Crystallographic Information and Data Management

A Satellite Symposium to the 28th European Crystallographic Meeting

Programme and background materials

Organised by COMCIFS, the IUCr Committee for the Maintenance of the CIF Standard

An important one-day Symposium to celebrate and develop the role of the CIF information exchange standard in crystallography: from diffraction images to structure solution, refinement, validation, publication and archiving. Speakers will explain how crystallography, particularly in its application to structure determination, is a field with supremely well developed practices in the collection, analysis, interpretation, publication and archiving of raw and processed data, and the structural information derived from diffraction experiments.

These practices are so well ingrained that many crystallographers are unaware of how well we do these things, compared with other scientific fields. The Symposium will remind the crystallographic community of its great track record in data handling, and also anticipate further advances in data management, integration, validation and curation.

Much of the community's success in information management arises from the Crystallographic Information Framework (CIF), a standard now entering its third decade, developed originally for small-molecule structural modelling but which is now used directly in many areas of crystallography, or informs the development of formal data definitions in others. New developments in the CIF standard will maintain our position at the leading edge of scientific information characterisation and exchange.

Complete documentation of CIF, the data exchange standard of crystallography, is in *International Tables for Crystallography Volume G: Definition and exchange of crystallographic data* (2005), edited by S. R. Hall and B. McMahon; corrected reprint published 2010 by John Wiley & Sons Ltd, Chichester, UK. It provides: information for scientists who wish to record or use the results of a single-crystal or powder diffraction experiment; the data ontology necessary for designing interoperable computer applications; accompanying software including machine-readable CIF dictionaries, libraries and utility programs.

Full documentation and freely downloadable dictionaries and software tools and libraries are also available at the IUCr web site http://www.iucr.org/resources/cif and, for macromolecular CIF (mmCIF), at http://mmcif.rcsb.org.

Welcome

Welcome to this COMCIFS Symposium on Crystallographic Information and Data Management. COMCIFS (the Committee for the Maintenance of the CIF Standard) was formed in 1993 as an IUCr Technical Committee charged with maintaining the definitions and software tools associated with the then new CIF (crystallographic information file) exchange and archival file format. Although COMCIFS has hosted a number of technical workshops (with significant consequences for CIF as it evolved) and Congress microsymposia, this is the first full-scale public meeting under its aegis.

This Symposium celebrates the advances in efficiency and accuracy of published crystallographic research that have taken place during the electronic era. Many – though by no means all – of these are due to the IUCr's commitment to standards and to CIF in particular. But of course the meteoric rise of the World Wide Web, the incredible advances during our lifetimes in computer hardware and information technology, and the astonishing quality of modern scientific equipment all play important roles too.

At the same time, economic and social changes challenge the way in which we are used to conduct scientific research and share the results of our discoveries. The use of bibliometric statistics to assess research impact and the consequent funding of research programmes can bias both the nature of the scientific problems that are investigated, and the balance of skills that can be called upon to develop a particular research area. Valuable work in areas such as service crystallography and informatics development may be under-resourced because their impact, on short timescales, is judged to be 'low'.

There is also the risk that excessive reliance on software analysis can lead relatively inexperienced researchers into egregious error; and there are worrying indications that some scientists, perhaps concerned about career prospects, are tempted towards falsification of their results. Vigilance, careful scrutiny and considered, critical analysis remain central elements of scientific method. It is essential not to lose our grasp of these fundamental qualities in the hectic pace of modern research and rapid communication, and under the sheer volume of data and information that we must handle in our routine lives as scientists.

These concerns must lie at the back of our minds as we settle back to enjoy today's programme. However, we can take some comfort – without, of course, feeling self-satisfied – that crystallography is particularly well placed to weather many of these current and forthcoming storms.

In our first session, **Standard information exchange formalisms**, we shall review the benefits that CIF has brought to three major areas of crystallography – small-unit-cell structure determination, biological macromolecular structures and powder diffraction. Small-molecule or inorganic crystallographers are very familiar with structural data sets in CIF format, and this is a standard well established across the different databases and journals in this field. Protein crystallographers are less familiar with CIF as a file format, because many of the community's working practices grew up around the already established PDB format. Nevertheless, the macromolecular CIF (mmCIF) data model forms the basis of the database schemas inplemented in the Protein Data Bank, and has been an effective means to promote interoperability with other data representations familiar in structural biology. The powder CIF (pdCIF) file is still not used universally in powder diffraction; but its more complicated data model reflects the complexity of the real-world structures that are often encountered in powder work. It is highly likely that this complexity will be further developed in new and exciting research into incommensurate structures, quasicrystals, nanocrystalline and other novel materials of potential technological impact.

While the emphasis of CIF in its first 20 years was on the description and validation of the derived structural model and the refinement of the structure factors or Rietveld profiles supporting that model, a more recent emphasis is further upstream, on the 'raw' data collected from the experiment, typically diffraction images, neutron counts, electron micrographs *etc*. The second session, **Improving the management of experimental data**, examines many of the current issues in storing, characterising and handling the increasing volumes of such data. There are technical challenges to face: how to optimise file size and data collection rates from very fast, very sensitive detectors; how to associate the raw measurements with a specific experiment, including the metadata that characterise that experiment and allow you to make any sense of the raw data. There are challenges for experimental facilities in the consistent handling of data sets from different type of instruments, associated with different types of experiments, for onward dissemination to very different communities of scientific users. There are policy challenges: how to ensure compliance with the data management mandates from funding organisations; how to manage security or free redistribution of different data sets; how much data to keep, and for how long; whether to facilitate – or mandate – access to experimental data as part of the publishing peer review process.

Our third session looks at established aspects of peer review for crystal structures. In **The integrity of published information** we review some of the ways in which crystallographic journals have improved the checking of crystal structures to reduce significantly the incidence of erroneous published structures. We also survey the tools provided by IUCr journals to facilitate the authoring of structure reports based on CIF – a data-friendly format that nevertheless is flexible enough to form the basis for article submissions. We also consider the added value in crystallographic publications of direct access to the underlying data sets as an integral part of the online publication.

The final session of the day, **Towards ever better science**, reminds us that the purpose of improving the quality of crystallographic information is not simply to polish an individual structure, but to make a real contribu-

tion towards scientific knowledge as a whole. Databases of high-quality structures form a more reliable basis for inference-driven scientific hypotheses based on data mining. Connecting different disciplines through interoperable machine-readable formalisms (whether CIF dictionaries, XML schemas or RDF ontologies) permits automated knowledge mining within and across subject boundaries. And CIF itself does not stand still. At the workshop immediately preceding this Symposium, we have been looking at the next generation of CIF dictionary, which will instil an even higher level of precision into the machine-readable definitions that form the heart of crystallography's data characterisation efforts.

CIF, of course, is not the complete answer to all problems in information and data management. No single format or formalism could be, and I am sure that will be clear from the discussions of real-world complexity that we will hear throughout the day. Nor is CIF perfect: I expect to hear many criticisms in the course of the day. COMCIFS will welcome such criticisms, because they will reflect the unmet needs of the community, and those needs will feed back into the continuing mandate of COMCIFS.

However, I hope that the day will give you some insight into the complexity of the issues that are faced by any effort to standardise, characterise and manage the high-information (= high-entropy!) description of natural phenomena. I hope the day will stimulate some of you to engage directly in this field – whether by contributing to CIF development and joining COMCIFS, by improving the input/output and exchange features of crystallographic software packages, by paying careful attention to the *significance* of checkCIF alerts, or simply by appreciating more deeply the power and potential of informatics.

In any case, I hope that *all* of you will enjoy this meeting.

