Designing Data Repositories to Support Preservation and Publication for the Chemistry Community

Simon Coles
EPSRC National Crystallography Service
School of Chemistry
University of Southampton
Social Networks for Chemists

Google generation: new behaviour and approach
Sharing Rich Media

- Video + Paper = Pubcast
New Approaches to ‘Sharing Experiments’

- Specialised domain-oriented innovations
Formation of Open Communities

- New approaches surfacing and growing FAST
• Immediate sharing of experimental information & data
New Information Exchange Environments

• Immersive alternative to conventional browsing & interaction
A Virtual Research Environment

Adoption of “Web2.0” philosophy and methodology to promote exchange of research information: The New eScience?!

- Community networking
- Targeted audiences
- Open or closed
- Discoverable
- Multimedia/complex formats
Data Deluge

• 40 years ago a PhD student would determine about 3 crystal structures during the course of their study – this can now be easily achieved in a day

• There are approx. 30 million known chemical compounds
• Approx. 2 million crystal structures have been determined
• There are less than 0.5 million published crystal structures residing in (licensed) curated databases
• There are just a few thousand ‘open’ crystal structures

• The primary cause of this is the current data publication process, which is tied to journal articles and peer review
Data Publishing & Open Access

- Short communications
- Electronic only
- Rapid publication
- Open access (01/2008)
- Highly cited
- Written in CIF
- Freely available tools
- Still cannot keep up!
- Journal ‘article’ format required
- Not all crystal structures are of primary importance to the underlying chemistry: by-products / unexpected results / tracking reactions
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
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
A Thorough Approach to Preservation

A study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation

eBank-UK Phase 3: WP4
September 2006 - June 2007
Final Version Revised: 2nd September 2007

Mungo Foul
University of Bath, UK
Snell Colen
National Crystallography Service
University of Southampton, UK

KEEPING RESEARCH DATA SAFE
A COST MODEL AND GUIDANCE FOR UK UNIVERSITIES

Neil Beagie, Julia Chruszcz, and Brian Lavoie
with case studies contributed by the Universities of Cambridge, Southampton, King’s
College London, and the Archaeology Data Service University of York.

Final Report - April 2008
Prepared by:
Charles Beagie Limited
www.beagie.com
A study funded by:
JISC

The text in this document reflects the views of the author(s) and is not necessarily that of the funder(s) or any other organization.

A thorough approach to preserving research data requires a well-thought-out strategy that takes into account the specific needs and challenges of the data being preserved. The importance of curation and preservation issues in the eCrystals Data Repository and Proposed Federation is highlighted in this study, which presents a comprehensive approach to ensuring the long-term viability of research data. The document outlines the key components of an effective data preservation plan, including the identification of data types, the development of a preservation strategy, and the implementation of technical solutions. The inclusion of case studies from various universities provides practical insights and demonstrates the applicability of the proposed guidance. This document serves as a valuable resource for researchers, data managers, and institutions seeking to develop robust strategies for preserving the integrity and accessibility of their research data.
Scaling Up: A Community Solution

Interviews & analysis

**Synthesis**: IR Policy & Practice, Laboratory Practice & Workflows, Technical Interoperability & Standards, Metadata Schema & Application Profiles, Semantic Interoperability, Data Citation, Identifiers & Linking, Federation Architectures & Third Party Services, Rights & Licensing, Data Quality & Validation, Preservation, Curation & Sustainability

**Recommendations, commentary**

**Matters Arising**: Diverse lab practice, LIMS and proprietary formats, Data policy should reflect lab practice & institutional model, Data quality criteria/validation, “Prior publication” problem, We need scalable assignment of “terms” for data discovery, No discipline preservation model

Scaling Up: Towards a Federation of Crystallography Data Repositories

**Document details**

<table>
<thead>
<tr>
<th>Author</th>
<th>Liz Lyon, Simon Coles, Monica Duke, Traugott Koch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>12th May 2008</td>
</tr>
<tr>
<td>Version</td>
<td>1.0 Final</td>
</tr>
<tr>
<td>Document Name</td>
<td>ebank-phase3-report-final.doc</td>
</tr>
<tr>
<td>Notes</td>
<td></td>
</tr>
</tbody>
</table>

http://wiki.ecrystals.chem.soton.ac.uk
n = 2 d sin θ
General Chemistry Issues: Data Generation

Synthesis

Characterisation
Shortfalls in Data Management

“Data from experiments conducted as recently as six months ago might be suddenly deemed important, but those researchers may never find those numbers – or if they did might not know what those numbers meant”

“Lost in some research assistant’s computer, the data are often irretrievable or an undecipherable string of digits”

“To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data”

“Data from Big Science is … easier to handle, understand and archive. Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science.”

