any scientific assertion should be

- **complete** (*i.e.* all data collected for a particular purpose should be available for subsequent re-use); and
- **precise** (the meaning of each datum is fully defined, processing parameters are fully specified and quantified, statistical uncertainties evaluated and declared).

Together, these properties include the criteria of Paragraph 8 of the Accord (long form), that open data should be *discoverable*, *accessible*, *intelligible*, *assessable* and *usable*. We note, however, that a full understanding of the data may depend on associated scientific publications that discuss the details of data processing where these differ from routine practice. The full linking of article and data is another key element of openness.

Science is best served when access barriers to data (and publications) are low. A major barrier to access is cost, and the phrase 'open access' is often used to characterize access to data and publications that involve no charge to the end-user. However, the maintenance of the highest levels of quality in collecting, evaluating, storing and curating data is a very expensive component of the scientific process, and care must be taken to understand how to obtain the maximum benefit from public funding of science. In many fields, it may indeed be cost-effective to provide direct funding to repositories or publishing platforms that require no further payment to access. In other fields, the situation is less clear cut.

Crystallography has a diverse ecosystem of disciplinary databases, data repositories, experimental facilities and publishers. Several of these are sustained through subscription-based access; but the other side of the coin is that they ingest, evaluate and publish data and information at no charge to the author/depositor, and without imposing any additional charge on the public purse. At the present time, this variety of approaches to sustainability and quality assurance serves this discipline well.

The SACLA X-ray free-electron laser facility in Hyogo, Japan, makes it possible to observe atoms and molecules in real time – generating vast amounts of data in the process.

Photo credit: RIKEN/XFEL

# APPENDIX: Annotated Accord

We illustrate the points made in the *Executive Summary* by annotating the relevant parts of the short form of the Accord (reproduced in red below). Where we make no explicit comment, it may be taken that we are in tacit agreement with that part of the Accord.

*This accord is presented as an outcome of "Science International 2015", the first of a series of annual meetings of four top-level representatives of international science (the International Council for Science – ICSU, the InterAcademy Partnership – IAP, The World Academy of Sciences – TWAS and the International Social Science Council – ISSC) that are designed to represent the global scientific community in the international policy for science arena. The accord identifies the opportunities and challenges of the data revolution as today's predominant issue for global science policy. It proposes fundamental principles that should be adopted in responding to them. It adds the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite in maintaining the rigour of scientific inquiry and maximising public benefit from the data revolution in both developed and developing countries. Science International partners will promote discussion and adoption of these principles and their endorsement by their respective members and by other representative bodies of science at national and international levels.*

The IUCr welcomes the interest of high-level international stakeholders in presenting a united voice that stresses the importance of scientific inquiry world-wide. In a world of expensive research programmes, often largely dependent for funding on income raised from public taxation, it is important that the needs and opportunities of small and developing countries are considered alongside those of the developed world.

***The IUCr welcomes the interest of high-level international stakeholders in presenting a united voice that stresses the importance of scientific inquiry world-wide***

The IUCr represents the worldwide community of scientists in the field of crystallography and related structural sciences. It comprises 50 Adhering Bodies representing 58 distinct nations. The IUCr itself is a member of ICSU and of CODATA, ICSU's Committee on Data for Science and Technology.

## 1. The Big Data World

*The digital revolution of recent decades is a world historical event as deep [as] and more pervasive than the introduction of the printing press. It has created an unprecedented explosion in the capacity to acquire, store, manipulate and instantaneously transmit vast and complex data volumes, with profound implications for science[1]. The rate of change is formidable. In 2003 scientists declared the mapping of the human genome complete. It took over 10 years and cost $1billion – today it takes mere days and a small fraction*

***Macromolecular structures in Protein Data Bank: > 125,000***

*of the cost ($1000). "Big data", in which unprecedented fluxes of data stream in and out of computational systems, and "Broad Data" in which numerous datasets can be semantically linked to create deeper meaning, are the engines of this revolution, offering novel opportunities to natural, social and human sciences.*

*[1] The word "science" is used to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.*

While the opening section correctly raises the subject of the large volume of research data routinely generated and collected nowadays, and its applicability across many subject areas, we note that the proper conduct of science has always depended on a deep understanding of the nature of the data collected in any research effort, and a careful and proper analysis of its accuracy, precision and validity.