Brian McMahon
Coordinating Secretary, COMCIFS

Timetable

9.00 am Introduction and welcome. James R. Hester

I. Standard information exchange formalisms

Chair: James R. Hester

9.05 am A coherent information flow in crystallography
Brian McMahon
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

9.30 am mmCIF and structural bioinformatics
John Westbrook
RCSB PDB, Rutgers University, Piscataway, NJ, USA

9.55 am pdCIF and the messy world of real data
Brian H. Toby
Advanced Photon Source, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439-4814, USA

10.20 am Coffee break

II. Improving the management of experimental data

Chair: Loes Kroon-Batenburg

10.40 am The data explosion and the need to manage diverse data sources in scientific research
Simon J. Coles
UK National Crystallography Service, Chemistry, Faculty of Natural and Environmental Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK

11.05 am Deposition and use of raw diffraction images
John R. Helliwell
School of Chemistry, University of Manchester, M13 9PL, UK

11.30 am Managing research data for diverse scientific experiments
Erica Yang
Scientific Computing, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Oxford, Didcot OX11 0QX, UK

11.55 am Managing crystallographic data in facilities using integrated CIF, HDF5 and NeXus
Herbert J. Bernstein
Dowling College, 1300 William Floyd Pkwy, Shirley, NY 11967, USA

12.20 pm Research data management and UK funding policies
Simon Hodson
CODATA, 5 Rue Auguste Vacquerie, 75016 Paris, France

12.45 pm Buffet lunch and soft drinks

III. The integrity of published information

Chair: John C. Bollinger

1.45 pm Publication of small-unit-cell structures in *Acta Crystallographica*
Michael A. Hoyland
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

2.10 pm Validating a small-unit-cell structure; understanding *checkCIF* reports
Anthony Linden
Institute of Organic Chemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

2.35 pm Writing a macromolecular structure paper with *publBio*
Manfred Weiss
Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography (HZB-MX), Albert-Einstein-Strasse 15, D-12489 Berlin, Germany

3.00 pm Deposition and validation of macromolecular structures at wwPDB
Sameer Velankar
EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SD, UK

3.25 pm Coffee break

IV. Towards ever better science

Chair: J. Mitchell Guss

3.45 pm Data quality and the value of structural databases
Colin Groom
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

4.10 pm Towards the semantic web of science
Peter Murray-Rust
Unilever Centre for Molecular Informatics, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK

4.35 pm Into the future with CIF
Nick Spadaccini
University of Western Australia, 35 Stirling Highway, Crawley, WA, Australia 6009

5.00 pm Close of Workshop

6.00 pm ECM28 Opening Ceremony

Abstracts

A coherent information flow in crystallography

Brian McMahon

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: bm@iucr.org

From its beginnings as a descriptive framework for structure determination experiments, CIF has grown to be a powerful framework for many aspects of crystallographic practice. It describes or is actively used in raw data collection (such as diffraction images), data reduction (*e.g.* as structure factors or Rietveld profiles), structure solution and refinement, and the publication, database curation and visualisation of crystal and molecular structures. Although not the only format in common use in the structural sciences, it is central to structure determination of small-unit-cell and macromolecular structures, and its ontologic approach informs the best practice of information and data handling throughout crystallographic researches.

Brian McMahon *is the Research and Development Officer at the International Union of Crystallography's offices in Chester, UK, and a former IUCr Representative to CODATA, the ICSU Committee on Scientific Data. He is Coordinating Secretary of COMCIFS and a Co-editor of International Tables for Crystallography Volume G: Definition and exchange of crystallographic data.*

---

mmCIF and structural bioinformatics

John Westbrook

RCSB PDB, Rutgers University, Piscataway, NJ, USA
Email: jwest@rcsb.rutgers.edu

A jewel in the crown of structural biology, the Protein Data Bank is an unparalleled archive of protein and nucleic acid structures. Its operation depends on information-rich relational databases, the schemas for which are efficiently expressed in the macromolecular CIF (mmCIF) formalism. Discrete entities that may be stored in the databases are characterised in CIF-format data dictionaries, which express ontologies widely used in areas as diverse as structural genomics, NMR, cryo-electron microscopy and protein production.

John Westbrook *is a Project Team Leader of the RCSB Protein Data Bank, and is based at Rutgers University. He has played key roles in creating and maintaining the PDB database schema, in developing many of the software tools that underpin PDB operation, and in developing formal ontologies with other structural biology communities. He is a member of COMCIFS.*

---

pdCIF and the messy world of real data

Brian H. Toby

Advanced Photon Source, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439-4814, USA
Email: Brian.Toby@ANL.GOV

Powder diffraction may well be the impoverished cousin of single-crystal diffraction, but for many materials and measurements there is no other choice; structural analyses have been performed from powders for materials ranging from simple salts to proteins. Many other types of material characterizations can be performed with powder diffraction. Excellent instruments are available that measure powder diffraction in different ways, depending on the needs of the study and the type of illumination source. Successful modeling requires that data be kept in a form close to the original measurement, which creates complexity for a powder CIF dictionary (pdCIF). It has been nearly 15 years since the initial pdCIF dictionary was completed and the successes and failures of pdCIF will be considered.

Brian Toby *has had extensive professional experience in the chemical industry, academic and government sectors, where the latter included both synchrotron and research reactor facilities. His research interests are in understanding how the arrangements of atoms in solids determine how the material functions, chemically or physically, and for the development and teaching of techniques for those studies. To do this, he works on software and instrument development as well as conducting measurements and analyzing the results. He has collaborated with researchers in academia, industry and government to produce 120 papers that have been cited nearly 6000 times. He is a former member of COMCIFS and was the leader of the powder CIF (pdCIF) dictionary development effort.*

Abstracts

The data explosion and the need to manage diverse data sources in scientific research

Simon J. Coles

UK National Crystallography Service, Chemistry, Faculty of Natural and Environmental Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK
Email: s.j.coles@soton.ac.uk

Crystal structure determination has become a high-throughput activity, and even at the level of the departmental laboratory, automation is increasingly important in managing the experiment and, crucially, the data collected from the experiment and its subsequent processing, analysis and dissemination. The UK National Crystallography Service is a medium-scale facility which needs to address issues of data management, accountability and dissemination, on top of its efforts to achieve best experimental practice. In providing a service to chemists as well as crystallographers, it has amassed considerable experience in cross-discipline ontology building, data publication *via* repository platforms, and integration with laboratory management systems.

Simon Coles *is Head of the UK National Crystallography Service and a member of staff of the Department of Chemistry at Southampton University.*

Deposition and use of raw diffraction images

John R. Helliwell

School of Chemistry, University of Manchester, M13 9PL, UK
Email: john.helliwell@manchester.ac.uk

The IUCr Executive Committee has charged a Working Group with an assessment of the potential benefits of depositing raw experimental data sets (with initial emphasis on X-ray diffraction images), and the cost, technical and structural ramifications of doing so. There are a number of potential locations for depositing raw images that allow their reuse in validation, re-refinements, reanalysis for new science, education and software development – for example, in discipline-specific data centres, in large-scale instrument facilities, or in institutional repositories. These are not necessarily exclusive (for example, a central data centre might archive only data sets associated with published structures), and initiatives such as the Australian TARDIS demonstrate approaches to federating separate repository platforms. Crucial to interoperability between such federated archives will be well-defined metadata and procedural standards.

John Helliwell *trained in physics and molecular biophysics and is now Emeritus Professor of Structural Chemistry at the University of Manchester. He is former Editor-in-Chief of the journals of the International Union of Crystallography and Past President of the European Crystallographic Association. His research involves crystallography methods developments applied to structural chemistry and biology. He is currently IUCr representative to CODATA (the ICSU Committee for Scientific Data) and ICSTI (the International Council for Scientific and Technical Information), and chairs the IUCr Diffraction Data Deposition Working Group. He is also a member of the CODATA/VAMAS Working Group on the description of nanomaterials.*

Managing research data for diverse scientific experiments

Erica Yang

Scientific Computing, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Oxford, Didcot OX11 0QX, UK
Email: erica.yang@stfc.ac.uk

The Rutherford Appleton Laboratory is a UK National Facility that handles a considerable amount of experimental research, and operates many instruments that generate large volumes of data such as the ISIS Pulsed Neutron Source, Central Laser Facility, and Diamond Light Source. It has developed a Core Scientific Metadata Model (CSMD) for the management of the data resources of the facilities in a uniform way. This talk discusses the emerging opportunities opened up by the availability of systematically managed data in a large laboratory setting and the technical barriers of making extensive use of them in real world open data infrastructure developments. The latest developments of CSMD will be presented to highlight the current direction we are undertaking in the context of the PaNData-ODI project, a collaborative European project involving thirteen major world class research laboratories that operate one or more neutron or photon sources in Europe.

