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

DDDWG Work Plan

Metadata

DDDWG Triennial Report 2011-2014

- 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

DDDWG Meeting Montreal 2014

- 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.

Finding datasets

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

- Store.Synchrotron (MyTARDIS)

vera183.its.monash.edu.au/experiment/view/40/

MyTardis Home About My Data Public Data Stats Help Register Log In

Experiment

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

16 | 14262 | 51.3 GB | 6th February 2013 | Public

Toggle Full Description

Mirrored from http://rawdata.chem.uu.nl/ ..

Description Metadata Sharing Transfer Datasets

Institution Utrecht University

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

Administrators Steve Androulakis

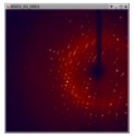
Download All TAR

16 Datasets

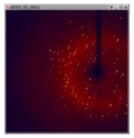
Download Selected

Just start typing to filter datasets based on descriptions

4DD0
361
6.0 GB



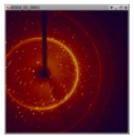
4DD1
615
626.5 MB



4DD1 unwarp
615
1.2 GB

4DD2
361
6.0 GB

4DD3
361
6.0 GB



4DD4
778
787.4 MB

4DD4 unwarp

Finding datasets

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

- Store.Synchrotron (MyTARDIS)
- Zenodo

16 June 2011

Dataset Open access

Simultaneous X-ray diffraction from multiple single crystals of macromolecules

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

(show affiliations)

X-ray diffraction datasets from the
publication: K. S. Paithankar, H. O. Sørensen, J. P. Wright, S. Schmidt, H. F. Poulsen and E. F. Garman*

Acta Cryst. (2011). D67, 608-618 doi:10.1107/S0907444911015617

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.

Files

Name	Date	Size	
HEWL_7_crystals.tar.lzma	13 Aug 2014	918.4 MB	Download
HEWL_4_crystals.tar.lzma	13 Aug 2014	913.7 MB	Download
HEWL_3_crystals.tar.bz2	13 Aug 2014	1.7 GB	Download
HEWL_2_crystals.tar.lzma	13 Aug 2014	1.9 GB	Download
README.txt	13 Aug 2014	906 Bytes	Download
insulin_4_crystals.tar.lzma	13 Aug 2014	964.4 MB	Download
insulin_3_crystals.tar.lzma	13 Aug 2014	1.8 GB	Download
insulin_2_crystals.tar.lzma	13 Aug 2014	1.8 GB	Download

Publication date:
16 June 2011
DOI
DOI 10.5281/zenodo.11277
Keyword(s):
macromolecular crystallography methods development radiation damage
Collections:
Communities > Macromolecular Crystallography Datasets
Open Access
License (for files):
Creative Commons Attribution Share-Alike
Uploaded on:
13 August 2014

Finding datasets

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

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

HEWL_cisplatin_aq_glycerol: 4dd0

Tanley, Simon

[Research data] version 46. 2013. The University of Manchester.

Access to files

cisplatin_aq_glycerol_1-70.zip (zip)
cisplatin_aq_glycerol_71-140.zip (zip)
cisplatin_aq_glycerol_141-210.zip (zip)
cisplatin_aq_glycerol_211-280.zip (zip)
cisplatin_aq_glycerol_281-360.zip (zip)

Abstract

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 lysozyme crystals 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. 35.3 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 depository, 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_aq_glycerol
Subtitle:	4dd0
Data creators:	Tanley, Simon
Data contributors:	Helliwell, John R; University of Manchester
Related publications:	uk-ac-man-scw:219230
Publisher:	The University of Manchester
Published year:	2013
Version:	46
External data resources:	doi:10.1107/S0021889812044172
Embargo period:	Immediate release
Release date:	13th February, 2014
Access state:	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:	13th March, 2015, 13:35:32

