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

A Satellite Workshop to the 29th European Crystallographic Meeting

Practical session: 2. Towards a distributed / virtual repository of primary datasets
• In the Madrid Congress the Commissions were charged at the DDD inaugural meeting to define the metadata that should accompany their raw data ...

• While each IUCr Commission needs 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.
DDDWG Work Plan

Metadata

• A centralised crystallographic repository of raw dataset metadata should be scoped and piloted.
• With such a repository in place, we should revisit the proposal that authors shall provide a permanent and prominent link from an article to the associated raw datasets.
Raw crystallographic datasets are now beginning to be deposited at a number of locations.

- Store.Synchrotron (MyTARDIS)
Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals

Simon W. M. Tanley, Antoine M. M. Schreurs, John R. Helliwell and Loes M. J. Kroon-Batenburg

Institution: Utrecht University

Licensing: This experiment data is licensed under Creative Commons Attribution 3.0 Australia (CC BY 3.0).

Administrators: Steve Androulakis

16 Datasets

Just start typing to filter datasets based on descriptions

- 4DD0 (6.8 GB)
- 4DD1 (628.5 MB)
- 4DD1 un warp (1.2 GB)
- 4DD2 (6.8 GB)
- 4DD3 (6.8 GB)
- 4DD4 (787.4 MB)
- 4DD4 un warp
Raw crystallographic datasets are now beginning to be deposited at a number of locations.

- Store.Synchrotron (MyTARDIS)
- Zenodo
16 June 2011

Simultaneous X-ray diffraction from multiple single crystals of macromolecules

Pathankar, Karthik; Sørensen, Henning; Wright, Jonathan; Schmidt, Soren; Poulsen, Henning; Garman, Elspeth


The potential in macromolecular crystallography for using multiple crystals to collect X-ray diffraction data simultaneously from assemblies of up to seven crystals is explored. The basic features of the algorithms used to extract data and their practical implementation are described. The procedure could be useful both in relation to diffraction data obtained from intergrown crystals and to alleviate the problem of rapid diffraction decay arising from the effects of radiation damage.

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Size</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEWL_7_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>918.4 MB</td>
<td>Download</td>
</tr>
<tr>
<td>HEWL_4_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>913.7 MB</td>
<td>Download</td>
</tr>
<tr>
<td>HEWL_3_crystals.tar.bz2</td>
<td>13 Aug 2014</td>
<td>1.7 GB</td>
<td>Download</td>
</tr>
<tr>
<td>HEWL_2_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>1.9 GB</td>
<td>Download</td>
</tr>
<tr>
<td>README.txt</td>
<td>13 Aug 2014</td>
<td>906 Bytes</td>
<td>Download</td>
</tr>
<tr>
<td>insulin_4_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>964.4 MB</td>
<td>Download</td>
</tr>
<tr>
<td>insulin_3_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>1.8 GB</td>
<td>Download</td>
</tr>
<tr>
<td>insulin_2_crystals.tar.lzma</td>
<td>13 Aug 2014</td>
<td>1.8 GB</td>
<td>Download</td>
</tr>
</tbody>
</table>
Finding datasets

Raw crystallographic datasets are now beginning to be deposited at a number of locations.

- Store.Synchrotron (MyTARDIS)
- Zenodo
- University of Manchester eScholar
Abstract

The International Union of Crystallography has for many years been advocating archiving of raw data to accompany structural papers. Recently, it initiated the formation of the Diffraction Data Deposition Working Group with the aim of developing standards for the representation of these data. A means of studying this issue is to submit exemplar publications with associated raw data and metadata. A recent study on the effects of dimethyl sulfoxide on the binding of cisplatin and carboplatin to histidine in 11 different hexameric crystal forms from two diffractometers led to an investigation of the possible effects of the equipment and X-ray diffraction data processing software on the calculated occupancies and B factors of the bound Pt compounds. 20.5 Gb of data were transferred from Manchester to Utrecht to be processed with EVAL. A systematic comparison shows that the largest differences in the occupancies, and B factors of the bound Pt compounds are due to the software, but the equipment also has a noticeable effect. A detailed description of and discussion on the availability of metadata is given. By making these raw diffraction data sets available via a local repository, it is possible for the diffraction community to make their own evaluation as they may wish.

Bibliographic metadata

Content type: Research data
Research data type:
Digital Object Identifier: 10.15127/1.219230
Manchester eScholar: uk-ac-man-scw-219230
PID:
Title: HEWL_cisplatin_aglycerol
Submit: 4ddo
Data creators:
Data contributors:
Related publications:
Publisher:
The University of Manchester
Published year: 2013
Version: 46
External data resources: 10.15127/1.219230
Embargo period: Immediate release
Release date: 13th February, 2014
Access status: Active

Institutional metadata

University researcher(s): Tanley, Simon
Academic department(s):
Faculty of Engineering and Physical Sciences
Faculty of Life Sciences
School of Chemistry

Record metadata

Manchester eScholar ID: uk-ac-man-scw-219230
Created by: Tanley, Simon
Created: 13th February, 2014, 14:49:03
Last modified by: Gibson, Christopher
Last modified: 15th March, 2015, 13:35:22
Finding datasets

Raw crystallographic datasets are now beginning to be deposited at a number of locations.