Erica Yang *is a senior computer scientist at the STFC Rutherford Appleton Laboratory, where she works in the Scientific Computing Department (SCD). She is also the national labs services liaison officer, responsible for developing a sustainable long term data services research and development strategy and roadmap for STFC's national laboratories. She has worked with the UK National Crystallography Service and the University of Cambridge to develop cross-organisation data*

Abstracts

infrastructure technologies for the UK structural science communities. She also manages and directs projects involving STFC's facilities and large scale data and HPC infrastructures. In EU FP7 project 'PaNData-ODI', she works closely with international facilities (e.g. ESRF, ILL, DESY) to define and develop a fully integrated and cross-facility data management roadmap and services for EU photon and neutron communities.

Managing crystallographic data in facilities using integrated CIF, HDF5 and NeXus

Herbert J. Bernstein

Dowling College, 1300 William Floyd Pkwy, Shirley, NY 11967 USA
Email: yayahjb@gmail.com

A specific aspect of interoperability within large scientific facilities is the management of different data formats associated with different fields. NeXus, HDF5 and CIF are scientific data formats that overlap in some areas (*e.g.* in the capture and management of X-ray diffraction images), and there is considerable benefit to be gained by harmonising their content and working towards maximum interoperability. Format conversion at the syntactic level can be a straightforward process, but truly interoperable software applications require functional mappings between different representation standards. An initial approach to this is being attempted by constructing a DDL2 dictionary, acting as a concordance between the existing NeXus and imgCIF formats. Full interconvertibility between NeXus and imgCIF relies on methods evaluation, and may be possible using the new dictionary definition language DDLm and an evaluation engine (such as dREL).

Herbert Bernstein *is Professor of Mathematics and Computer Science at Dowling College, Oakdale, NY. He is a member of COMCIFS, Chair of the imgCIF dictionary working group, and lead developer of CIFtbx, a Fortran library for handling CIF data. He is also a member of the NeXus International Advisory Committee (NIAC).*

Research data management and UK funding policies

Simon Hodson

CODATA, 5 Rue Auguste Vacquerie, 75016 Paris, France
Email: ExecDir@codata.org

Although this meeting focuses on the technical aspects of sound data management, because we believe the benefits of this are self-evident to scientists who are used to taking the greatest care in analysing and using data, we should be aware also of the general research policy framework in which so much modern research is located. At the level of national and international science policy, funding bodies increasingly require evidence for best practice in data management; likewise, open-access mandates for publications may become a reality. This presentation contextualises the practice of our science within the evolving policy framework for publicly funded research.

Simon Hodson *is Executive Director of CODATA, http://www.codata.org, an organisation whose mission is to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use. He also sits on the Board of Directors of the Dryad data repository, http://datadryad.org, a not-for-profit initiative to make the data underlying scientific publications discoverable, freely reusable, and citable. From 2009 to 2013, as Programme Manager, he led two successive phases of Jisc's innovative Managing Research Data programme, http://researchdata.jiscinvolve.org.*

Publication of small-unit-cell structures in *Acta Crystallographica*

Michael A. Hoyland

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: mh@iucr.org

Structural journals published by the IUCr have an efficient workflow built around the CIF standard. For small-unit-cell structures, authors submit their articles in this format, building on files created directly by their structure solution/refinement software. The processed experimental data from which the structure has been determined (structure factors, Rietveld profiles) are also uploaded in this format, allowing stringent technical peer review of the quality of the structural modelling. All published structures are accompanied by the underlying data that would permit their redetermination, and can be explored in the online publication through three-dimensional visualisation and analysis tools.

Mike Hoyland *is a Systems Developer at the International Union of Crystallography, Chester, UK. He maintains the checkCIF web service for validating small-unit-cell structures, and is lead developer of the submission and review system for authors of IUCr journals.*

Abstracts

Validating a small-unit-cell structure; understanding *checkCIF* reports

Anthony Linden

Institute of Organic Chemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Email: alinden@oci.uzh.ch

When small structures are submitted for publication in IUCr journals, they are analysed as part of the peer review process by the software suite *checkCIF*. The goal of this analysis is to provide an objective assessment of the quality and reliability of the published structure, with reference both to the available experimental data and the level of consistency with known chemistry. For this to work well, validation protocols have to keep up with advances in structure determination methodologies. Consequently, it is important that CIF tools and definitions are both practical and extended and revised regularly. The purpose and output of validation also have to be understood easily by users, reviewers and journal editors, some of whom may not be expert crystallographers.

Tony Linden *is a Research Group Leader at the X-ray Crystallography Facility, Institute of Organic Chemistry, University of Zurich, and Section Editor of Acta Crystallographica Section C.*

Writing a macromolecular structure paper with *publBio*

Manfred Weiss

Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography (HZB-MX), Albert-Einstein-Strasse 15, D-12489 Berlin, Germany
Email: msweiss@helmholtz-berlin.de

For biological macromolecules, where structures and experimental data have traditionally been deposited in a central archive (the Protein Data Bank), integrating structural data with derivative publications is more complex. IUCr journals have developed an online publication tool that can extract deposited macromolecular data from the archive, and prompt the author for the additional information required for the fullest characterisation of a macromolecular structure determination. Again, the goal is to maximise the integrity of the scientific discussion.

Manfred Weiss *works at the Institute for Soft Matter and Functional Materials of the Helmholtz Zentrum Berlin, and is a Section Editor of Acta Crystallographica Section F. He has been a Co-editor of Acta Crystallographica Section D since 2002, is a member of the IUCr Commission on Crystallographic Teaching and a Consultant for the IUCr Commission on Biological Macromolecules.*

Deposition and validation of macromolecular structures at wwPDB

Sameer Velankar

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SD, UK
Email: sameer@ebi.ac.uk

Increasingly, the structural biology community is aware that the quality of structures deposited in the Protein Data Bank (PDB) needs critical assessment using informative, well-defined and community-accepted validation methods. To obtain recommendations on validation methods and criteria to be applied to newly deposited as well as existing structures in the PDB, the Worldwide Protein Data Bank (wwPDB) has convened several Validation Task Forces (VTFs). The validation methods and criteria suggested by the wwPDB X-ray VTF will be used to validate all X-ray structures deposited in the PDB as well as all crystal structures already in the archive. The talk will highlight the importance of validation and describe the implementation of the recommendations of the X-ray VTF. Salient features of the validation reports generated by the new wwPDB X-ray validation pipeline will also be discussed. It is hoped that all journals that publish biomacromolecular structure papers will make submission of these reports mandatory whenever they receive a manuscript describing a new structure.

Sameer Velankar *is a team leader at the EBI, responsible for content and integration of the Protein Data Bank in Europe (PDBe) resource. He earned his PhD in Structural Biology from the Indian Institute of Science in Bangalore, India in 1997, working on protein crystallographic studies of thymidylate synthase and triose phosphate isomerase. He then joined Dale Wigley's group in Oxford for post-doctoral research on the elucidation of the mechanism of DNA helicase. He has worked with and contributed to all parts of the PDBe team's operations, from annotation of newly deposited structures to the development of advanced PDBe services.*

Data quality and the value of structural databases

Colin Groom

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK
Email: groom@ccdc.cam.ac.uk

Abstracts

The Cambridge Structural Database of organic and metal–organic structures is one example of the curated archives of structural data that have been crucially important crystallographic and chemical resources for many decades. Many structural analyses based on the holdings in such databases rely on the quality and reliability of the individual structures. Historically, CCDC's Scientific Editors have worked hard to validate the individual depositions, a task greatly helped by the adoption of CIF as a community standard format. CCDC now applies probabilistic computational tools to determine the 'chemistry' represented by a CIF alongside expert human analysis.

Colin Groom *is the Executive Director of the Cambridge Crystallographic Data Centre. After a career in academia in the UK and in New Zealand, Dr Groom joined Pfizer where he established the protein crystallography group in the UK. He subsequently held various computational and informatics roles in the UK and US. Following this he joined Celltech/UCB, leading investigational chemistry and computer-assisted drug design groups.*

Towards the semantic web of science

Peter Murray-Rust

Unilever Centre for Molecular Informatics, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK
Email: pm286@cam.ac.uk

The development of CIF has allowed IUCr journals to capture the structured semantic content of research articles to a very high degree. Hyperlinking between components of the scholarly article (the text, structural data model, experimental data) provides a sound technical basis for mining, validating and reusing scientific data with the help of high-volume robotic tools.