Spectroscopic analysis is often performed to ensure a reaction is proceeding according to plan – as a result <5% are published (via a process with heavy information loss).
A New Kind of Electronic Lab Notebook

the Smart Tea Project
A General Chemistry Laboratory Repository

Create new compound (parent record)

Add new experiment type

Add metadata and upload data files
A General Chemistry Laboratory Repository

• Probity: A process to assert originality of a data record
Analysis & Discussion: Blogging Experiments

A repository can...

- Allow one to **put, store and get**
- Provide **search** and **browse** functionality
- **NOT** provide the **presentation** and **discussion** functions essential to working up a scientific study

![Investigations into neutral drift](image-url)
Facilitating Research

- Enables ‘geographically distributed collaborative research’
- Can be open or private
- A useful approach for sharing ‘failed’ experiments?

---

**PCR of beta-galactosidase third attempt** by Jennifer Hale
December 2006 @ 11:10

Unfortunately the purification appears not to have gone well. Though I also can’t get any consistency from the figures given by the nano-drop. These are the results I got:

<table>
<thead>
<tr>
<th></th>
<th>reading 1</th>
<th>reading 2</th>
<th>reading 3</th>
<th>reading 4</th>
<th>reading 5</th>
<th>reading 6</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR product before*</td>
<td>282.3 ng/μL</td>
<td>283.4 ng/μL</td>
<td>281.1 ng/μL</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>282.3 ng/μL</td>
</tr>
<tr>
<td>PCR product after*</td>
<td>7.8 ng/μL</td>
<td>12.9 ng/μL</td>
<td>17.6 ng/μL</td>
<td>85.4 ng/μL</td>
<td>22.4 ng/μL</td>
<td>12.8 ng/μL</td>
<td>?</td>
</tr>
</tbody>
</table>

*Both reactions combined together after PCR

I’m going to do another PCR again. That step is working really well. I’m just not sure what to do about purifying it. The only other thing I can try is eluting in TE buffer rather than water (which it says you can also elute into)

In this purification I used preheated water and followed the instructions closely. Perhaps the DNA will elute into TE more effectively.

---

**Re: PCR of beta-galactosidase third attempt** by David Neylon
14th December 2006 @ 18:32

I would definitely compare these on a gel so as to see whether it is just the nanodrop that is the problem. It might help also if you are explicit about how much solution you are trying to purify and what the final volume is.
Automatic Blogging by Machines
Automatic Blogging by Sensors

- Continuous log of ‘environmental’ conditions in a laboratory
- Instant detection of erroneous events
- Correlate with inconsistencies in datasets
Comments and Collaborative Tools

• Annotation tools allow comments and foster collaboration and / or communication
• Need for more advanced Blog tools / technology around data
Packaging and Interoperability

• New moves in Digital Libraries community to enable distributed repositories to exchange content
• OAI-ORE (Open Archives Initiative – Object Reuse and Exchange)
• [http://www.openarchives.org/ore/](http://www.openarchives.org/ore/)
• Describes an aggregation of objects in an exchangeable format
• Microsoft funded eChemistry testbed project
Towards a New Model for Chemical Information Exchange
A solid foundation for Open/Self-Publishing of Chemistry Data???

Thanks to:

• Jeremy Frey, Andrew Milsted, Richard Stephenson, Cameron Neylon, Jamie Robinson, Steven Wilson

• Dave DeRoure, Les Carr, Chris Gutteridge, Tim Myles-Board, Arouna Woukei

• Liz Lyon, Rachel Heery, Monica Duke, Michael Day, Traugott Koch, Manjula Patel