
Link to official URL (if available): http://dx.doi.org/10.1109/e-Science.2009.27

Opus: University of Bath Online Publication Store
http://opus.bath.ac.uk/

This version is made available in accordance with publisher policies. Please cite only the published version using the reference above.

See http://opus.bath.ac.uk/ for usage policies.

Please scroll down to view the document.
The Role of Representation Information in the Digital Curation of Crystallography Data

Manjula Patel¹, Simon Coles², David Giaretta³, Stephen Rankin³, Brian McIlwrath³

¹UKOLN, University of Bath, UK
²EPSRC NCS, University of Southampton, UK
³Science & Technology Facilities Council, UK

5th IEEE International Conference on e-Science
9-11th December 2009
Oxford, UK
Data Deluge

40 years ago a PhD student would determine about 3 crystal structures for their thesis – this can now be easily achieved in a day.

The primary cause is the current data publication process, which is tied to journal articles and peer review.
The Solution

Intellect & Interpretation (Journal article, report, etc)

Underlying data (Institutional data repository)
The eCrystals Data Repository

- Quick & simple to deposit
- Software tools
- Laboratory archive
- Community involvement
- ‘Embargo’ facility
- Structured foundations
- Discoverable & harvestable
- Number of file formats

[http://ecrystals.chem.soton.ac.uk](http://ecrystals.chem.soton.ac.uk)
A Thorough Approach to Dissemination

• Using simple Dublin Core protocol (OAI-PMH)
  • Crystal structure
  • Title (Systematic IUPAC Name)
  • Authors
  • Affiliation
  • Creation Date
• Additional chemical information through Qualified Dublin Core
  • Empirical formula
  • International Chemical Identifier (InChI)
  • Compound Class & Keywords
• Specifies which ‘datasets’ are present in an entry
• Application Profile http://www.ukoln.ac.uk/projects/ebank-uk/schemas/
• DOI links http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145
• Rights & Citation http://ecrystals.chem.soton.ac.uk/rights.html
Building a Federation of Repositories
Crystallography Data Commons

The cdlInk Sakai site is a collaboration site for the development of repository requirements and frameworks for raw and derived crystallographic data.

- Generic publication & dissemination
- Domain & context
- Management
- Preservation Data for Crystallography Data
- Core Scientific Data Model
OAIS Background

- Development led by the Consultative Committee for Space Data Systems (CCSDS)
- Adopted as ISO 14721:2003
- “Open” refers to development of the model in an open forum
- Reference Model, not a blueprint for implementation
- Establishes a common framework of terms and concepts
- Identifies the basic functions of an OAIS
- Defines an information model
- Three major areas of influence:
  - Preservation metadata schemas
  - Architecture and system design
  - Conformance criteria for archival repositories
OAIS Definition and Selected Concepts

• **OAIS:** “An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community”

• **Designated Community:** Community of stakeholders and users that the OAIS serves

• **Knowledge Base:** A set of information, incorporated by a user or system, that allows that user or system to understand the received information

• **Information Object:** Data Object + Representation Information

• **Representation Information:** any information required to render, interpret, use and understand digital data

• **Information Package:** Content Information + Preservation Description Information + Packaging Information (Submission, Archival and Dissemination Information Packages)

• **Preservation Description Information:** Provenance, Context, Reference, Fixity information
**OAIS Functional Entities**

- **Ingest:** services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS.

- **Archival Storage:** services and functions used for the storage and retrieval of AIPs.

- **Data Management:** services and functions for populating, maintaining, and accessing a wide variety of information.

- **Administration:** services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis.

- **Preservation Planning:** services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community.

- **Access:** services and functions which make the archival information holdings and related services visible to Consumers.
• **Information Object** is composed of a Data Object that is either physical or digital, as well as the Representation Information that allows for the full interpretation of the data into meaningful information.

• **Representation Information** is *any* information required to render, interpret, use and understand data.
• Types of RI
  - **Structure** e.g. file formats for text, images, audio, moving images, datasets, 3D models
  - **Semantic** e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri
  - **Other** e.g. software, algorithms, standards, time dependent information, actions, processes

• RI is recursive in nature; using one element of RI in a meaningful manner may well require further RI, resulting in a RI Network
  - Recursion is terminated based on the designated community’s knowledge base
  - Essential that RI itself is curated and preserved to maintain access to data
Registry/Repository of RI (RRoRI)

- Development started under the DCC-Development team
- Work now being undertaken jointly with the CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (Integrated Project co-funded by EU FP6 Programme, April 2006)
- Representation Information is the key to long-term access
- RRoRI should itself be a trustworthy OAIS
- Repository: some RI is stored; Registry: links to external RI
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI for third party use
RRoRI: RI Label & CPID

• Idea of RI is the key
  – Information Object: a specific object to be archived/preserved/curated
  – RI: all information required to render, interpret, use and understand the object
  – RI Label: used to connect RI to an Information Object

• RI Label serves as a mechanism for accessing RI in RRoRI
  – Label is used to identify relevant RI
  – Provides mechanism for recording individual RI components

• RI Label has a Curation Persistent Identifier (CPID)
  – Used to connect the digital object to the RI Label
Use of CPID

1. User gets data from archive. Data has associated Curation Persistent Identifier (CPID).

2. User unfamiliar with data so requests RI using CPID.

3. User receives RI – which has its own CPID in case it is not immediately usable.

- The Digital Object could have some RI packed with it, as well as a CPID.
- CPID supports automated access & processing.
Capturing RI: Crystallography Data

- Crystallography data are highly structured
- Convention is to share derived or reduced data, access to raw data is rare
- Crystallography Information File (CIF) is a de facto exchange standard
- CIF maintained by International Union of Crystallography (IUCr)
- Open standards and software e.g. CIF, checkcif, CML, INChI
- Culture for sharing/depositing data (CCDC)
- Well-established workflow for crystallography experiments
A client with a GUI for ingest, search and retrieval of RI and RI labels
Conclusions

• A preservation strategy based on RI depends on a global, well-engineered, distributed infrastructure of RI
  - Needs coordination, collaboration and globally shared effort
  - Mining of RI networks for inference purposes
• Creation of robust RI networks requires domain expertise
• Likely to be gaps in global networks of RI
  - Business case for using a store of RI is clear, however the case for submitting RI to the global effort is less clear (commercial, IPR etc.)
Thank You

Questions?

Manjula Patel, Simon Coles

eCrystals Federation Project
http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page