Peter Murray-Rust *was until 2012 Reader in Molecular Informatics at the Unilever Centre for Molecular Informatics, University of Cambridge. He is co-developer with Henry Rzepa of Chemical Markup Language (CML) and a Consultant to COMCIFS.*

Into the future with CIF

Nick Spadaccini

University of Western Australia, 35 Stirling Highway, Crawley, WA, Australia 6009
Email: Nick.Spadaccini@uwa.edu.au

When CIF was adopted as an information exchange standard by the IUCr in 1991, it was, if anything, ahead of the then state of the art. Free-format, extensible, low-overhead – it offered an easy route to implementation by scientific software, much still written in Fortran. By externalising the semantics of its tags to external dictionaries, it allowed ontology development to be decoupled from format, and thus had an important role to play in interoperability between purely crystallographic applications and the wider worlds of structural biology, chemical informatics and laboratory data management, as illustrated elsewhere in this symposium. The latest enhancements to CIF, with the introduction of an even more powerful dictionary definition language that support methods definitions and multiple data models, provide unrivalled potential for taking computer ontologies into completely new territories.

Nick Spadaccini *is one of the main authors of the STAR File format, the STAR methods dictionary development language (DDLm) and the dictionary methods evaluator software dREL.*

Session Chairs

James Hester *is an Instrument Scientist at the Bragg Institute of the Australian Nuclear Science and Technology Organisation (ANSTO), a leading facility in the use of neutron scattering and X-ray techniques to solve complex research and industrial problems in many important fields. He is Chairman of COMCIFS.*

Loes Kroon-Batenburg *is a member of the crystallography group in the Laboratory of Crystal and Structural Chemistry of the University of Utrecht. Her current research projects focus on accurate data collection from X-ray diffraction using the EVAL method.*

John Bollinger *is Crystallographic Informatics Scientist at St Jude Children's Research Hospital, Memphis, Tennessee, one of the world's foremost paediatric treatment and research facilities. He is a member of COMCIFS.*

Mitchell Guss *is Professor of Structural Biology in the School of Molecular Bioscience of the University of Sydney. He was one of the Founding Editors of Acta Crystallographica Section F: Structural Biology and Crystallization Communications, of which he is still a Co-editor. He is a Member of the IUCr Executive Committee.*

A CIF primer

CIF is an acronym for the **Crystallographic Information Framework**, a set of data descriptors defining many application fields of crystallography. In practice, CIF is usually understood to mean the **Crystallographic Information File**, a particular implementation of the framework most familiar to small-molecule crystallographers, but also forming the basis for the database schemas of the Protein Data Bank and the Bilbao modulated structures database, the image format CBF which is becoming widely adopted for new detector installations, and a number of other applications.

The concrete implementation of the Crystallographic Information File comprises:

1 An ASCII free-format line orientated file type.

Discrete sets of information in 'data' blocks

Tags ('data names') begin with underscore

Data names have **one** associated value, *or*, . . .

. . . if collected with related data names and preceded by a ' loop_' directive, they act as the equivalent of table headings and thus identify a column of values

A given data name may occur at most *once* within a data block

```
data_I
_chemical_formula_moiety            'C17 H15 Cl O2 S'
_chemical_formula_weight                    318.80
_chemical_melting_point_gt                     432
_chemical_melting_point_lt                     433
_symmetry_cell_setting                   triclinic
_symmetry_space_group_name_H-M             'P -1'
loop_
    _symmetry_equiv_pos_site_id
    _symmetry_equiv_pos_as_xyz
           1      'x, y, z'
           2      '-x, -y, -z'
_cell_length_a                             6.043(2)
_cell_length_b                            11.716(4)
_cell_length_c                            12.217(5)
_cell_angle_alpha                         117.99(2)
_cell_angle_beta                           92.00(2)
_cell_angle_gamma                          97.42(2)
_cell_volume                              752.9(5)
```

2 Machine-readable dictionaries of data names.

Standard data names are defined in machine-readable dictionaries in essentially the same format as CIF data files. This allows software to be written that can validate certain attributes of values in data files according to the specifications in the dictionary.

Currently two variants of a dictionary definition language (DDL) are in use. The example uses DDL2, a standard that assigns **categories** to related data names (the data names themselves use a period character to indicate the category part of the name), and that enforces stronger constraints on data types and organisation than the more elementary DDL1.

```
save__cell.length_b_esd
    _item_description.description
;               The standard uncertainty (estimated
                standard deviation) of _cell.length_b.
;
    _item.name                    '_cell.length_b_esd'
    _item.category_id               cell
    _item.mandatory_code            no
    loop_
        _item_dependent.dependent_name
                                  '_cell.length_a_esd'
                                  '_cell.length_c_esd'
    _item_related.related_name    '_cell.length_b'
    _item_related.function_code     associated_value
    _item_sub_category.id           cell_length_esd
    _item_type.code                 float
    _item_units.code                angstroms
     save_
```

3 A machine-readable dictionary of the data names used in dictionaries.

The machine-readable definitions in the CIF dictionaries are organised using a base set of data names, that are in turn defined in *DDL dictionaries*. This hierarchy, *data file* > *data dictionary* > *DDL dictionary*, may be of little interest to most crystallographers, who are used only to the data files for exchanging data files between computer programs, or for submitting structure reports to journals. However, in principle it allows software developers to create validation tools that are ever more efficient at enforcing consistency and logical integrity of data sets in a growing range of subject areas.

```
save__item_range.minimum

    _item_description.name              '_item_range.minimum'
    _item_description.description
;    Minimum permissible value of a data item
     or the lower bound of a permissible range.
     ( minimum value <  data value)
;
    _item.name                          '_item_range.minimum'
    _item.category_id                     item_range
    _item.mandatory_code                  no
    _item_type.name                     '_item_range.minimum'
    _item_type.code                       any
     save_
```

4 Software tools, libraries and web services.

Many resources are available for software developers and for end users of CIF. The IUCr web site has links to many such network resources, or hosts archive copies of programs for: validating CIFs; manipulating the files in many ways; visualising structures or performing crystallographic calculations using CIF; and editing files or article submissions built on CIF data sets. See http://www.iucr.org/resources/cif/software

The web service *checkCIF* (http://checkcif.iucr.org) is freely available for validating small-molecule structural data sets.

The web service *printcif* (http://publcif.iucr.org/services/tools/printcif.php) is a free web service for creating a PDF and an interactive structure report, formatted according to a choice of styles. The styles reflect the article formats of IUCr journals as well as a generic format displaying all the publication-related experimental and geometric data content in the CIF.

The Protein Data Bank has an extensive collection of software tools for managing PDB data in mmCIF format, a DDL2-based CIF implementation for biological macromolecular crystallography. See http://mmcif.rcsb.org

The Image-supporting CIF (imgCIF) specification is used for portable diffraction images, and is normally implemented in an isomorphous binary file format, the Crystallographic Binary File (CBF). A library of ANSI-C functions for accessing imgCIF and CBF files is maintained on SourceForge (http://sourceforge.net/projects/cbflib).

A brief history of CIF

The acronym CIF is used both for the **Crystallographic Information File**, the data exchange standard file format of Hall, Allen & Brown (1991), and for the **Crystallographic Information Framework**, a broader system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations, including, but not restricted to, Crystallographic Information File and XML.

CIF was developed by the IUCr Working Party on Crystallographic Information in an effort sponsored by the IUCr Commission on Crystallographic Data and the IUCr Commission on Journals. CIF was adopted in 1990 as a standard file structure for the archiving and distribution of crystallographic information. It is now well established and is in regular use for reporting crystal structure determinations to *Acta Crystallographica* and other journals. It is often cited as a model example of integrating data and textual information for data-centric scientific communication.

As a granular, structured format, CIF was well suited to the telegraphic style of structure reports required by *Acta Crystallographica Section C: Crystal Structure Communications*, and was immediately adopted by IUCr journals as a submission medium for this journal. It soon became the mandatory submission format to *Acta C*, and the mandatory format for supplementary structural data, and subsequently for structure factors, for all IUCr journals.

Automated procedures could then be developed for checking the submitted structural data. This allowed routine technical assessment of all submitted structures (although manual validation was already carried out by many conscientious Co-editors), and the number of erroneous space-group determinations and other technical errors in structure determinations declined significantly. The validation procedures were subsequently developed into the web-based *checkCIF* service, supported and used by other publishers and available as a general resource to the community for independent validation and assessment of the technical quality of a structure determination.