While very high data volumes and data acquisition rates may make it increasingly difficult to treat data with the care and respect that it needs, it is nevertheless essential to good science that every effort is made to do so. The ubiquity of data is no substitute for proper and critical analysis.

## 2. The Opportunities

*The scientific opportunities of this data-rich world lie in discovering patterns that have hitherto been beyond our reach; in linking and correlating different aspects of systems better to understand their behaviour; in characterising complexity; and in iterating between descriptions of the state of a complex system and simulations that forecast its dynamic behaviour. There are many areas of research where such capacities are deeply relevant: in weather and climate forecasting; in understanding the workings of the brain; in the behaviour of the global economy; in evaluating agricultural productivity; in demographic forecasts; in unravelling histories; and in many contemporary global challenges such as those of environmental change, infectious disease and mass migration that require combined insights and data from many disciplines.*

***The proper conduct of science has always depended on a deep understanding of the nature of the data collected … and a careful and proper analysis of its accuracy, precision and validity***

Crystallography abounds in examples of scientific laws and applications being derived from collecting and correlating data. Historic classification of naturally occurring crystal forms led to the understanding of lattice symmetries, packing arrangements and energetics. Subsequent probing of crystal structures by X-ray diffraction and other techniques

***Molecular structures in Cambridge Structural Database: > 800,000***

yielded vital information on the nature of chemical bonds, molecular structures and solid-state properties of materials. The elaboration of the structures of DNA and proteins created great insights into biological processes, and the availability of large and growing databases of nucleic acid and protein structures feeds into the enormous advances being made in genetics and therapeutics. Pioneering developments in time-resolved structural dynamics using synchrotrons and X-ray free-electron lasers probe the very nature of chemical reactions. Each of these developmental phases has involved

***Inorganic and organic structures in Crystallography Open Database: > 365,000***

ever-growing volumes and complexity of data that challenged the science of the time. We see the current 'Data Deluge' as just the latest (albeit large) step up in this evolutionary process. We welcome both the challenges it brings and the prospects for future discovery.

## 3. The Challenges

*Grasping these opportunities poses serious challenges to the way science is done and organised. Open data are the common, enabling threads.*

### The Open Data Imperative

*The fundamental role of publicly funded research is to add to the stock of knowledge and understanding that are essential to human judgements, innovation and social and personal wellbeing. The technologies and processes of the digital revolution provide a powerful medium through which scientific productivity and creativity can be enhanced by permitting data and ideas to flow openly, rapidly and pervasively through the networked interaction of many minds. If this social revolution in science is to be realised it is vital that we adopt a default position that publicly funded data should be made publicly accessible and re-usable when a*

*research project through which the data have been collected is completed.*

While the bulk of this Accord focuses on publicly funded research – appropriately, as its main purpose is to help shape public policy – many of the principles it discusses are important to the proper conduct of science in any manifestation, including its practice within the private sphere. Some privately funded research

***Inorganic crystal and powder diffraction data sets in Powder Diffraction File: > 384,000***

does contribute directly to the public good, *e.g.* through academic publications; some is harnessed for commercial gain. While there are some additional pressures that reduce the ways in which such data are openly shared outside of the originating stakeholder, we nevertheless feel that the principles stated in this Accord should be considered as ideals towards which all scientific effort should aspire. We see below that, even in the area of publicly funded research, there can be factors that moderate the practical open dissemination of data. We urge the worldwide community of scientists, whether publicly or privately funded, always to have the starting goal to divulge fully all data collected or generated in experiments, and to temper this goal only so far as is absolutely necessary to allow the basic enterprise to be maintained in a sustainable way and with full scientific and ethical integrity.

## Maintaining self-correction

*Openness of the evidence (the data) for scientific claims is the bedrock of scientific progress. It permits the logic of an argument to be scrutinised and the reproducibility of observations or experiments to be tested, thereby supporting or invalidating those claims. When a paper making a scientific claim is published, it is essential that the evidentiary data, the related metadata that permit their re-analysis, and the codes used in computer manipulation are made concurrently open to scrutiny to ensure that the vital process of self-correction is maintained. Recent demonstrations in several disciplines of high rates of non-reproducibility of results of published papers emphasise the crucial need to re-invigorate open data processes for a big data world. Openness is not however enough. Data must be intelligently open, meaning that they should be: discoverable, accessible, intelligible, assessable and (re-)usable.*

The reference to *evidentiary* data is a useful reminder that data collected in the course of a scientific inquiry may play a variety of roles. Our summary of ways in which crystallographic data has led to novel scientific hypotheses and conclusions (Section 2) largely refers to data sets that are themselves derived scientific models. (They are tabulations of atomic positional and displacement parameters, with associated information about chemical nature, modelling constraints or restraints, and many other metadata relating to provenance, analytical procedures, calculated precision of derived values, *etc.*) These models are constructed from experimental data, typically the diffracted intensities from a scattered collimated radiation or particle beam.

