Challenges and opportunities in curating one million crystal structures

Amy A. Sarjeant, Suzanna C. Ward, Ian J. Bruno
American Crystallographic Association Annual Meeting, Cincinnati, OH / Covington KY
Sunday July 21st 2019
Outline

• Brief historical perspective
• Modern curation practices
• Making data findable and accessible
• Challenges of ensuring data integrity
• Opportunities to educate and assist
• Where do we go from here?
The CCDC

CCDC is the home of small molecule crystallography data and leader in software for pharmaceutical discovery, materials development, research and education.

- Originated in the Department of Chemistry at the University of Cambridge
- Fully independent non-profit institution & a registered charity
- >50 years of scientific expertise
- More than 750 peer-reviewed publications
The Vision

We clearly recognised even in those early days, that data banks have three principal functions. Firstly they must gather together existing knowledge and make it readily available to the scientific community. Secondly they can be used to reduce a large number of observations to a small set of constants and rules, and in this way transform a data base to a knowledge base. Such a knowledge base may obviate the need for further individual experiments in specific areas. Thirdly, they facilitate the comparison and collective analysis of individual results to gain insight into new or as yet unexplained phenomena. These ideas have been at the heart of the work of the Cambridge Crystallographic Data Centre and the driving force for improving methods of data collection, storage and dissemination. Most importantly they influenced development of computer programs and methodologies which are needed for the analysis and transformation of the accumulated information. (5)
Historical Perspective on Crystallographic Data Sharing

- Crystallographic Journals
  - Z. Krystallog. (1877)
  - Acta Cryst. (1948)

- Crystal structure databases
  - JCPDS/ICDD (1941)
  - CCDC (1965)
  - PDB (1971)

- Data Analysis and Archival
  - ACA SIG (2018)
  - Transactions Symposium (2019)
The early days of data sharing

- Hand-drawn structures
- Tabulated data as main text or SI
- As few details as possible
  - Space is limited
- Full experimental details not shared

Modern data management

- CIF File – everything “important” about the experiment
  - Chemistry / Sample Characteristics
  - Geometry / Connectivity
  - Experimental / Refinement
  - Publication Information
  - [structure factors]
  - [refinement instructions]

- Institutional archives
- Curated databases...

The Cambridge Structural Database (CSD)

One million structures and counting

Suzanna Ward

Monday July 22\textsuperscript{nd} 2019
Room: Ballroom B
1:30 PM - 1:55 PM
Curating the CSD

- Each dataset expertly curated
- Datasets enhanced
  - Chemical connectivity
  - Compound names
  - 2D chemical diagrams
  - Additional experimental data
  - Bibliographic information
Curation and Chemistry Assignment
Making data findable and accessible
Published data

Sources of Publication Information:
- Pre-publication metadata communicated by journal publisher feeds
- Full publication metadata communicated and updated by journal publisher feeds

Manual CCDC Actions:
- Reviewing publication details
- Publication information updated via journal scanning by CCDC staff
- Publication information communicated by researchers wanting to access data
Published data

Pre-publication metadata communicated to CCDC by journal publishers

Full publication metadata communicated to CCDC by journal publishers

Publication information updated via journal scanning by CCDC staff

Data Publication

Pre-publication metadata added to data record

Full publication metadata added to data record

CCDC

Data Access

Data Deposition
(Ingest)

Data Publication

Data Curation

Metadata and terms of deposition confirmed by the depositor

CIF, HKL and FCF data deposited via the CCDC deposition and validation service

Depositor's responses to checks and checkCIF reports added to deposit record.