Extensions of CIF were rapidly developed to describe powder diffraction, modulated structures and electron density studies. Other extensions have followed over the years, including, recently, a description of crystallographic restraints and constraints, and, currently under review, an extension for crystallographic twinning.

An ambitious project to extend CIF to the complete description of protein structure experiments and models resulted in major enhancements to the underlying data model. The resulting **mmCIF** format formed the basis for the PDB database of protein structures as refactored by the Research Collaboratory for Structural Biology (RCSB) in 1999. Deposit of structure factors to the PDB became possible with the adoption of an mmCIF-based format for experimental data. The PDBx exchange format continues to build on the original mmCIF dictionary and interfaces with extensions for non-crystallographic methods and procedures in protein structure determination and characterisation (Westbrook *et al.*, 2005). An XML format based on the same underlying data model is routinely used to maintain synchronicity between the international partners of the Worldwide Protein Data Bank.

Another important development, beginning in the late 1990s, has been the specification of **imgCIF** and its corresponding binary format **CBF** (the 'Crystallographic Binary File') for capturing and exchanging image data (Bernstein & Hammersley, 2005). This provides a common format across the diversity of detector manufacturers, and is currently under close consideration for its possible role in promoting strategies for the routine deposition of primary diffraction image data.

In 2006 the importance of CIF and the value of *checkCIF* were recognised by the Award for Publishing Innovation of the Association of Learned and Professional Society Publishers (ALPSP). In their report, the judges '*were impressed with the way in which CIF and checkCIF are easily accessible and have served to make critical crystallographic data more consistently reliable and accessible at all stages of the information chain, from authors, reviewers and editors through to readers and researchers. In doing so, the system takes away the donkeywork from ensuring that the results of scientific research are trustworthy without detracting from the value of human judgement in the research and publication process*'.

Research has been under way in recent years to develop a new formalism within the CIF framework for specifying data definitions with greater precision, and with machine-readable methods for expressing relations between distinct data items (Spadaccini & Hall, 2012). This will allow automated generation of derivable data that are absent from a particular file, provided all the relevant parent data are present. This new formalism is not intended in the short term to replace the existing CIF format in routine practice, but it does have the potential to provide a unifying computational framework for applications requiring CIF input from different subject areas.

References

Bernstein, H. J. & Hammersley, A. P. (2005). Specification of the Crystallographic Binary File (CBF/imgCIF), in *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data*, S. R. Hall & B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. pp. 37–43.

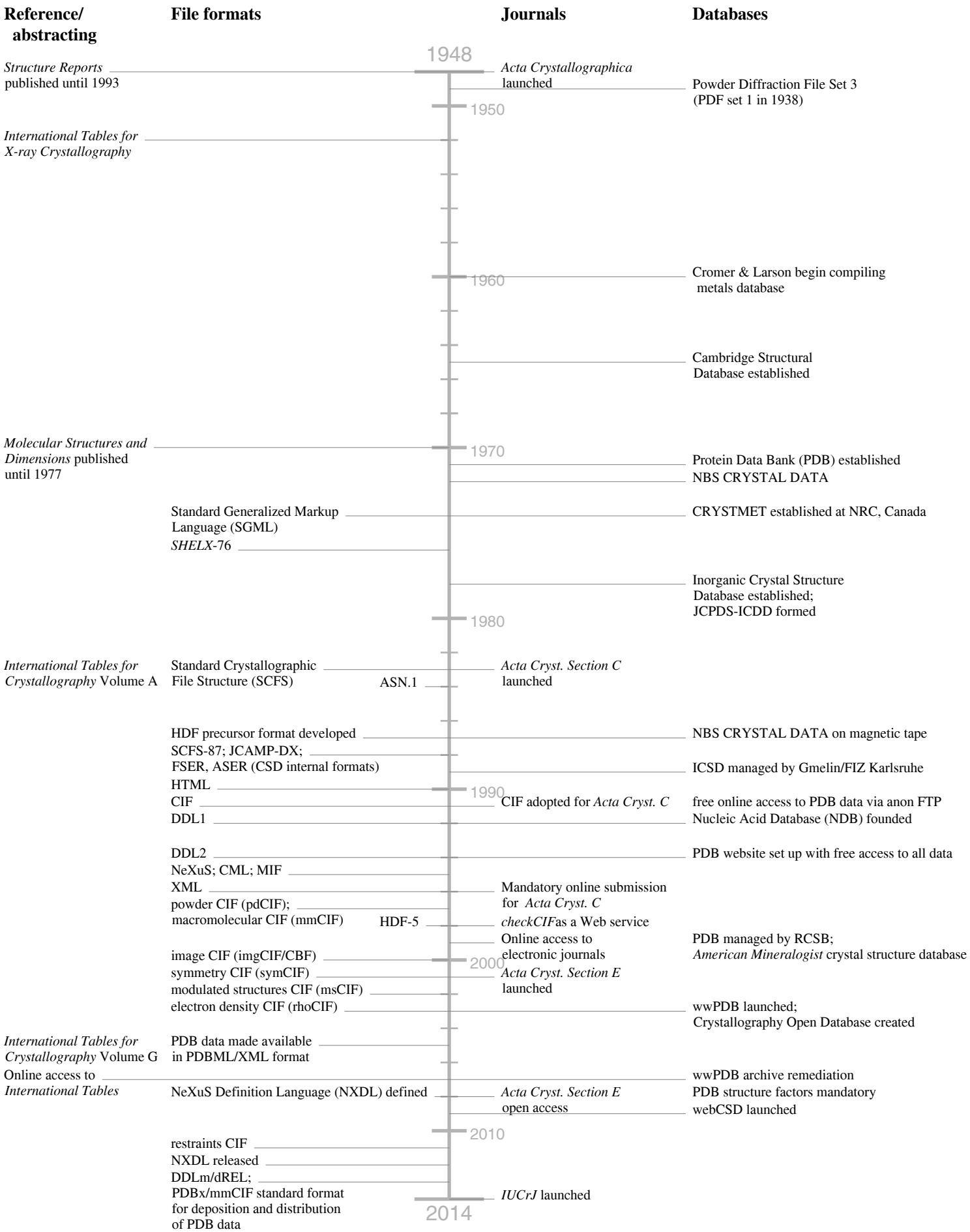
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography. *Acta Cryst.* A**47**, 655–685.

Spadaccini, N. & Hall, S. R. (2012). DDLm: a new dictionary definition language. *J. Chem. Inf. Model.* **52**, 1907–1916.

Westbrook, J., Yang,, H., Feng, Z. & Berman, H. M. (2005). The use of mmCIF architecture for PDB data management, in *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data*, S. R. Hall & B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. pp. 539–543.

Timeline of Crystallographic Information

An extended interactive version of this timeline is at http://www.iucr.org/resources/cif/comcifs/symposium-2013/timeline



Activities of the IUCr Diffraction Data Deposition Working Group 2011–2013

Terms of reference

It is becoming increasingly important to deposit the raw data from scattering experiments; a lot of valuable information gets lost when only structure factors are deposited. A number of research centres, *e.g.* synchrotron and neutron facilities, are fully aware of the need and have established detector working groups addressing this issue.

The IUCr is the natural organization to lead the development of standards for the representation of data and associated metadata that can lead to the routine deposition of raw data. A Working Group on these matters has thereby been launched by the IUCr Executive Committee, to which the Working Group will report, to be Chaired by Professor John R. Helliwell. Its provisional title is 'Diffraction Data Deposition Working Group of the IUCr'.

Discussion forum

Discussion forums were established on the IUCr web site (http://forums.iucr.org), including one for 'Public input on diffraction data deposition' to which all interested parties are invited to contribute (http://forums.iucr.org/viewforum.php?f=7). The DDD Forum has been very actively harnessed in collating comments at three different levels of input: the public at large, IUCr Officers including its Commissions, and the DDD WG itself. Documents from ICSU, CODATA and so on also feature.