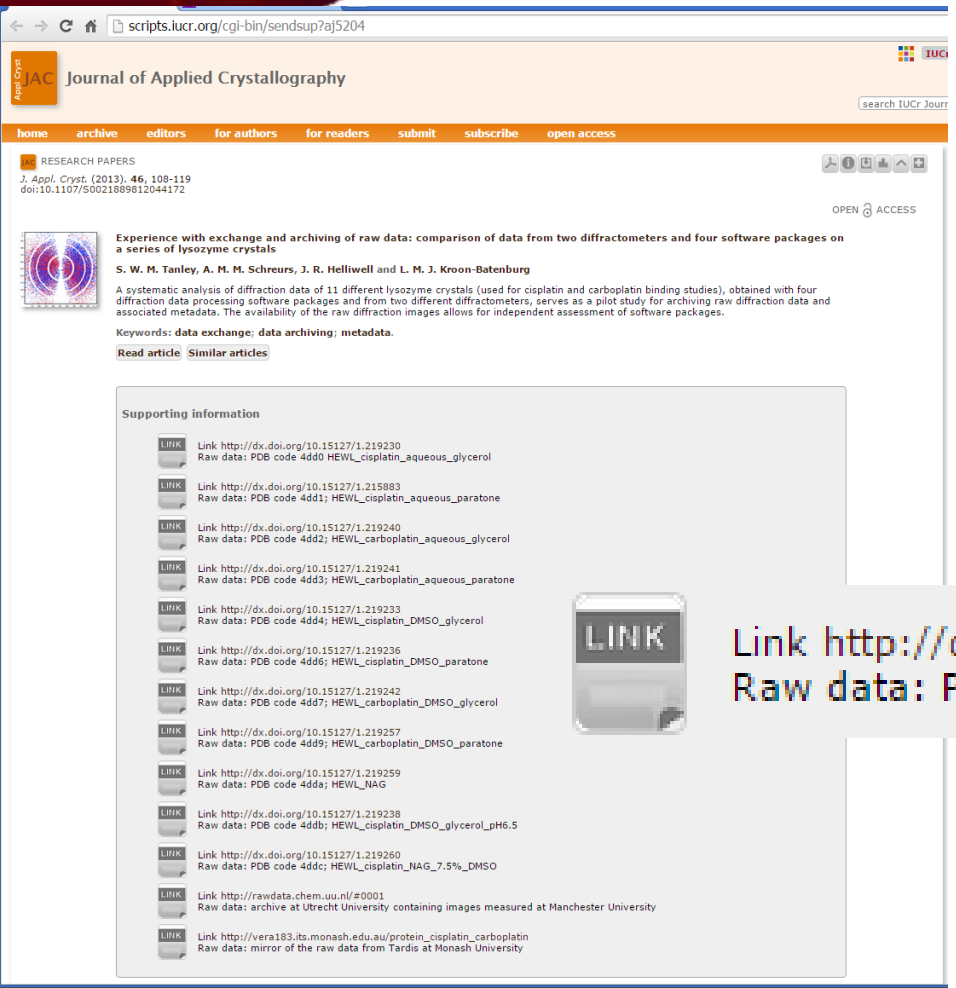
Finding datasets

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.

