CCDC Metadata Initiatives

Suzanna Ward, Ian Bruno, Colin Groom

Workshop on Metadata for raw data from X-ray diffraction and other structural techniques

Session IV: Data in the Wider World - From Laboratory to Database

Sunday 23rd August 9:35-10:00
From data to knowledge

Experimental Data

$C_{10}H_{16}N^+, Cl^-$

Structural Knowledge
From raw data to model

<table>
<thead>
<tr>
<th>Raw</th>
<th>Processed</th>
<th>Derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image data</td>
<td>Positions and intensities of diffracted beams</td>
<td>Modelled and refined from electron density</td>
</tr>
<tr>
<td>Structure factor data</td>
<td>Quantities needed to derive electron density</td>
<td></td>
</tr>
</tbody>
</table>

Primary emphasis has typically been on deposition of derived data only
# The Cambridge Structural Database

The Cambridge Structural Database (CSD) is a comprehensive collection of crystal structures, containing over 1 million entries as of 2023. It is a valuable resource for researchers in chemistry and materials science, providing detailed information about the atomic arrangements of molecules in the solid state.

## CIF deposition and validation service

This web service enables you to submit CIF files and associated structure factor files to the CCDC and for your structures to be included in the Cambridge Structural Database. Deposition allows you to correct small errors, check the integrity of your data, and add additional data.

Please include structure factor data for all structures.

- Files should be in CIF, CIF or XML format and may be
- At least one CIF file must be included in the submission
- All files submitted on one form must correspond to one
- There is a limit of 100 MB on the total size of files uploaded
- For more information please see our Structure Deposit

---

### Deposited CIF

<table>
<thead>
<tr>
<th>Atom</th>
<th>Symbol</th>
<th>Type</th>
<th>Fract X</th>
<th>Fract Y</th>
<th>Fract Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>C</td>
<td>0.31594(37)</td>
<td>0.75375(20)</td>
<td>0.70189(16)</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>C</td>
<td>0.19196(41)</td>
<td>0.77066(23)</td>
<td>0.76436(17)</td>
<td></td>
</tr>
<tr>
<td>O1</td>
<td>O</td>
<td>0.22259(32)</td>
<td>0.80293(21)</td>
<td>0.83459(13)</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>C</td>
<td>0.03051(40)</td>
<td>0.74837(26)</td>
<td>0.73926(18)</td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>H</td>
<td>-0.05699(40)</td>
<td>0.75910(26)</td>
<td>0.77720(18)</td>
<td></td>
</tr>
</tbody>
</table>

---

### CSD Entry

- **Identifier**: BEP8592
- **Author(s)**: En-Chieh Yang, W.Wernsdorfer, L.N.Zeichner, Y.Karaki, A.Yamaguchi, F.Mitsui, Guo-Di Lu, S.A.Wilson, A.L.Phillips, H.Hidemoto, D.N.Hendrickson
- **Formula**: C21H30Cl4N4O10
- **Space Group**: P62c
- **Cell Angles**: α = 16.142(6) deg, β = 16.142(6) deg, γ = 29.469(14) deg
- **R-factor**: 5.21
- **Disorder**: The water molecule is disordered by symmetry.
ZOYBIA – a co-crystal of vanillic acid and theophylline - the 750,000th structure in the CSD
Metadata

A set of data that describes and gives information about other data

- Data that describes the substance studied
  - Important for discovery, data analysis and mining
- Data that describes the dataset as a whole
  - Important for provenance (structure of record), attribution (citation)
  - Data that describes the associated publication
- Data that describes the experiment
  - Where was it done, how it was done, who did it, who funded it
Recent metadata developments at CCDC

Web Deposit  
- Launched 2014  
  - Updated Web Interface with greater interactivity  
  - Aim to capture as much input from depositor as possible

CSD-Xpedite  
- Deployed 2013  
  - Manages depositions and creation of CSD entries  
  - Designed to adapt to changing internal and external needs

CCDC DOIs  
- Implemented 2014  
  - Unambiguous and persistent identification of datasets  
  - Foundation for formalising data citation and interoperability