Working group members: Steve Androulakis *Representative of TARDIS (Australian repositories for diffraction images)* ● Sol Gruner *Diffuse scattering specialist and Synchrotron Radiation Facility Director* ● John R. Helliwell, Chair. *IUCr Representative to CODATA and to ICSTI; Chair, IUCr Commission on Journals 1996–2005* ● Loes Kroon-Battenburg *Data processing software developer and user* ● Brian McMahon *Coordinating Secretary, COMCIFS* ● Tom Terwilliger *Representative of IUCr Commission on Biological Crystallography* ● John Westbrook *Representative of wwPDB (Worldwide Protein Data Bank)* ● Hans-Josef Weyer *Synchrotron Radiation and Neutron Facility user* ● **Consultants:** ● Alun Ashton *Diamond Light Source* ● Herbert Bernstein *Head, imgCIF Dictionary Maintenance Group and member of COMCIFS* ● Frances Bernstein *Observer on data deposition policies* ● Gerard Bricogne *Active software and methods developer* ● Bernhard Rupp *Macromolecular crystallographer*

Report on the Bergen Workshop

This is an edited version of the full report by John Helliwell, Tom Terwilliger and Brian McMahon that appears at http://forums.iucr.org/viewtopic.php?f=21t=102

A full-day Workshop was organised by the DDD WG as a Satellite of the ECM27 meeting in Bergen, Norway, on August 6th 2012. Its purpose was to review progress during the Working Group's first year of activity, and to help frame a policy to be drafted by the IUCr DDD WG on raw diffraction data deposition for final approval by the IUCr Executive Committee.

Presentations on the following topics are available at http://www.iucr.org/resources/data/dddwg/bergen-workshop

- The IUCr Diffraction Data Deposition Working Group Activities since IUCr Madrid *J. R. Helliwell and Brian McMahon*
- Motivations, challenges, horror stories and opportunities: Experiences of diffraction data management, archival and publication at the UK National Crystallography Service. *Simon J. Coles*
- Report on several important EU projects: CRISP, PaNdata, NMI3, Biostruct X, HDRI and CALIPSO *Heinz J. Weyer*
- Linking raw experimental data with scientific workflow and software repository: some early experience in the PanData-ODI project *Erica Yang and Brian Matthews*
- Ten years and change: the MX data archive at ALS 8.3.1 *James Holton*
- Continuous improvement of macromolecular crystal structures *Thomas C. Terwilliger*
- Towards policy for archiving raw data for macromolecular crystallography: Recent experience *Loes M. J. Kroon-Batenburg, Antoine M. M. Schreurs, Simon W. M. Tanley and John R. Helliwell*
- Some Economic Considerations for Managing a Centralized Archive of Raw Diffraction Data *John Westbrook*
- A vision involving raw data archiving via local archives as a supplement to the existing processed data archives (PDB, CSD, ICDD *etc.*) *John R. Helliwell, Brian McMahon and Thomas C. Terwilliger*

The need to have clarity on DDD issues has two main aspects. First, crystallographers have obligations to securely and properly retain the raw data that they measure ('loss of data is viewed as research malpractice'). Second, the reader of a published article involving crystallography can and should have access to the raw data on which the article is based ('don't take my word for it; try the data yourself and see directly the research results').

The actions and recommendations arising from the Bergen Workshop can be summarised as follows:

- The Workshop noted that there is an enthusiasm and encouragement to archive more than derived or processed data in many areas of science besides our own.
- The crystallographic community prides itself in making its processed data accompany its publications; indeed, this has been obligatory in IUCr journals for over 10 years.
- We, the crystallographic community, basically have three practical options in the near future to extend these principles to our raw data;
- via a local Data Archive
- via synchrotron or neutron or X-ray laser (or other large-scale experimental facility) data storage
- or via the corresponding author setting up a personal link to datasets underpinning publications on their personal websites. [At the Workshop the Protein Data Bank (John Westbrook) offered that the PDB would help to coordinate DOI registration in cases where the raw data could be hosted on a reliable public site.]
- So we suggest that we encourage all three practical options and recommend to the IUCr Executive Committee that:

– *Authors should provide a permanent and prominent link from an article to the raw data sets underpinning a journal publication*, with a view to making this a formal requirement on authors at such time as the community has adopted raw data deposition as a routine procedure.

Post meeting note; The IUCr Executive Committee endorsed this proposal from the DDD WG but replaced the word 'should' by 'may'. This is indeed still a positive step forward as it endorses the commitment of resources, for example by IUCr Journals, in assisting authors with this. See, for example, the article of Tanley et al. (2013) cited below.

There is an urgent need to be clear about the metadata required for the types of experiment and their raw data. John Westbrook of the RCSB stressed the importance of this: 'if the metadata details required are not standardised then there will be datasets which are nothing more than a mess and which would not be effectively usable by someone retrieving [them]'.

Some recent publications attempt to describe in detail the technical metadata required. (1) A protein crystallography article entitled 'Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals' by Simon W. M. Tanley, Antoine M. M. Schreurs, John R. Helliwell and Loes M. J. Kroon-Batenburg. *J. Appl Cryst.* (2013), **46**, 108–119. (2) An article defining data formats in X-ray absorption spectroscopy entitled 'Towards data format standardization for X-ray absorption spectroscopy' by B. Ravel, J. R. Hester, V. A. Solé and M. Newville. *J. Synchrotron Rad.* (2012), **19**, 869–874. (3) Data and metadata definitions have been published also for SAXS and SANS: 'Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution' by D. A. Jacques, J. M. Guss, D. I. Svergun and J. Trewhella. *Acta Cryst.* (2012), D**68**, 620–626.

While IUCr Commissions need to specify 'technical' metadata – *i.e.* those specific to their experimental raw data – there is also a need to review 'generic' metadata – *e.g.* who 'owns' a data set, details of research grants, embargo periods etc. A higher-level classification of the domain of study may be needed. *E.g.*, a synchrotron facility might need to define different data storage policies for, say, X-ray diffraction images versus X-ray tomography images. Such policies could be automatically implemented if data sets had characteristics identifying what sort of scientific study they represent. We feel that it would be beneficial to form a specialist group analysing these requirements. Members of this sub-group would be specialists able to represent different subject areas and experimental facilities. It would probably be a sub-group of the IUCr DDD WG.

One way to encourage satisfactory clarification of metadata technical definitions and standards is for the IUCr Executive Committee to require its Commissions to provide metadata recommendations as soon as possible.

An exemplar of good practice demonstrating access to raw data is at the ISIS UK neutron source. In the workshop Dr Erica Yang showed an example STFC DOI landing page for a particular data set and discussed the ISIS data management policy, from which we highlight a couple of points: • *[5.4] PIs and researchers who carry out analyses of raw data and metadata are encouraged to link the results . . . with the raw data / metadata using the facilities provided by the on-line catalogue. Furthermore, they are encouraged to make such results publicly accessible.* • *[3.3.3] Access to raw data and the associated metadata . . . from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible . . . [unless a PI can] make a special case to the Director of ISIS.*

In the Workshop, discussion of the concept of 'The Living Publication' led to the suggestion that journals need a new category name for articles stemming from a 'starter article and data set'. '*Ad extensum*' was offered as a suggested article type name. Subsequent investigation revealed that a mechanism already exists in the electronic publishing world to let the reader know that an article is related to other articles and data, and this can be accompanied by metadata explaining the relationships. So perhaps there is no need for an '*Ad extensum*' designation; but for derivative articles there could instead be an agreed set of metadata across publishers. Specifically, the **CrossMark** scheme of CrossRef, an organisation defining such standards across publishers, is an example of an attempt to get publishers to collaborate in this way.

Next step: special articles

A short series of articles has been commissioned to appear in *Acta Crystallographica Section D: Biological Crystallography* to bring some of the relevant issues to a wider community.

Tentative authors and titles for this series are as follows:

- Gerard Bricogne – Why deposition of diffraction data is important
- James Holton – What data should be deposited for macromolecular crystallography?
- John Westbrook – Practicalities of storage and deposition of image data
- John Helliwell and Loes Kroon-Batenburg – Experience with making image data available. What metadata do we need to archive?
- Tom Terwilliger and Gerard Bricogne – Continuous improvement of macromolecular structures
- Mitchell Guss and Brian McMahon – How to make deposition of images a reality

Publication is expected in late 2013 or early 2014.