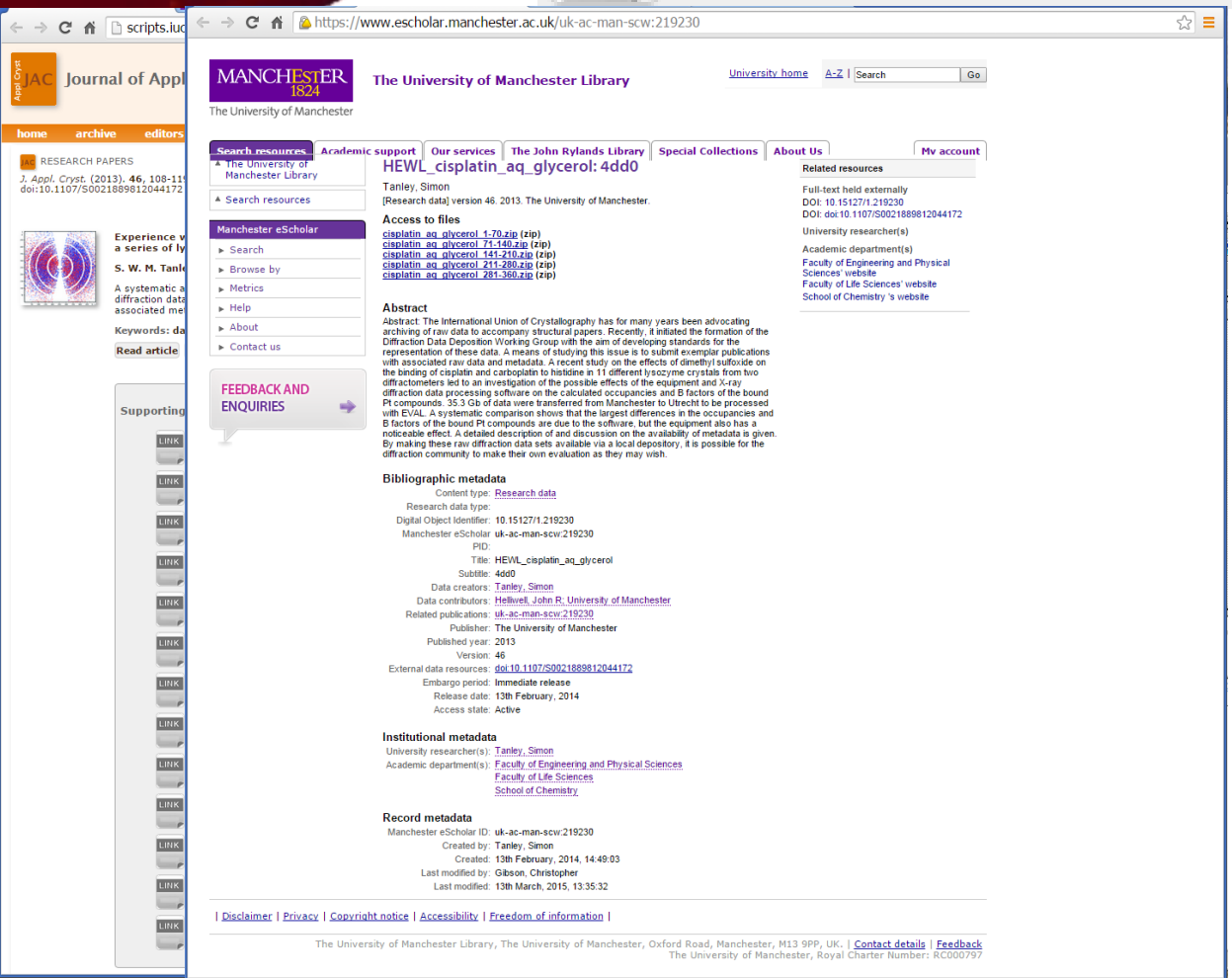


Tanley, S. W. M., Schreurs, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals. *J. Appl. Cryst.* **46**, 108–119.

Link http://dx.doi.org/10.15127/1.219230
Raw data: PDB code 4dd0 HEWL_cisplatin_aqueous_glycerol

Finding datasets

Link http://dx.doi.org/10.15127/1.219230
Raw data: PDB code 4dd0 HEWL_cisplatin_aqueous_glycerol