***We urge the worldwide community of scientists, whether publicly or privately funded, always to have the starting goal to divulge fully all data collected or generated in experiments***

***Inorganic crystal structures in Inorganic Crystal Structure Database: > 185,000***

These evidentiary (experimental) data supporting each structural data set should also be made available as part of the scientific record. For biological macromolecular structures, the community expects that such evidentiary data should be deposited in the same curated database (the Worldwide Protein Data Bank) as the structures themselves. These tabulations of intensities ('structure factors') are also required for small-molecule or inorganic structures published by the journals of the IUCr. They are not always required as part of the submission of structural results by other journal publishers, but their deposition in the relevant structural databases is increasingly encouraged as best community practice. There has long been a convention in crystallography that individual structural data sets and supporting sets of structure factors may be freely downloaded, even from journals or databases where subscriptions are required to access the published articles or complete database.

***Pauling File: > 41,000 phase diagrams, > 290,000 crystal structures, > 106,000 physical property entries***

More recently, there has been growing interest within the field of crystallography to retain the raw data for each structure determination experiment. These most commonly take the form of a collection of two-dimensional images capturing the diffracted beams as the crystal sample is rotated through all orientations relative to the incident beam. It is from these images that the more concise set of processed diffraction data (structure factors) is derived. While a set of structure factors is typically a few megabytes (MB) in size, the raw diffraction images may occupy many gigabytes (GB). For the traditionally small-scale data volume requirements of structural science, this does amount to a foray into 'Big Data'. Rapid improve-

ments in detector technology and developments in dynamical structure elucidation with intense synchrotron and X-ray laser sources will increase the volume of raw data collected by further orders of magnitude.

There has been considerable discussion in the last few years of the need to archive the primary data sets [1, 2]. Many researchers consider that the structure factors are adequate (in most cases) to validate the derived structural model. To the extent that crystal structure determination experiments are largely homogeneous in their methods and equipment, a high degree of confidence can be placed in the reduction processes that lead to the structure factors. However, errors of interpretation are sometimes made, and access to the raw data can help to mitigate this. Furthermore, diffuse scattering in the original images contains information about internal molecular dynamics or the correlated dynamics over different distance scales in the crystal. This is generally ignored during standard data reduction, and so potentially valuable scientific information is lost. There is a growing sense that at least some proportion of raw data images should therefore be retained for purposes of validation, new scientific discovery, and development or testing of novel software methods.

**DDDWG**

The Diffraction Data Deposition Working Group of the IUCr has been active since 2011, scoping the demand and practical requirements for routine deposition of diffraction images, the raw experimental data sets from many crystallographic experiments. Activities also involve the characterisation of essential metadata for describing the great variety of crystallographic and related structural experiments. Links to Workshops, discussion forums and other activities are at http://www.iucr.org/resources/data/dddwg.

There is also value in retaining primary data as a safeguard against the publication of results that are fraudulently derived. However, this should not be overstated; individuals motivated to fabricate scientific evidence may still find ways of doctoring primary experimental data. Although access to primary data can help to discourage unethical scientific behaviour, it cannot act as a complete preventative. This example serves to highlight the fact that more data, by itself, does not change the need to treat all data with appropriate care, respect and critical analysis.

***CrystMet database of metals, alloys and intermetallics: > 161,000 entries***

***There has been growing interest within the field of crystallography to retain the raw data for each structure determination experiment***

IUCr-sponsored data exchange standards provide for very detailed metadata to define precisely the content and context of a data set, and recent efforts have focused on the need to define (for all experiments) the metadata needed to understand fully the data collected, and to permit reproducibility of the experiment [3].