Run up to IUCr Congress, Montreal

In the run up to the IUCr Congress 2014 we anticipate further progress by IUCr Commissions in clarifying their metadata needs to accompany raw data relevant to them. Secondly the proactive efforts of authors at 'grass roots' level and the IUCr Executive at 'top down' level should help contribute to making available raw data in general (and diffraction data images in particular). Initiatives of this type are likely to be increasingly appropriate in the 'open access' era, which extends beyond the written word to the data that form the firm platform on which science is based. Raw data availability will be a natural extension to our existing practice, over several decades, of making available in an organised way processed data (structure factors) and derived data (molecular coordinates).

Further reading

IUCr publishing activities

Abrahams, S. C. & Matula, R. A. (1988). Crystallographic publishing in retrospect and prospect. *Acta Cryst.* A**44**, 401–410.
Kamminga, H. (1989). The International Union of Crystallography: its formation and early development. *Acta Cryst.* A**45**, 581–601.
Cruickshank, D. W. J. (1998). Aspects of the History of the International Union of Crystallography. *Acta Cryst.* A**54**, 687–696.
Strickland, P. R. & McMahon, B. (2008). Crystallographic publishing in the electronic age. *Acta Cryst.* A**64**, 38–51.

The Crystallographic Information File

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.
Bourne, P., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). Macromolecular Crystallographic Information File. *Methods Enzymol.* **277**, 571–590.
Hall, S. R. & McMahon B. (eds) (2005). *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data,* Dordrecht: Springer. Corected reprint (2010). Chichester: Wiley.

General articles about CIF

Hall, S. R. (1998). Data Languages and Dictionaries for Crystallography. *Acta Cryst.* A**54**, 820–832.
Westbrook, J. D. & Bourne, P. E. (2000). STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics*, **16**, 159–168.
Brown, I. D. & McMahon, B. (2002). CIF: the computer language of crystallography, *Acta Cryst.* B**58**, 317–324.
Brown, I. D. & McMahon, B. (2006). The Crystallographic Information File (CIF). *Data Science Journal* **5**, 174–177.

Reference papers for the STAR File format, DDL and dREL

Hall, S. R. (1991). The STAR file: a new format for electronic data transfer and archiving. *J. Chem. Inf. Comput. Sci.* **31**, 326–333.
Hall, S. R. & Cook, A. P. F. (1995). STAR dictionary definition language: initial specification. *J. Chem. Inf. Comput. Sci.* **35**, 819–825.
Hall, S. R. & Spadaccini, N. (1994). The STAR File: detailed specifications. *J. Chem. Inf. Comput. Sci.* **34**, 505–508.
Spadaccini, N. & Hall, S. R. (2012). Extensions to the STAR File Syntax. *J. Chem. Inf. Model.* **52**, 1901–1906.
Spadaccini, N. & Hall, S. R. (2012). DDLm: A New Dictionary Definition Language. *J. Chem. Inf. Model.* **52**, 1907–1916.
Spadaccini, N., Castleden, I. R., Du Boulay, D. & Hall, S. R. (2012). dREL: A Relational Expression Language for Dictionary Methods. *J. Chem. Inf. Model.* **52**, 1917–1925.

Semantic physical science

Murray-Rust, P. (1998). The Globalization of Crystallographic Knowledge. *Acta Cryst.* D**54**, 1065–1070.
Murray-Rust, P. & Rzepa, H. S. (2012). Semantic Physical Science. *J. Cheminformatics*, **4**:14
McMahon B. (2012). Appllied and implied semantics in crystallographic publishing. *J. Cheminformatics*, **4**:19

Definition of essential metadata for primary experimental data sets

Ravel, B., Hester, J. R., Solé, V. A. & Newville, M. (2012). Towards data format standardization for X-ray absorption spectroscopy. *J. Synchrotron Rad.* **19**, 869–874.
Tanley, S. W. M., Schreuers, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals. *J. Appl Cryst.* **46**, 108–119.
Jacques, D. A., Guss, J. M., Svergun, D. I. & Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. *Acta Cryst.* D**68**, 620–626.

Crystallographic Databases

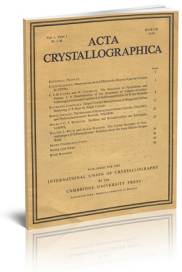
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst.* B**58**, 380–388.
Berman, H. M., Henrick, K. & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.* **10**, 98.
Aroyo, M. I., Perez-Mato, J. M., Capillas, C., Kroumova, E., Ivantchev, S., Madariaga, G., Kirov, A. & Wondratschek, H. (2006). Bilbao Crystallographic Server I: Databases and crystallographic computing programs. *Z. Krist.* **221**, 15–27.
Aroyo, M. I., Kirov, A., Capillas, C., Perez-Mato, J. M. & Wondratschek, H. (2006). Bilbao Crystallographic Server II: Representations of crystallographic point groups and space groups. *Acta Cryst.* A**62**, 115–128.
Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). Crystallography Open Database - an open-access collection of crystal structures. *J. Appl. Cryst.* **42**, 726–729.

The journals of the International Union of Crystallography

Publication of its own research journals was an integral part of the core mission of the International Union of Crystallography, and *Acta Crystallographica* dates from the very beginnings of the Union. Over the last 60 years, the portfolio of publications has expanded to keep pace with the growing field, in the spirit of its Founding Editor's first editorial:

Acta is intended to offer a central place for publication and discussion of all research in this vast and ever-expanding field. It borders, naturally, on pure physics, chemistry, biology, mineralogy, technology and also on mathematics, but is distinguished by being concerned with the methods and results of investigating the arrangement of atoms in matter, particularly when that arrangement has regular features.

Editor-in-Chief: S. S. Hasnain

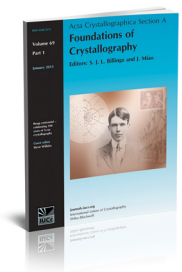


Acta Crystallographica

Founded 1948 Founding Editor: P. P. Ewald

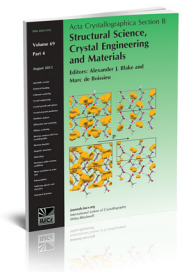
IUCrJ

Launching in 2014 Main Editors: E. N. Baker, G. R. Desiraju, C. R. A. Catlow, S. Larsen, J. C. H. Spence



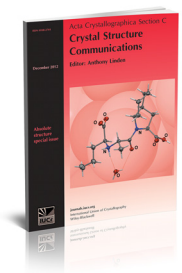
Acta Crystallographica Section A: Foundations of Crystallography

Launched 1968 Section Editors: S. J. L. Billinge, J. Miao



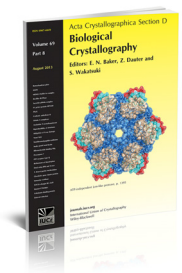
Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials

Launched 1968 Section Editors: A. J. Blake, M. de Boissieu



Acta Crystallographica Section C: Crystal Structure Communications

Launched 1983 Section Editor: A. Linden

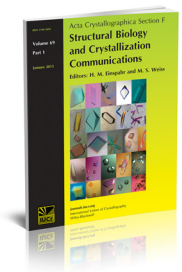


Acta Crystallographica Section D: Biological Crystallography

Launched 1993 Section Editors: E. N. Baker, Z. Dauter, S. Wakatsuki

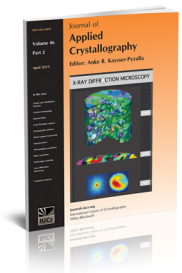
Acta Crystallographica Section E: Structure Reports Online

Launched 2001 Section Editors: W. T. A. Harrison, H. Stoeckli-Evans, E. R. T. Tiekink, M. Weil



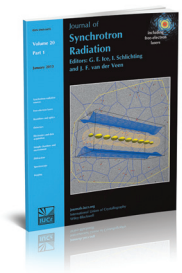
Acta Crystallographica Section F: Structural Biology and Crystallization Communications

Launched 2005 Section Editors: H. M. Einspahr, M. S. Weiss



Journal of Applied Crystallography

Launched 1968 Editor: A. R. Kaysser-Pyzalla



Journal of Synchrotron Radiation

Launched 1994 Editors: G. E. Ice, I. Schlichting, J. F. van der Veen

From the beginning of 2014, all journals of the IUCr will be published online only

About our sponsors

We acknowledge the generosity of many corporate sponsors who have made possible this Symposium and associated technical Workshop. The Workshop focussed on the next generation of CIF dictionaries, based on a new dictionary definition language that supports computational methods. Successful development of this new approach opens the possibility of generic data validation software, which can analyse the self-consistency of *any* data set and generate missing derivable values using only the relevant machine-readable data dictionaries.