- Store.Synchrotron (MyTARDIS)
- Zenodo
- University of Manchester eScholar
- eCrystals / Atlas data store
- Protein Data Bank
- Wladek Minor Laboratory, U. Virginia
- Experimental facilities
Finding datasets

Raw crystallographic datasets are now beginning to be deposited at a number of locations.

Finding datasets

Raw data: PDB code 4dd0 HEWL_cisplatin_aqueous_glycerol

Link http://dx.doi.org/10.15127/1.219230
Finding datasets

Link http://rawdata.chem.uu.nl/#0001
Raw data: archive at Utrecht University containing images measured at Manchester University
Finding datasets

Link: http://vera183.its.monash.edu.au/protein_cisplatin_carboplatin
Raw data: mirror of the raw data from Tardis at Monash University

Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals

1. Experience with exchange and archiving of lysozyme crystals

| PDB  | Sample image | Snap | In original format | Cc | O | P | D | E | A | C | D | E | F | G | H | I | J | K | L |
|------|--------------|------|-------------------|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4DD0 | 4DD0_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD1 | 4DD1_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD2 | 4DD2_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD3 | 4DD3_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD4 | 4DD4_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD5 | 4DD5_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD6 | 4DD6_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD7 | 4DD7_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4DD8 | 4DD8_01_0001 |          |                   |    |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Description: Metadata, Sharing, Transfer Datasets

Institution: Unicraft University

License: This experiment data is licensed under Creative Commons Attribution 3.0 Australia (CC BY 3.0).

Administrators: Steve Androuakis

Downloaded All / 16 Datasets

Just starting to filter datasets based on descriptions
Requirements

- Identification
- Provenance
- Disambiguation
- Categorization
- Context
- Relationship
- Size
- Licence
The OAI-Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. ... The metadata that is harvested may be in any format that is agreed by a community ... although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or "aggregated" data. The link between this metadata and the related content is not defined by the OAI protocol. It is important to realise that OAI-PMH does not provide a search across this data, it simply makes it possible to bring the data together in one place. ... To provide services, the harvesting approach must be combined with other mechanisms.

Although the OAI-PMH is technically very simple, building coherent services that meet user requirements remains complex. The OAI-PMH protocol could become part of the infrastructure of the Web, as taken-for-granted as the HTTP protocol now is, if a combination of its relative simplicity and proven success by early implementers in a service context leads to widespread uptake by research organisations, publishers, and "memory organisations".
Communicating with repositories

Dublin Core

Dublin Core Metadata Element Set Version 1.1

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

- Identification
- Provenance
- Disambiguation
- Categorization
- Context
- Relationship
- Size
- Licence
• So, Dublin Core could serve as a metadata descriptor language in repositories containing crystallographic data

• But . . .

The OAI-Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. ... The metadata that is harvested may be in any format that is agreed by a community ... although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or "aggregated" data. The link between this metadata and the related content is not defined by the OAI protocol. It is important to realise that **OAI-PMH does not provide a search across this data**, it simply makes it possible to bring the data together in one place. ... **To provide services, the harvesting approach must be combined with other mechanisms.**
Finding the data

OAI-PMH has three potentially useful features

1. Sets
   • A set is an optional construct for grouping items for the purpose of selective harvesting. Repositories may organize items into sets. Set organization may be flat, i.e. a simple list, or hierarchical. Multiple hierarchies with distinct, independent top-level nodes are allowed.

2. Friends
   • The friends container . . . is used by repositories that want to point harvesters to other repositories, by listing their base URLs. Usage of the friends container is recommended; it may support harvesters in discovering the network-location of repositories.

3. Metadata negotiation
   • ListMetadataFormats [is a verb] used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific item. If this argument is omitted, then the response includes all metadata formats supported by this repository. Note that the fact that a metadata format is supported by a repository does not mean that it can be disseminated from all items in the repository.
1. Sets

- Define one or more set specifications that characterise crystallographic data, and educate repository managers to use these consistently, *e.g.*
  - subject:crystallography-images

- **For**: establishes and supports a natural classification/organizational hierarchy
- **Against**: requires construction of the classification scheme; must ensure impossibility of conflict with other *ad hoc* schemes
2. Friends

- Crystallographic repositories should include a friends container listing other known crystallographic repositories

- **For**: allows incremental growth of network from local knowledge
- **Against**: possible duplication or multiple misallocation of related repositories; may benefit from a central registry of known ‘friends’ in this discipline
3. Metadata negotiation

- The crystallographic community (this WG?) should define an appropriate ‘middle-layer’ metadata scheme for optimising handling of high-level crystallographic metadata by repositories.

- **For**: The existence of such a scheme in a repository guarantees that you are retrieving the desired sort of data; freedom to build complex systems.

- **Against**: need to devise, test and implement such a scheme; may be needlessly duplicating other, more general efforts [e.g. Jisc Research at Risk consultation identified need for a discipline neutral metadata scheme – http://researchdata.jiscinvolve.org/wp/2015/07/03/research-data-metadata/ (thanks to Chris Gibson, U. Manchester)]
What should be the outcome of a mechanism for locating distributed data sets?

A central database of deposited data sets

- Identifier (DOI)
- Provenance
- Link to publication
- Link to structure deposition (CSD, PDB, COD, ICSD)
- Nature (raw, processed, derived; X-ray, electron, neutron; diffraction, microscopy, NMR)
- Quality (resolution, completeness, level of interest)
- Size
- Licence / access
Stakeholders

Who benefits?
Who pays?
Who develops?
Who maintains?

IUCrData?