## Adapting scientific reasoning

*Many of the complex relationships that we now seek to capture through big- or broad- linked data lie far beyond the analytical power of many classical statistical methods. They require deeper mathematical approaches including topological methods to ensure that inferences drawn from big data and broad data are valid. Data-intensive machine-analysis and machine-learning are becoming ubiquitous, and have major implications for scientific discovery. The complexity of patterns that machines are able to identify are not easily grasped by human cognitive processes, posing profound issues about the human-machine interface and what it might mean to be a researcher in the 21st century.*

As indicated in the introductory comments, crystallography is already an established leader in deriving complex relationships from extensive data collections, albeit much of this research successfully uses classical statistical methods. Further advances will be facilitated by harnessing the potential of 'broad-linked' data, *i.e.* by permitting text mining of publications and data mining of their associated data sets (including, as appropriate, the raw or processed experimental data that underpin the structural data models). Furthermore, automated discovery and analysis are assisted by the curation of a discipline-specific machine ontology (the Crystallographic Information Framework, CIF) [4] and the development of software that can use this ontology directly to test and follow linkages between granular concepts expressed within the data or associated publications.

In the case of IUCr journals, open-access articles are published through a Creative Commons attribution licence that permits text mining. For other articles whose publication is financed through journal subscriptions (a 'paywall'), the IUCr will provide free access for text-mining robots to *bona fide* researchers.

Work continues under the sponsorship of the IUCr to increase interoperability between the growing family of crystallography related ontologies ('CIF dictionaries') and cognate ontologies in related areas of science (*e.g.* Chemical Markup Language, CML, in the description of chemical structures and reactions; NMRStar for protein NMR conformation studies; macromolecular

structure application profiles in NeXus/HDF5 image acquisition files).

***Bilbao Incommensurate Structures Database: > 130 incommensurate modulated and composite structures***

## Ethical constraints

*The open data principle has ethical implications for researchers and research subjects. It can appear to override the individual interests of the researchers who generate the data, such that novel ways of recognising and rewarding their contribution need to be developed. The privacy of data subjects needs to be protected. In a regime of open sharing in which data are passed on from their originators, there is loss of control over future usage, whilst anonymisation procedures have been demonstrated to be unable to guarantee the security of personal records.*

### Crystallographic Information Framework

The Crystallographic Information Framework is a suite of machine-readable ontologies, data exchange formats and software applications and services developed by the IUCr since 1991 for data definition and exchange in crystallography and related structural sciences. The standards are fully documented on the web (cif.iucr.org) and in print (*International Tables for Crystallography Volume G: Definition and exchange of crystallographic data*). Specific ontologies exist for single-crystal and powder X-ray diffraction, biological macromolecular structures (proteins and nucleic acids), modulated and composite structures, electron density, twinning, symmetry and diffraction images.

## Open global participation

*Big data and open data have great potential to benefit less affluent countries, and especially least developed countries (LDCs). However, LDCs typically have poorly resourced national research systems. If they cannot participate in research based on big and open data, the gap could grow exponentially in coming years. They will be unable to collect, store and share data, unable to participate in the global research enterprise, unable to contribute as full partners to global efforts on climate change, health care, and resource protection, and unable fully to benefit from such efforts, where global solutions will only be achieved if there is global participation. Thus, both emerging and developed nations have a clear, direct interest in helping to fully mobilize LDC science potential and thereby to contribute to achievement of the UN Sustainable Development Goals.*

A major project of the International Year of Crystallography in 2014 was to launch a series of capacity-building 'open laboratories' in many developing countries [5]. These often involved the loan of equipment from commercial vendors and hands-on training in the use of the equipment and the proper handling of generated data. For research in chemical crystallography, many results will be generated in local laboratories. Most equipment and software uses the open CIF standard; the IUCr and other crystallographic institutions provide free or open-source software for standard reduction and analysis of the experimental data, for characterizing the derived structural data sets, and for preparing articles for publication. Open-access article processing charges are reduced or waived for authors from developing countries. For larger-scale or more complex experiments, often conducted in synchrotron or neutron facilities, there are initiatives to develop regional resources (*e.g.* in the Middle East and Africa) that will provide access to the necessary equipment for LDCs. Through liaison with established facilities, IUCr working groups aim to encourage common modes of practice amongst the larger facilities with respect to data management and archiving.