Many of the companies and organizations below will be familiar to an audience of crystallographers. They provide the tools of our trade, in laboratory equipment, instrumentation or databases. The appearance of other names as sponsors of activities at a crystallography conference may be more surprising. Their inclusion demonstrates the importance of the orderly dissemination of scientific information, not only within a single community, but across all fields of science, and as a general contribution to global knowledge. We are grateful to them all.

The **International Union of Crystallography** 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.

The **University of Western Australia** (UWA) is recognised internationally as an excellent research-intensive university. Established in 1911, the University's ground-breaking research, quality academic staff and state-of-the-art facilities combine to offer a vibrant student experience.

The **Worldwide Protein Data Bank** (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for Protein Data Bank (PDB) data. Members are: RCSB PDB (USA), PDBe (Europe) and PDBj (Japan), and BMRB (USA). The wwPDB's mission is to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

The **Cambridge Crystallographic Data Centre** (CCDC) is dedicated to the advancement of chemistry and crystallography for the public benefit through providing high quality information, software and services.

Chemists in academic institutions and commercial operations around the world rely on the CCDC to deliver the most comprehensive and rigorous molecular structure information and powerful insights into their research.

The CCDC is a non-profit organisation and a registered charity, supported entirely by software subscriptions from its many users. The CCDC compiles and distributes the Cambridge Structural Database (CSD), the world's repository of experimentally determined organic and metal–organic crystal structures. It also develops knowledge bases and applications which enable users quickly and efficiently to derive huge value from this unique resource.

The **Digital Curation Centre** (DCC) is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management across the UK's higher education research community.

The Digital Curation Centre provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data.

The DCC provides access to a range of resources including our popular How-to Guides, case studies and online services. Our training programmes aim to equip researchers and data custodians with the skills they need to manage and share data effectively.

We also provide consultancy and support with issues such as policy development and data management planning.

The **British Library** is the national library of the United Kingdom. Its mission is: Advancing the world's knowledge.

Its vision: In 2020, the British Library will be a leading hub in the global information network, advancing knowledge through our collections, expertise and partnerships, for the benefit of the economy and society and the enrichment of cultural life.

About our sponsors

CODATA, the Committee on Data for Science and Technology, is an interdisciplinary Scientific Committee of the International Council for Science (ICSU), established in 1966 to promote and encourage, on a world-wide basis, the compilation, evaluation and dissemination of reliable numerical data of importance to science and technology.

The mission of CODATA is to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use.

It works to improve the quality, reliability, management and accessibility of data of importance to all fields of science and technology. CODATA provides scientists and engineers with access to international data activities for increased awareness, direct cooperation and new knowledge. It is concerned with all types of data resulting from experimental measurements, observations and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science, ecology and others. Particular emphasis is given to data management problems common to different disciplines and to data used outside the field in which they were generated.

WILEY

Wiley's Scientific, Technical, Medical, and Scholarly (STMS) business, also known as Wiley-Blackwell, serves the world's research and scholarly communities, and is the largest publisher for professional and scholarly societies. Wiley-Blackwell's programs encompass journals, books, major reference works, databases, and laboratory manuals, offered in print and electronically. Through Wiley Online Library, we provide online access to a broad range of STMS content: over 4 million articles from 1,500 journals, 9,000+ books, and many reference works and databases. Access to abstracts and searching is free, full content is accessible through licensing agreements, and large portions of the content are provided free or at nominal cost to nations in the developing world through partnerships with organizations such as HINARI, AGORA, and OARE.

RSC | Advancing the Chemical Sciences

The **Royal Society of Chemistry** is the largest organisation in Europe for advancing the chemical sciences. Supported by a worldwide network of members and an international publishing business, our activities span education, conferences, science policy and the promotion of chemistry to the public.

Elsevier is a global provider of science and health information, headquartered in Amsterdam. Its mission statement declares that Elsevier is committed to making genuine contributions to the science and health communities by providing:

World class information: Elsevier publishes trusted, leading-edge Scientific, Technical and Medical (STM) information – pushing the frontiers and fueling a continuous cycle of exploration, discovery and application.

Global dissemination: Elsevier disseminates and preserves STM literature to meet the information needs of the world's present and future scientists and clinicians – linking thinkers with ideas.

Innovative tools: Elsevier develops electronic tools that demonstrably improve the productivity and outcomes of those we serve – we are dedicated to helping them make a difference.

Working together: Elsevier works in partnership with the communities we serve to advance scholarship and improve lives. This interrelationship is expressed in our company's Latin motto, Non Solus, 'not alone'.

ACS Chemistry for Life®

The Publications Division of the **American Chemical Society** provides the worldwide scientific community with a comprehensive collection of journals in the chemical and related sciences, publishing more than 40 journals.

The ACS Publications Division supports the ACS Vision and its complementary Mission Statement:

Improving people's lives through the transforming power of chemistry to advance the broader chemistry enterprise and its practitioners for the benefit of Earth and its people.

About our sponsors

Dectris is the technology leader in X-ray detection. The DECTRIS photon counting detectors have transformed basic research at synchrotron light sources, as well as in the laboratory and with industrial X-ray applications. DECTRIS aims to continuously improve measurement quality, thereby enabling new scientific findings. This pioneering technology is the basis of a broad range of products, all scaled to meet the needs of various applications. DECTRIS also provides solutions for customer developments in scientific and industrial X-ray detection.

DECTRIS was awarded the 2010 Swiss Economic Award in the High-Tech Biotech category, the most prestigious prize for start-up companies in Switzerland.

The **Crystallography Open Database** is a project that aims to collect inorganic, metal–organic and small organic molecule structural data into a single database, available through an open-access model. The COD currently contains close to a quarter of a million entries in CIF format.

Since its inception in 1951, **Rigaku** has been at the forefront of analytical and industrial instrumentation technology. Today, with hundreds of major innovations to their credit, the Rigaku Group of Companies are world leaders in the fields of general X-ray diffraction (XRD), thin film analysis (XRF, XRD and XRR), X-ray fluorescence spectrometry (TXRF, EDXRF and WDXRF), small angle X-ray scattering (SAXS), protein and small molecule X-ray crystallography, Raman spectroscopy, X-ray optics, semiconductor metrology (TXRF, XRF, XRD and XRR), laboratory automation, X-ray sources, computed tomography, nondestructive testing and thermal analysis.

Crystal Impact's basic goal is to develop high quality software which allows even non-specialist users to apply most recent scientific and software technologies. Key areas of activity are crystal structure solution, visualization, and phase identification. Chemists and material scientists from industry and academic institutions in 57 countries all over the world use Crystal Impact's innovative software tools to determine, visualize and understand the crystal structures of their compounds.

Crystal Impact started life as a project based in the University of Bonn, where G. Bergerhoff's research group established the ICSD (Inorganic Crystal Structure Database) in 1983. Researching new methods for detection of systematically hidden patterns in structural databases, we collected a great lot of experiences in crystallographic programming. In the early 90s we developed completely new retrieval and visualization software for the ICSD, which was awarded an international software prize in 1993.

Bruker Corporation has been driven by the idea to always provide the best technological solution for each analytical task for more than 50 years.

Today, worldwide more than 6,000 employees are working on this permanent challenge at over 90 locations on all continents. Bruker systems cover a broad spectrum of applications in all fields of research and development and are used in all industrial production processes for the purpose of ensuring quality and process reliability.

Bruker continues to build upon its extensive range of products and solutions, its broad base of installed systems and a strong reputation among its customers. Being one of the world's leading analytical instrumentation companies, Bruker is strongly committed to further fully meet its customers' needs as well as to continue to develop state-of-the-art technologies and innovative solutions for today's analytical questions.

The **International Year of Crystallography 2014** (IYCr2014) commemorates not only the centennial of X-ray diffraction, which allowed the detailed study of crystalline material, but also the 400th anniversary of Kepler's observation in 1611 of the symmetrical form of ice crystals, which began the wider study of the role of symmetry in matter.

The timeline of crystallographic information that appears in this programme booklet will be developed as part of the interactive timelines project to be hosted on the IYCr2014 web site iycr2014.org.