Finding datasets

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

rawdata.chem.uu.nl/c001/index.html

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

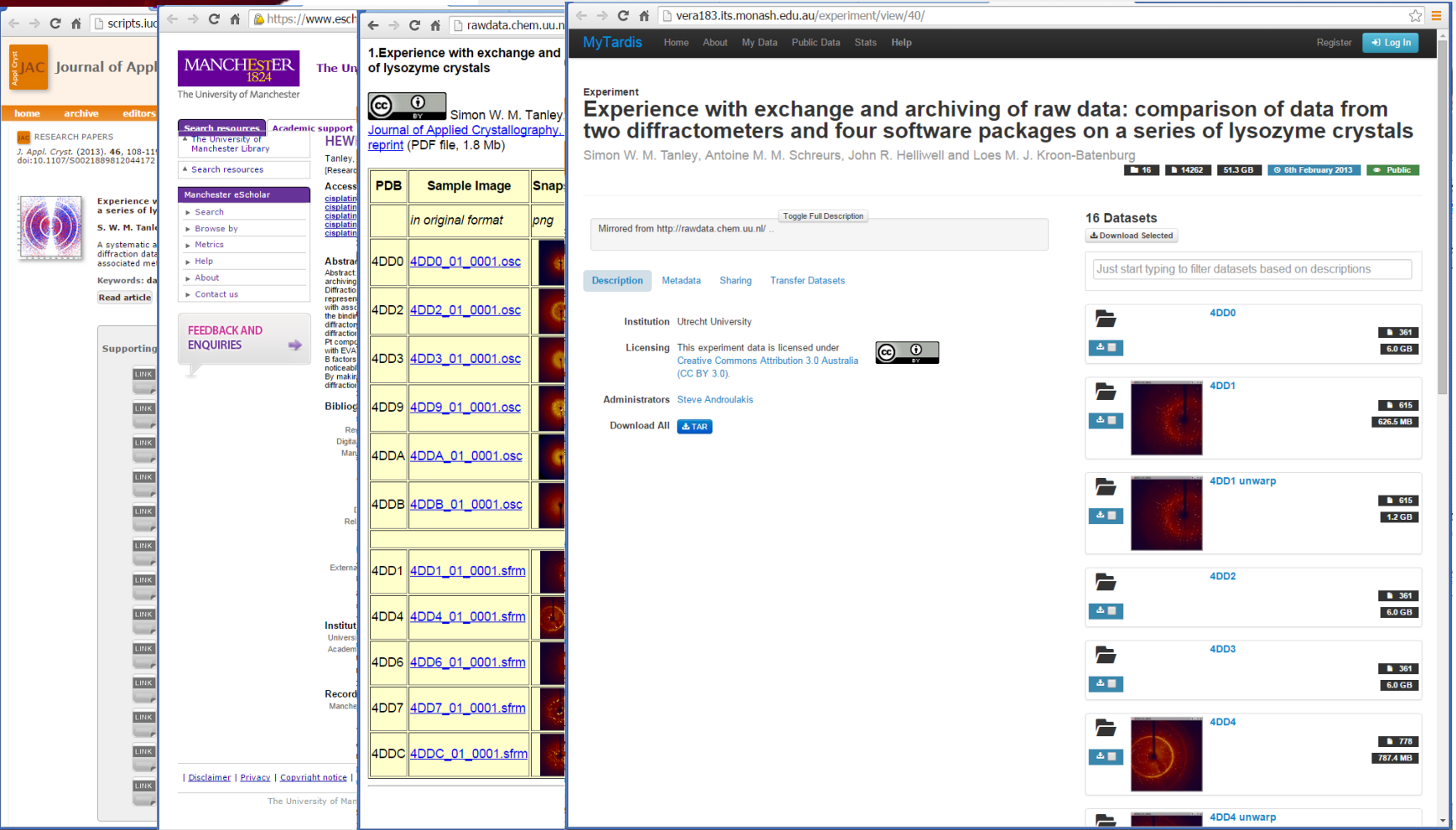
Simon W. M. Tanley, Antoine M. M. Schreurs, John R. Helliwell and Loes M. J. Kroon-Batenburg
Journal of Applied Crystallography, 2013, Volume 46, pages 108-119
reprint (PDF file, 1.8 Mb)

PDB	Sample Image	Snapshot	Nr of Scans	Nr of Images	Tarfile(s)	Size (Mb)	Expanded Size (Mb)	Diffractometer
	in original format	*png*			*X.tar.gz unpacks into subdirectory X*			
4DD0	4DD0_01_0001.osc		1	360	4DD0.tar.gz	1465	6191	Rigaku R_AXIS IV
4DD2	4DD2_01_0001.osc		1	360	4DD2.tar.gz	2657	6191	
4DD3	4DD3_01_0001.osc		1	360	4DD3.tar.gz	2293	6191	
4DD9	4DD9_01_0001.osc		1	360	4DD9.tar.gz	2716	6191	
4DDA	4DDA_01_0001.osc		1	180	4DDA.tar.gz	1036	3096	
4DDB	4DDB_01_0001.osc		1	360	4DDB.tar.gz	2249	6191	
4DD1	4DD1_01_0001.sfrm		1	614	4DD1.tar.gz 4DD1unwarp.tar.gz	480 554	630 1236	Bruker PLATINUM[135]
4DD4	4DD4_01_0001.sfrm		5	777	4DD4.tar.gz 4DD4unwarp.tar.gz	591 624	792 1564	
4DD6	4DD6_01_0001.sfrm		2	1440	4DD6.tar.gz 4DD6unwarp.tar.gz	1067 1117	1464 2897	
4DD7	4DD7_01_0001.sfrm		4	1862	4DD7.tar.gz 4DD7unwarp.tar.gz	1464 1566	1913 3746	
4DDC	4DDC_01_0001.sfrm		2	1440	4DDC.tar.gz 4DDCunwarp.tar.gz	1149 1220	1480 2897	proteumnac.p4p

Last modified: 6/19/2015, 1:29:22 PM

Finding datasets

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



Requirements

- Identification
- Provenance
- Disambiguation
- Categorization
- Context
- Relationship
- Size
- Licence

Communicating with repositories

OAI-PMH

Protocol for Metadata Harvesting (OAI-PMH)

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.