## Seizing the opportunity

*Effective open data can only be realised if there is systemic action at personal, disciplinary, national and international levels. Although science is an international enterprise, it is done within distinctive national systems of responsibility, organisation and management, all of which need to respond to the opportunity. Research funders and research performing institutions should fund and implement processes that lighten the burden on researchers of making data intelligently open and that support open data processes. Increasing numbers of research communities have discovered the benefits of sharing data, in fields as varied as linguistics, bio-informatics and chemical crystallography, and have made major strides in realising benefit for their disciplines through international collaboration in facilitating access and use of open data. Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA), its World Data System (WDS) and the Research Data Alliance (RDA), to promote and support developments of the systems and procedures that will ensure international data access, interoperability and sustainability.*

***A major project of the International Year of Crystallography in 2014 was to launch a series of capacity-building 'open laboratories' in many developing countries***

***Discrete data items defined in the core CIF dictionary: 802***

The IUCr formally documents every aspect of its data standardization programme on its website and through its journals and reference works [6]. It is an active member of CODATA, and seeks synergies with WDS and the RDA, and other international organizations such as the International Council for Scientific and Technical Information, ICSTI.

## Open science and public knowledge

*The idea of "open science" has developed in recognition of the need for stronger dialogue and engagement by the scientific community with wider society in addressing many current problems through reciprocal framing of the issues and the collaborative design, execution and application of research. There are, of course, legitimate limits to openness, such as the need to protect security, privacy and proprietary concerns through judiciously applied mechanisms. There are also countervailing trends towards privatisation of knowledge that are at odds with the ethos of scientific inquiry and the basic need of humanity to use ideas freely. If the scientific enterprise is not to founder under such pressures, an assertive commitment to principles of open data, open information and open knowledge is required from the global scientific community.*

***Discrete data items defined in the macromolecular CIF extension dictionary: 5631***

The IUCr included many public outreach activities in its programme for the International Year of Crystallography, and is committed to maintain and expand such activities [7].

There are precedents (especially in structural biology) for retaining exclusive rights to access experimental data sets for a finite period of time. While such embargoes are permitted *e.g.* by the Worldwide Protein Data Bank, the relevant IUCr bodies are supportive of a move towards minimizing or removing them altogether.

We recommend caution in the use of terms such as 'privatisation'. While the IUCr supports the ideal of full open access to scientific data and knowledge, the proper maintenance and curation of databases, data repositories and publications is expensive. For a variety of reasons, not all such facilities are funded directly or fully from the public purse, and scientific endeavour remains a diverse ecosystem in terms of funding and business models. IUCr Journals, first published in 1948, grew according to the universal subscription model of the time. Although there is movement in the direction of open-access publication (the IUCr has two fully open-access titles), there are still many authors who will not or cannot pay the article processing charges necessary to sustain this publishing model. Therefore we currently offer a hybrid model where individual articles may be open access or behind a subscription paywall. Similarly, the most comprehensive database of small-molecule chemical structures (the Cambridge Structural Database, CSD) is funded through subscriptions, again for historical reasons: a previous national Government insisted that public funding of the academically based Cambridge Crystallographic Data Centre that maintains the CSD be replaced by a self-sustaining business model. In cases such as this, the 'private' status of scientific service providers does not imply that their primary objective is other than to advance the cause of science.

***Structural data sets freely available from IUCr journals: > 58,800***

Note also our comments near the start of Section 3 where we commend the principles set out in this Accord as equally applicable to publicly and privately funded research.

## 4. Principles of Open Data

*Such is the importance and magnitude of the challenges to the practice of science from the data revolution that Science International believes it appropriate to promote the following statement of principles of open data.*

***The IUCr formally documents every aspect of its data standardization programme on its website and through its journals and reference works***

### Responsibilities

***Scientists***

***i.*** *Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed.*

***ii.*** *The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form[2]. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To the extent possible, data should be deposited in well-managed and trusted repositories with low access barriers.*

[2] *Refer to the full text document at http://www.science-international.org*

We reiterate that 'low access barriers' may involve payment from the end-user. This should be at a rate that ensures sustainability of the repository, and that allows for appropriate levels of quality control of the deposited data and associated services.

We note also that 'low access barriers' include the technical facilitation of reuse

through open standards, well-documented APIs (application programming interfaces), and rich metadata describing the nature, use and relevance of each data set – that is to say, the 'intelligently open' form alluded to in this principle.