Get Structures  
- Launched 2014  
  - Update to previous “structure request” service  
  - Enhanced web interface for accessing individual structures
Crystallographic Information Framework (CIF)

A standard format for crystallographic data
- derived results
- raw and processed data
- experimental conditions
- publication data

International Union of Crystallography
Commission on Crystallographic Data
Commission on Journals
Working Party on Crystallographic Information

The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography*

BY SYDNEY R. HALL

Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

FRANK H. ALLEN

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND I. DAVID BROWN

Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada

(Received 8 April 1991; accepted 26 June 1991)

Abstract

There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of national and international networks, have fuelled the demand for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited effective use. This is true even in fields where the data requirements are well defined. Problems of data exchange are further exacerbated if the number and types of data change rapidly and continuously. Under these conditions specialized and local file formats have evolved. This diversity was tolerated when electronics.

The growth of electronic deposition

Hand-typed tables of coordinates

Table 1. Crystallographic data and details of data collection and processing for \(\text{MgNO}_2\), with M = Ce (11), Ni (12), and Cu (13)

<table>
<thead>
<tr>
<th>Structure</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>(\text{Pna}_21)</td>
<td>(\text{Pna}_21)</td>
<td>(\text{Pna}_21)</td>
</tr>
<tr>
<td>(a)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
</tr>
<tr>
<td>(b)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
</tr>
<tr>
<td>(c)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
<td>5.760(2)</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
<td>90(^\circ)</td>
</tr>
</tbody>
</table>

- **Table 2.** Atomic coordinates with estimated standard deviations in parentheses

<table>
<thead>
<tr>
<th>Atom</th>
<th>(x)</th>
<th>(y)</th>
<th>(z)</th>
<th>Atom</th>
<th>(x)</th>
<th>(y)</th>
<th>(z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ce1</td>
<td>0.000(1)</td>
<td>0.000(1)</td>
<td>0.500(1)</td>
<td>Ce2</td>
<td>0.500(1)</td>
<td>0.500(1)</td>
<td>0.000(1)</td>
</tr>
<tr>
<td>Ni1</td>
<td>0.000(1)</td>
<td>0.000(1)</td>
<td>0.500(1)</td>
<td>Ni2</td>
<td>0.500(1)</td>
<td>0.500(1)</td>
<td>0.000(1)</td>
</tr>
<tr>
<td>Cu1</td>
<td>0.000(1)</td>
<td>0.000(1)</td>
<td>0.500(1)</td>
<td>Cu2</td>
<td>0.500(1)</td>
<td>0.500(1)</td>
<td>0.000(1)</td>
</tr>
</tbody>
</table>

doi:10.1039/DT9850002177

Structures Digitally Deposited cf Published

CSD entries with 3D coordinates (1936-2013)
CCDC Number used as an indication of digital deposit
The growth of electronic deposition

Hand-typed tables of coordinates

doi:10.1039/DT9850002177
IUCr checkCIF and CSD-Deposit

- Deposition to CCDC now includes:
  - Generation of checkCIF reports
    - Available to depositor during submission
    - Stored at CCDC at submission
    - Soon to be available to publishers and referees during peer review process
  - An easy way to embed validation responses into CIF
  - Retrieval of edited CIFs and checkCIF reports
- **Launched July 2015**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level A</td>
<td>Most likely a serious problem - resolve or explain</td>
</tr>
<tr>
<td>Level B</td>
<td>A potentially serious problem, consider carefully</td>
</tr>
<tr>
<td>Level C</td>
<td>Check. Ensure it is not caused by an omission or oversight</td>
</tr>
<tr>
<td>Level G</td>
<td>General information/check it is not something unexpected</td>
</tr>
</tbody>
</table>

http://www.ccdc.cam.ac.uk/NewsandEvents/News/Pages/NewsItem.aspx?newsid=40
Capturing additional metadata

Example 1.cif
- data_5_93GPa

Example 2.cif
- data_Bu10kbar
How you capture metadata is important

- Important to capture metadata in a semantic or defined way.
- Free text is not necessarily a good thing.
- There can also be confusion over meaning for example the morphology as grown or as cut?