CCDC checks run on deposited files:
- Structure factor check
- IUCr checkCIF
- Unit Cell Check

Data deposited by data producer

Data stored for long-term preservation

CCDC automatic validation and duplicate check of data files

Does data pass automatic validation?
- Yes
- No

Email sent to depositor requesting additional information/data

Pre-publication metadata communicated to CCDC by journal publishers

Full publication metadata communicated to CCDC by journal publishers

Publication information updated via journal scanning by CCDC staff

Data Publication

Pre-publication metadata added to data record

Full publication metadata added to data record

Full publication metadata added to data record

Publication information updated via journal scanning by CCDC staff

Does data pass manual validation?
- Yes
- No

Email sent to depositor requesting additional information/data

Pre-publication metadata communicated to CCDC by journal publishers

Full publication metadata communicated to CCDC by journal publishers

Publication information updated via journal scanning by CCDC staff

Data Publication

Pre-publication metadata added to data record

Full publication metadata added to data record

Full publication metadata added to data record

Publication information updated via journal scanning by CCDC staff

Does structure meet criteria for curation into the CSD?
- Yes
- No

Data transferred to FIZ for curation into the ICSD

Published data
Links from Articles to CCDC Data


Footnote

This journal is © The Royal Society of Chemistry 2009
Making data accessible though the CSD

Data not published in a scientific journal can be curated into the CSD and made available to the community as a CSD Communication.

Structures from your PhD thesis can be made publicly available through the CSD.

https://www.ccdc.cam.ac.uk/Community/csd-communications/
Provenance and attribution
Joint CSD and ICSD Services

Over 180,000 entries from the Inorganic Crystal Structure Database (ICSD) now available through Access Structures
Accessing data from general resources

Links to specialist resources
The challenge of data integrity
A community responsibility

• Fraud in the crystallographic community
• Increased vigilance
• Increased sharing

Regrettably, this editorial is to alert readers and authors of Acta Crystallographica Section E and the wider scientific community to the fact that we have recently uncovered evidence for an extensive series of scientific frauds involving papers published in the journal, principally during 2007. Although several thousands of structures published in Acta Crystallographica Section E every year will continue to reflect results of serious scientific work, the extent of these problems is significant with at least 70 structures demonstrated to be falsified and meanwhile acknowledged by the authors as such. Our work is ongoing and it is likely that this figure will rise further.

From Data? to Knowledge?

Experiment

\[ C_{10}H_{12}N^+, Cl^- \]

Knowledge

CCDC
Maintaining Data Integrity

- **Integrity** – Completeness, consistency and trustworthiness

- **Data completeness** – Trends in reporting of metadata
  - Identify CSD Deposit checks and enhancements
  - Identify new filters to allow CSD users to better select fit for purpose data

- **Consistency** – Looking at experimental metadata to identify trends in information supplied

- **Trustworthiness** – Establishing automatic identification of potential cases of misconduct – including fraudulent and plagiarised data

Research integrity is much more than misconduct. *Nature*, 2019, 570, 5-5. DOI:10.1038/d41586-019-01727-0
Following Standard Ethical Practises

- CCDC is now a Member of the Committee on Publication Ethics.
- COPE’s objective is "to educate and advance knowledge in methods of safeguarding the integrity of the scholarly record for the benefit of the public".
- Membership gives us access to COPE resources and COPE advice – helping us deal with publication ethics and data integrity and issues.

https://publicationethics.org/about/governance
Revisiting Data

Targeted improvements allow improved integrity, consistency, discoverability and value of data.
Opportunities to educate and assist
Guidelines

The CCDC CIF Deposition Guidelines

When preparing your CIF for deposition please include as much information as possible and check it carefully. This is especially true for CSD Communications where there is no paper to describe the chemistry and experimental details leading to your structure. If you choose to publish your data as a CSD Communication please remember to provide all the authors/crystallographers/chemists who contributed to the crystallographic experiment as authors of the data. If we are unable to validate your structure from the information you have provided we may contact you. If we cannot resolve the issue, unfortunately, we may not be able to add your structure to the CSD.