***iii. Research institutions and universities** have a responsibility to create a supportive environment for open data. This includes the provision of training in data management, preservation and analysis and of relevant technical support, including library and data management services. Institutions that employ scientists and bodies that fund them should develop incentives and criteria for career advancement for those involved in open data processes. Consensus on such criteria is necessary nationally, and ideally internationally, to facilitate desirable patterns of researcher mobility. In the current spirit of internationalisation, universities and other science institutions in developed countries should collaborate with their counterparts in developing countries to mobilise data-intensive capacities.*

***iv. Publishers** have a responsibility to make data available to reviewers during the review process, to require intelligently open access to the data concurrently with the publication which uses them, and to require the full referencing and citation of these data. Publishers also have a responsibility to make the scientific record available for subsequent analysis through the open provision of metadata and open access for text and data mining.*

For chemical crystallography, IUCr journals require all derived structural models and the processed experimental data sets underpinning them to be submitted for peer review (and subsequent publication). The journals provide an automated service, *checkCIF*, that rigorously tests the completeness and internal consistency of the submitted data [8]. This service is openly available to authors in advance of submission as well as to the reviewers. In some of the journals, the synoptic *checkCIF* report on a structure is also provided as a supplement to the published article. Furthermore, because the service is openly available, anyone may generate a post-publication validation report to assess the precision of the determined structure. *checkCIF* is also used by other publishers of chemical crystallographic data sets.

***Experimental intensity data sets (structure factors) freely available from IUCr journals: > 58,400***

***'low access barriers' include the technical facilitation of reuse through open standards, well-documented APIs (application programming interfaces), and rich metadata***

For macromolecular structures, a validation report is created by database curators when a structural data set is deposited. (Within this discipline, such deposition typically occurs in advance of submission of research articles.) IUCr journals require authors to provide the validation report upon submission. Processed experimental data are also deposited with the structural databases; increasingly reviewers request this (and the raw experimental data) from authors. This is a voluntary process, but there is evidence that the community increasingly considers it as a necessary practice.

***v. Funding agencies** should regard the costs of open data processes in a research project to be an intrinsic part of the cost of doing the research, and should provide adequate resources and policies for long-term sustainability of infrastructure and repositories. Assessment of research impact, particularly any involving citation metrics, should take due account of the contribution of data creators.*

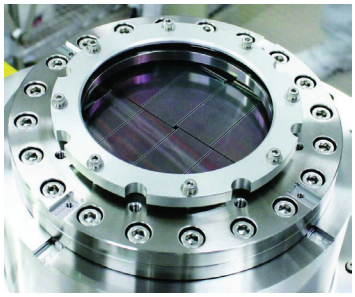
IUCr journals assign unique identifiers (DOIs) to all information supporting a publication, including derived and experimental data sets. This helps in providing citations for data. The IUCr has also launched a new service in 2016, *IUCrData*, which provides a fully citable form of short reports on crystallographic data.

***vi. Professional associations, scholarly societies and academies** should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.*

The IUCr does this consistently through its journal editorial guidelines, through the activities of advisory committees, working groups and Representatives on data and publishing organisations, through the community guidance and interactions of its Commissions, and since 2011 through the coordination and development activities of its Diffraction Data Deposition Working Group.

## A data deluge

The camera head of a multi-port charge-coupled detector developed for serial femtosecond crystallography [Hatsui & Graafsma (2015), *IUCrJ* **2**, 371–383].



Time-resolved crystallography at high-flux radiation sources with a resolution of femtoseconds can generate hundreds of gigabytes of raw data per experiment.

***vii. Libraries, archives and repositories** have a responsibility for the development and provision of services and technical standards for data to ensure that data are available to those who wish to use them and that data are accessible over the long term.*

The IUCr develops metadata standards within its ontological framework that facilitate data characterization, archiving, validation and exchange. Although managed in a format that is almost unique to the discipline ('CIF'), care is taken to ensure ready interoperability with generic metadata standards in the library and repository worlds.

## Boundaries of openness

***viii.** Open data should be the default position for publicly funded science. Exceptions should be limited to issues of privacy, safety, security and to commercial use in the public interest. Proposed exceptions should be justified on a case-by-case basis and not as blanket exclusions.*

All crystallographic data published by IUCr journals are openly available. Limited embargo practices are found in some areas, but the IUCr encourages full disclosure of all supporting data.

## Enabling practices

***ix. Citation and provenance:** When, in scholarly publications, researchers use data created by others, those data should be cited with reference to their originator, to their provenance and to a permanent digital identifier.*

See comments under (v) regarding the assignment of permanent digital identifiers and opportunities for citation. The IUCr is a formal signatory to the Force11 principles on data citation [9].