Study Temperature relative to Melting Point

Plot based on data as initially deposited and before validation at CCDC

Predominately due to MP reported in °C not K.
Over 70,000 structures deposited annually

The CCDC’s internal Systems automate most steps involved in processing deposited CIFs but manual intervention sometimes needed.
### CSD-Xpedite - Metadata used internally

<table>
<thead>
<tr>
<th>CSD-Xpedite</th>
<th>Deposition metadata</th>
<th>CSD entry metadata</th>
</tr>
</thead>
</table>

- **Deployed 2013**
- **Based on CRM**
- **Used to manage all the transactions that go on behind the scenes**
- **Number of different entities/record types**
- **Each entity has metadata associated with it**
  - Entities include contact, deposit, CSD entry, journal, publication, documentation
Assignment of chemistry

- Automated processes based on structural knowledge
- Manual validation by expert editorial staff

\[
P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}
\]

Assignment of chemistry is important for discovery, data mining, analysis and interoperability.
Data that describes the substance

Compound names based on IUPAC conventions using ACDname and manual curation

2D chemical diagram generated from existing CSD entries, CCDC diagram generation software and manual curation
Data that describes the dataset as a whole

• Important for provenance and attribution

• Includes:
  – Publication data
    • Could be in the scientific literature and/or repository
  – Authorship data
    • Creation of the sample, data collection, data analysis
Metadata - Associating CSD data to publications

Metadata exchanges

- Automated workflows with:
  - Publishers
  - Third Parties

Publication metadata at CCDC

<table>
<thead>
<tr>
<th>Journal</th>
<th>CrystEngComm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2015</td>
</tr>
<tr>
<td>Volume</td>
<td>17</td>
</tr>
<tr>
<td>First page</td>
<td>5377</td>
</tr>
<tr>
<td>IUCr Data Validation</td>
<td>--</td>
</tr>
<tr>
<td>Manuscript</td>
<td>CFC</td>
</tr>
<tr>
<td>Authors</td>
<td>Conrad W. Ingram, Geoffrey Kibakaya, John Bacsa, Stephan R. Mathis, Jr., Alvin A. Holder, Varma H. Rambaran, Brandon Dennis, Esmeralda Castanea, Julianne S. Robbins, Z. John Zhang</td>
</tr>
</tbody>
</table>

Metadata from publishers

<doi>doi:10.1107/S2056989015014383</doi>
<title>Crystal structure of 6,6,12,12-tetrachlorotricyclo[8.2.0.4,7]dodecane-5,11-dione</title>
<journal>unm-issn:2056-9890</journal>
<cnor>zu2962</cnor>
<author>Turan Alun, E.</author>
<author>Hokecik, T.</author>
<volume>71</volume>
<issue>9</issue>
<firstpage>1000</firstpage>
<lastpage>1002</lastpage>
<published>2015-08-06</published>
<ccdc>1415865</ccdc>

Publication metadata from third parties
Metadata enabling referee access to data

From: 418e7b.0b9f8125@editorialmanager.com
Sent: 04 March 2015 03:45
Subject: Manuscript [redacted] for review

In view of your expertise I would be very grateful if you could review the following manuscript ...

If this paper includes new structural information, the data should be available from one of the following databases:

CCDC (organic and metal-organic compounds): [redacted]

Fiz Karlsruhe (inorganic crystal structures): [redacted]

PDB (proteins): [redacted]
Metadata used to link articles and data

Links are in place for ACS, RSC, Elsevier, Wiley and IUCr journals
CCDC crystal structure DOIs

- Over 560,000 dataset DOIs registered since April 2014
- DataCite metadata includes related article DOIs
- Foundation for formalising data citation and interoperability
- Data Citation Principles - Data should be considered legitimate, citable products of research...