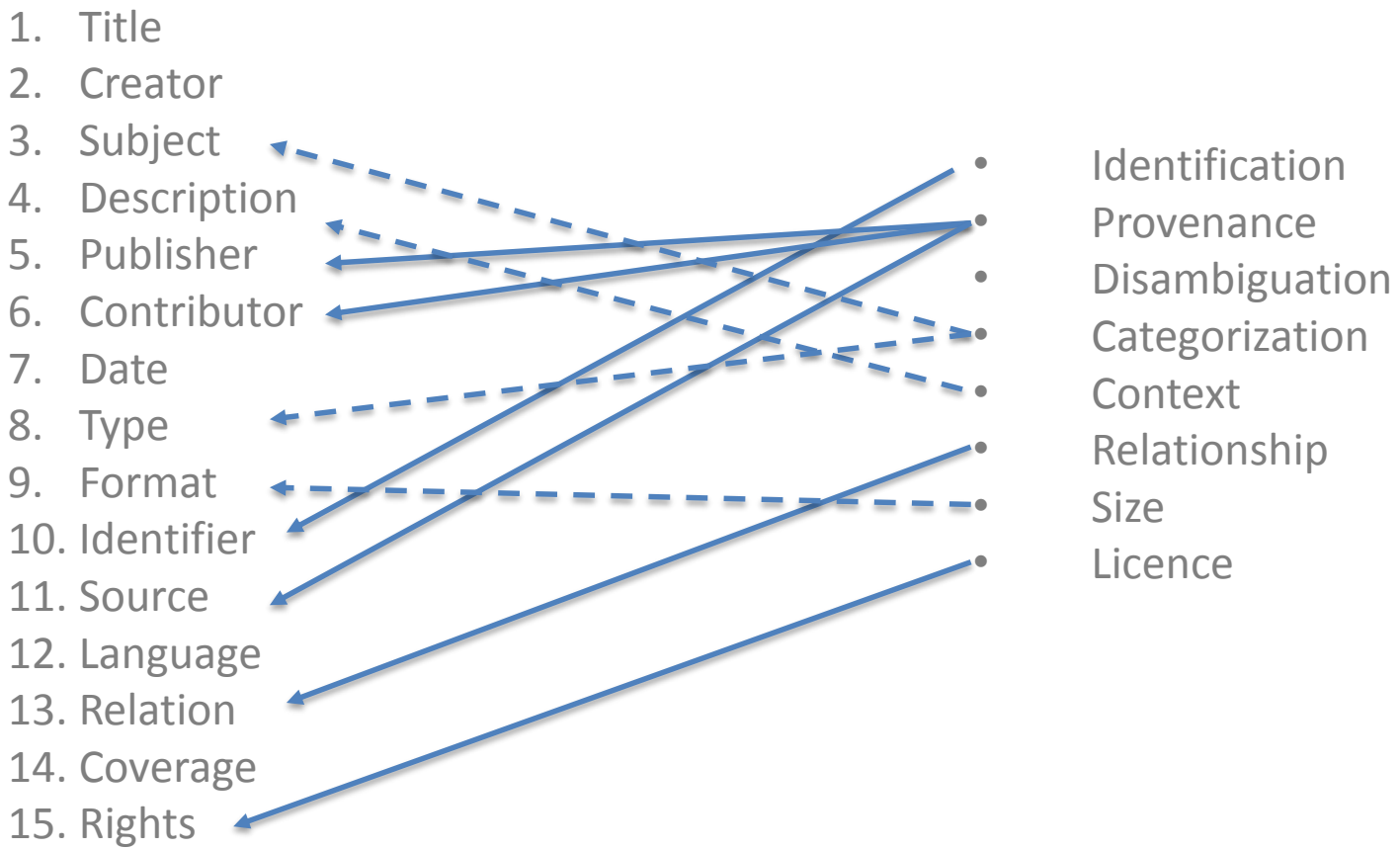
Although **the OAI-PMH is technically very simple,** building coherent services that meet user requirements remains complex. The OAI-PMH protocol could become part of the infrastructure of the Web, as taken-for-granted as the HTTP protocol now is, if a combination of its relative simplicity and proven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

http://www.oaforum.org/tutorial/english/page1.htm

Communicating with repositories

Dublin Core

Dublin Core Metadata Element Set Version 1.1



Communicating with repositories

OAI-PMH

- *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.

Finding the data

Recommendations

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 *or*
     discipline:crystallography:category:experimental:type:diffraction-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

Finding the data

Recommendations

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

Finding the data

Recommendations

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)

Outcome

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?