eBank UK: Dissemination of research data using EPrints

Simon Coles, School of Chemistry, University of Southampton
Overview

• Scholarly communications in Chemistry
  ▪ Data, information, workflows and provenance

• The data publication bottleneck
  ▪ e-Science and chemistry

• eBank UK
  ▪ Information architecture, data flow and interoperability

• Challenges for the future
  ▪ Expansion into other disciplines and data formats
The scholarly knowledge cycle.

Data creation / capture / gathering: laboratory experiments, Grids, fieldwork, surveys, media

Research & e-Science workflows

Presentation services: subject, media-specific, data, commercial portals

Data analysis, transformation, mining, modelling

Resource discovery, linking, embedding

Aggregator services: eBank UK

Repositories: institutional, e-prints, subject, data, learning objects

Learning object creation, re-use

Learning & Teaching workflows

Institutional presentation services: portals, Learning Management Systems, u/g, p/g courses, modules

Data curation: databases & databanks

Resource discovery, harvesting, embedding

Deposit / self-archiving

Publication

Validation

Searching

Harvesting

Embedding

Quality assurance bodies

Linking

Resource discovery, linking, embedding

Deposit / self-archiving

Peer-reviewed publications: journals, conference proceedings

Validation
Current chemistry publishing protocols

Ideas and interpretations

A challenge for green chemistry: designing molecules that readily dissolve in carbon dioxide

Hooks into the literature

Results & derived data

Raw data!
Chinese Journal of Chemistry
ISSN: 1001-804x
Subject: Chemistry (General) --- Analytical Chemistry --- Inorganic Chemistry --- Organic Chemistry
Publisher: The Chinese Chemical Society and Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences
Language: English, Chinese
Keywords: chemistry
Start Year: 2000

Croatica Chimica Acta
ISSN: 1334417x
Subject: Chemistry (General)
Publisher: Croatian Chemical Society
Language: English
Keywords: chemistry
Start Year: 1998

Current Science
ISSN: 00113891
Subject: Chemistry (General) --- Biology --- Environmental Sciences
Publisher: Current Science Association; Indian Academy of Sciences
Language: English
Keywords: chemistry, science, environmental science, biology
Start Year: 1998

Edeltrica Quimica
ISSN: 01004670
EISSN: 16784618
Subject: Chemistry (General)
Publisher: Fundação Editora da Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP
Language: English, Portuguese
Keywords: Earth and Earth Sciences
The data deluge

Data
Overload!

How do we disseminate?

EPSRC National Crystallography Service
Establishing common ground…

• Understand the data creation process
• Terminology and definitions
  – Data
  – Metadata
  – Datafile
  – Dataset
  – Data holding
• Different views
  – Digital library researchers, computer scientists, chemists
  – Generic vs specific
  – Modeller vs practitioner
• Aim for a common ontology
• Modelling the domain
• Creating a metadata schema
Crystallography workflow

- **Initialisation**: mount new sample on diffractometer & set up data collection
- **Collection**: collect data
- **Processing**: process and correct images
- **Solution**: solve structures
- **Refinement**: refine structure
- **CIF**: produce CIF (Crystallographic Information File format)
Deposition into the archive
An Archive entry

2-(N-Ferrocenylmethylcarbamoyl)-5-(methoxycarbonyl)-3,4-diphenylpyrrole

Susanne L. Huth, Michael B. Hursthouse, Simon J. Coles, Mark E. Light, Peter N. Horton, Phil A. Gale, G. Danault and C. N. Warriner.

University of Southampton

Chemical formula: C30 H26 Fe N2 O3

Crystal system: Orthorhombic

Space group symbol: Pbc a

Cell length a: 6.081(4)

Cell length b: 24.850(3) (15)

Cell length c: 31.120(8)

Cell angle alpha: 90.00

Cell angle beta: 90.00

Cell angle gamma: 90.00

Data collection temperature: 120(2)

Available Files

Final Result

02soc064.CIF  19k

02soc064.cml  8k

02soc064_checkcif.html  14k

Refinement

02soc064.REF  9k

Solution

02soc064.PRP  6k

Processing

02SOT04.HTM  6k

02soc064.HKL  338k

Other Files

02soc064.DOC  113k

02soc064.LST  43k
Access to the underlying data
Some metadata issues

- Using simple and qualified Dublin Core
- Additional chemical information in schema for harvesting e.g. empirical formula
- Schema contains International Chemical Identifier (InChI)
- Links to all datasets associated with an experiment
- Links to individual datasets within an experiment
- Links to EPrints (and other published literature) derived from the data
- Using vocabularies specific to crystallography
- Engaging the broader scientific community to ensure different schemas are compliant and standards can emerge
Harvesting: OAIster
Linking and aggregating
Embedded in a science portal
Current situation

- Version 2.0 eBank metadata schema
- Pilot institutional e-data repository for harvesting (raw, derived, results data) using EPrints.org software
- Exports records as ebank_dc and oai_dc
- Validation of schema & discussion with International Union of Crystallography for final developments and wider deployment
- Pilot eBank UK aggregator service
- Developing search interface Version 1.0
- Testing with PSIgate physical sciences portal – embedding eBank UK
What’s next?

- Progress towards generic metadata schemas
- Validation against other schema (CCLRC Model)
- Eprints.org software: allow for more generic scientific data and schemas?
- Metadata enhancement: keywords based on knowledge of keywords in related publications?
- Investigate identifiers: International Chemical Identifier
- Explore context sensitive linking
- Full embedding into chemical and crystallographic research and publishing
- e-Learning embedding and pedagogic evaluation
- Feasibility study in related domains
Breakout Session?

• Describing non ‘Dublin Core’ terms
  ▪ Qualified Dublin Core
  ▪ Complex object formats: METS vs MPEG-21 DIDL
  ▪ Set & Friends containers

• Compliance between schemas
  ▪ One generic schema
  ▪ Develop multiple schemas

• Rights
  ▪ Use / reuse
  ▪ Publisher

• Linking & aggregating
  ▪ DOI
  ▪ Keyword ontologies
  ▪ Identifiers
  ▪ Context sensitive linking