***Experimental powder profile data sets freely available from IUCr journals: > 1030***

***x. Interoperability:** Both research data, and the metadata which allows them to be assessed and reused, should be interoperable to the greatest degree possible.*

This has been a principle of the IUCr since its inception in 1947, practised in the publication of derived and experimental data (in hard-copy form) since its journals were first published in 1948, and facilitated in the electronic age by the development of successive machine-readable standards, such as the Standard Crystallographic File Structure in 1981 [10], and the Crystallographic Information File in 1991 [11].

***xi. Non-restrictive reuse:** If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be re-used with no more arduous requirement than that of acknowledging the producer.*

***xii. Linkability:** Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value.*

***IUCr journals assign unique identifiers (DOIs) to all information supporting a publication, including derived and experimental data sets***

## Notes and References


[1] Significant community discussion moderated by the IUCr Diffraction Data Deposition Working Group is archived in an IUCr discussion forum http://forums.iucr.org/viewforum.php?f=21

[2] Terwilliger, T. C. (2014). Archiving raw crystallographic data. *Acta Cryst.* D**70**, 2500–2501.

[3] See the record of the two-day workshop on 'Metadata for raw data from X-ray diffraction and other structural techniques', http://www.iucr.org/resources/data/dddwg/rovinj-workshop

[4] Hall, S. R. and McMahon, B. (2016). The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Sci. J.* **15**, p. 3. DOI: http://doi.org/10.5334/dsj-2016-003

[5] http://www.iycr2014.org/openlabs

[6] Hall, S. R. and McMahon, B. eds. (2005). *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data.* First edition Dordrecht: Springer.

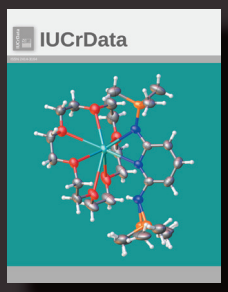
[7] http://www.iycr2014.org/into-the-future/conference/resolution

[8] Spek, A. L. (2009). Structure validation in chemical crystallography. *Acta Cryst.* D**65**, 148–155.

[9] https://www.force11.org/datacitation/endorsements

[10] Brown, I. D. (1988). Standard Crystallographic File Structure-87. *Acta Cryst.* A**44**, 232.

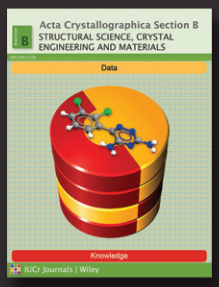
[11] Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Cryst.* A**47**, 655–685.

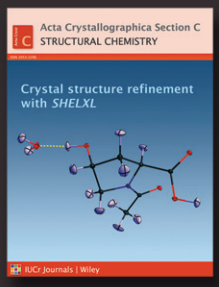


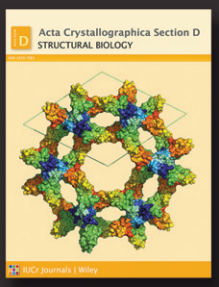

*IUCrData* is a peer-reviewed open-access data publication from the International Union of Crystallography (IUCr), first published in 2016. This innovative publication aims to provide short descriptions of crystallographic data sets and data sets from related scientific disciplines, as well as facilitating access to the data. The primary article category is *Data Reports*; these describe crystal structures of inorganic, metal-organic or organic compounds. Information on each crystal structure includes the crystallographic data (CIF and structure factors), a data validation report, figures and a text representation of the data.

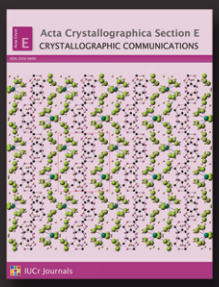
International Union
of Crystallography
2 Abbey Square
Chester CH1 2HU
UK
http://www.iucr.org

The IUCr is an International Scientific Union. Its objectives are to promote international cooperation in crystallography and to contribute to all aspects of crystallography, to promote international publication of crystallographic research, to facilitate standardization of methods, units, nomenclatures and symbols, and to form a focus for the relations of crystallography to other sciences.

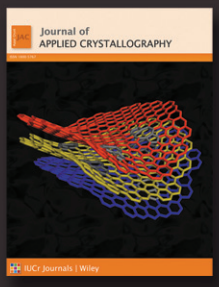



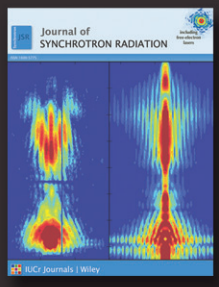






journals.iucr.org