Guidelines in Chinese

All experimental CIF files (including those from powder diffraction experiments) should contain an R-factor. This should be consistent with the crystallography being performed correctly and to the best ability that would be expected from the material and equipment used. We would like all experimental CIFs to contain:

- R-factors (R1, wR2, Rint)
- GoOF
- Shift/ESD (to show that the refinement has converged)
- Explanation of any problems with numbers of reflections and parameters
- Any residual electron density
- Details of squeeze/solvent masking
- Atomic Displacement Parameter (ADP) values
- Temperature – cell and data collection temperatures match
- Experimental set up including mounting device and instrument type
- HKL included
- RES included

We would encourage you to take advantage of the IUCr checkCIF reports built in to the deposition page. This can highlight issues to check with your structure that can be clarified in the validation reply form, particularly in the case of A- or B-level alerts. Ideally, treatment of disorder or partial occupancy atoms should be clear and of course, no non-positive definite atoms.

To allow us to create the most accurate representation of your structure please provide as much additional information on the “Enhanced Data” page as is appropriate for your structure. Some chemical issues we commonly encounter when processing data into the CSD are:

- Given formula and crystal formula don’t agree. Particular attention should be paid to hydrogen atoms which may not be located in the experiment. It would be very helpful to us to have a complete moiety formula (including unlocated hydrogens and any SQUEEZE/MASK species not located, if known)
- Charge balance, particularly for variable metal oxidation states and radicals
- Missing hydrogen atoms, especially on oxygen atoms that could be hydroxy/oxy/aqua ligands and for polyoxometalate structures
- Unusual bonding, tautomers or metal-metal bonding
- Poorly handled or unmodelled disorder
- Unexplained void space not accounted for by SQUEEZE or MASK procedures

Further information that will benefit the users of your structure and that will enable the correct identification of any previous versions of your structure are:

- Sterechemical determination method, if relevant
- Crystallisation solvent/conditions
- Melting point
- Details of re-refinement – please tell us if the structure is a re-refined version of an existing CSD entry
- Refcodes or CCDC numbers of any known related structures, i.e. by temperature / stereochernistry/ pressure, e.g. “high temperature determination of REFCODE”

If you have any further queries, please contact us via our Enquiries Page.
Adoption by the Community

- More structures are deposited with embedded HKL/Refinement data
- Increase dramatically since 2014
- Software made it part of the workflow
- Encouraging publishers to mandate this inclusion
What Else Could We Do?

• Improved peer review
  • Mandate crystallographic review of all structure-containing papers
  • Educate reviewers on how to referee crystallographic data

• File requirements
  • CIF + structure factors
  • Refinement instructions
  • CheckCIF report

• Validation checks
  • CheckCIF integration
  • Unit cell checks (with HKL checks? Or chemistry check?)
  • Geometry analysis?

• Additional files available to reviewers?
• Quality indicators attached to each entry?
Where else can we go?

• Predicted/calculated structures
  • Vast landscape of structural data
  • Exists in its own context
  • What is the most meaningful way to provide this to the community?

• Structure/property databases
  • What properties are best included?
  • What challenges exist to the standard curation of new experimental data

• Raw data archival?
  • That’s a question for another day...
Acknowledgements

- Suzanna Ward, Eric Rogers and the Database Group
- Ian Bruno for an introduction to the world of data best practices
- John Helliwell, Simon Coles and the IUCr CommDat

- The community, for providing a rich source of data and allowing us to help you share it with the world.

Thank you!
Abstract

As the Cambridge Structural Database (CSD) is poised to add its 1 millionth structure this year, we stop to reflect on the challenges involved in curating the world’s repository of small molecule crystal structure data and the opportunities that such big data can afford. Ensuring the completeness and quality of the data entrusted to us, accommodating data from new sources and experimental techniques, and continuing to provide relevant data in an ever-growing database are among a few of the challenges we face. The power of the CSD to inform various fields of chemical and biochemical research, stems from validation and enrichment of the data, as well as placing those data into a broader context. Only by continuing to evolve are we able to meet the needs of such a dynamic research community.

This talk will explore the part we can play in helping the community set and adhere to new best practices for data management that will enable researchers to get the most from crystal structure data. It will focus on the steps the CCDC has made in the past to address challenges, as well as some of the projects we are investigating as we seek to add the next million structures to the CSD.