Synthesis, Structure, and Catalytic Studies of Palladium and Platinum Bis-Sulfoxide Complexes

E. Postgrad, A. Crystallographer and R. Supervisor

Abstract
DOI: 10.1021/jacs.5b03972

Coordination-directed self-assembly has become a well-established technique for the construction of functional supramolecular structures. In contrast to the most often exploited transition metals, trivalent lanthanides LnIII have been less utilized in the design of polynuclear self-assembled structures despite the wealth of stimulating ....[...]

Supporting Information

Crystallographer doesn’t always get credited for their contribution.

Article DOI primarily identifies the paper and not the data.

Crystal structure data often buried amongst the supporting information.
Provenance and attribution

CSD DOI – Access to CSD entry & data

Information added to structure of record

Dataset publication

http://dx.doi.org/10.5517/cc5x3w1
DOIs – facilitating interoperability

Facilitates formal citation AND interoperability. Currently adding indication of full citation and resourceType to metadata

Coverage of CCDC data by the Thomson Reuters Data Citation Index achieved via the DataCite metadata store.
DOIs – facilitating interoperability

UK Research Data Discovery Service: Aim is to aggregate metadata for research data held within UK universities and national, discipline specific data centres.

RDA-WDS Publishing Data Services prototype taking a feed of CCDC article-dataset links from DataCite.

Data literature inter-inking service (beta) that ingests data article -inks. Inspection portal: http://dliservice.research infrastructures.eu/
Linking based on chemistry

Bulk identification of links between ChemSpider molecules and CSD Crystal Structures (~50,000) facilitated by InChIs.
Other crystal data repositories

- PDB
  - match PDB ligands to best representative CSD molecules
  - working group on ligand validation

29 July, 2015

Data correspondences between the PDB and CSD archives now available

The Worldwide Protein Data Bank and the Cambridge Crystallographic Data Centre (CCDC, http://www.ccdc.cam.ac.uk) are pleased to announce the availability of a new data resource containing correspondences between the biopolymer components and ligand molecules found in the PDB archive that exactly match small-molecule X-ray structures in the Cambridge Structural Database (CSD) archive.

The chemical structure of every unique molecule in the Protein Data Bank is described in the PDB Chemical Component Dictionary. The new PDB Chemical Component Model data file complements information in the PDB by providing the following CSD information for matching molecular entries: accession code correspondences, Cartesian coordinates and R-value, data-collection temperature and a disorder flag, SMILES and InChI descriptors, and a Digital Object Identifier (DOI) for the citation associated with the CSD entry.

At present, there are 20,077 chemical components in the PDB Chemical Component Dictionary, and for 1,418 of these exact match structures have been identified in the CSD. The new PDB Chemical Component Model file is available from the PDB FTP archive via:
Linking experiment to data

ISIS: Neutron and muon instruments at the UK Science and Technology Facilities Council Rutherford Appleton Laboratory.

ICAT

The ISIS data catalogue

All data produced at ISIS is catalogued into ICAT – the ISIS data catalogue. This catalogue is made available as a searchable website, via Mantid and as an API.

ISIS will issue citable DOIs for all experiments carried out at ISIS. You are encouraged to cite these in publications relating to ISIS experiments. For more information see DOIs

CCDC and STFC investigating the potential for linking between:
• raw data stored at STFC
• modelled structure stored at CCDC

Challenge: reliably associating results with original experiment
Metadata about people and institutions

Assigning DOIs through DataCite allows researchers to add their data to researcher IDs such as ORCiD.

Funders are wanting to know the output of their funded research. This will require linking between people, publications, data and funders.

Could be used to find institutions research output. Ringgold IDs now in our internal systems and we want to work towards reliably identifying institutions at point of deposition.
What next?

CSD Deposition Portal

- Integration of Researcher and Institutional IDs
- Exposure of CCDC services during deposition
  - Ability to edit CSD entry metadata such as 2D chemical diagram
  - Extend validation and integrity checks
- Capturing other metadata
  - Topological descriptions, funders, experimental data, links to associated data, etc
- Integration with third party software

Interoperability, links and programmatic access

- Extension of links....
- Extend CSD Python API to access deposited data

CSD-Community Access

- Extension of search functionality
- Addition of InChI lookups
Thank you