The data explosion

...and the need to manage diverse data sources in scientific research

Simon Coles (s.j.coles@soton.ac.uk)

Director, UK National Crystallography Service
Why manage?

• Volume
  – Day to day coping
  – How to “publish”??

• Scientific Responsibility
  – Accurate recording of the whole experiment
  – Archival & curation
  – Enabling future science

• Diversity
  – Heterogeneous data
  – Disciplinary differences
  – Institutional boundaries

• Accountability
  – Auditing and charging
  – Funder mandates
What can structures be used for?

• For example...
  – Materials science – link to properties, starting point for calculations
  – Insight into / control of crystallisation & crystallisability
  – Starting point for computational chemistry

• Data management is key, because it's all about context and linking...

• Good crystallographic practice requires information about the experiment not covered wholly by CIF

• Not possible to predict how a crystal structure might be used in the future – measure everything you possibly can!
Increasing complexity

- Even for pure crystallographic studies, volumes of data are getting bigger.
- ‘Informatics’ approaches being adopted in order to develop fundamental understanding, rules, etc.
- Must be able to reliably evaluate quality (provenance) to incorporate results of others.
The whole research process
How many standards?!
Data management (eco)systems

- Publication Repository  [Results Data]
- LIMS  [User & Sample Data]
- ELN  [Supporting Experiment Data]
- Datastores  [Underlying Diffraction Data]
- Desktop/Laptop  [Everything!]
- iPad/iPhone  [Anything!]

- How to integrate these??
Structured Data.

i.e. using LIMS, Datastores & Repositories
LIMS: Lifecycle Management

Proposal → Approval → Admin → Submission → Analysis → Experiment → Publishing → Reporting

Application for an Allocation
Period: 1st May 2013 - 31st October 2013

Personal Details
Name: Dr Mike Coogan
Email Address: m.coogan@lancaster.ac.uk
Department: School of Chemistry
Institution: Lancaster University
Address: BL Faraday Building
Lancaster University
Lancaster LA1 4YB
United Kingdom

Funding
Funding Source: Not Funded
Local Facilities: No
Reason(s) why additional facilities are required: No local facilities, this is a new department just starting up and we intend to buy a diffractometer, this will not be until next year at the earliest.

Last Allocation

<table>
<thead>
<tr>
<th>FSA</th>
<th>DSG</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Could you indicate the percentage of these:
- Have not been worked up
- Have been fully refined
- Have been written up into a report or thesis
- Have been published (please give full details of paper(s) in the next section)

Outline of Scientific Program

We have developed a range of complexes based around luminescent transition metal fragments which are useful in cell imaging applications. Some of these are macromolecular structures which can act as hosts for smaller molecules or ions and the luminescence is in many cases modulated upon encapsulation. As well as expanding the range of metals which we can use in cell imaging to the early TMs (Zr, Hf) as well as the more traditional late TMs we are looking at new host macromolecules: hetero-cycle appended calixarenes which form interesting hydrogen bonded structures as well as complexes with a range of transition metals; also complexes based around expanded structures based on polyglycylide in dimeric, trimeric and larger assemblies. All these complexes show interesting photophysical properties, e.g. acting as sensors, acting as imaging agents. The balance between crystallinity and solubility/solubility (essential for crossing cell membranes) is difficult to build in or predict (especially in the macromolecules) with such macromolecules and in many cases it is difficult to obtain high quality single crystals of these species, and thus as well as access for data collections only, we have requested full structure solution for difficult cases where the expertise of the service will be essential.

Publications

Additional Information

Please accept my apologies for the lack of samples outputs for the last period this is a result of the development of the new chemistry department at Lancaster having been more complex that originally anticipated, so no lab work has been possible until this week as major refurbishments were undertaken to bring the facilities to standard.

Attached Files
LIMS: Lifecycle Management

- Proposal
- Approval
- Admin
- Submission
- Reporting
- Analysis
- Experiment
- Publication
LIMS: Lifecycle Management

Proposal → Approval

Admin → Submission

Publication → Reporting

Analysis → Experiment

http://ecrystals.chem.soton.ac.uk
Management Across Boundaries

- Management across facilities (ICAT Information Model)
The Core Scientific Metadata model forms the information model for ICAT & is designed to describe facilities-based experiments.
CSMD as a Starting Point

- Forms the basis for extensions:
  - To derived data
  - To laboratory based science
  - To secondary analysis data
  - To preservation information
  - To publication data
But every crystallographic experiment has ‘unstructured’ data
Getting Mobile in the Lab

Package & Sample Tube

Bulk Sample & Manipulation

 Mounted Sample

Trial Diffraction Pattern
Getting Mobile in the Lab

The completed record...
Unstructured data.

Invariably the context for a crystal structure

Or

The context a crystal structure has to fit into!
Laboratory Notebooks

• All this is not a new problem...

Faraday's notebooks: the active organization of creative science

Ryan D Tweney

Faraday's notebooks constitute one of the largest and most revealing archives left to us by a major scientist. These records reveal a good deal of systematic invention and exploration of recording techniques by Faraday, work that reveals much about his thinking about science, as well as of the role of memory in general in scientific thinking.

Scientists are students—students of nature, to be sure, but, like all students, dependent for their success on taking notes. In even the most routine of scientific research, scientists must preserve external records of their work. Most externalize far more than just data, making records of their hypotheses, readings of the literature, wild speculations and the like. Thus, scientific diaries, laboratory notebooks, indeed the entire range of recording techniques, constitute an important topic for a full understanding of just what scientists do (Holmes 1987).

The study of Faraday's notebooks has left us a rich documentary legacy of thought that exists for perhaps any other scientific figure in history. Faraday's daily laboratory notebooks, diaries and commonplace books, almost all of which were carefully bound by Faraday himself, and almost all of which are held in the Archives of the Royal Institution of Great Britain and the Institution of Electrical Engineers are a rich source for the historian. As a rough guess, he left us records of about 30,000 experiments, both successful and unsuccessful, as well as a large number of speculative ideas, bibliographies, index cards, scrap-books, etc. etc. What is known as 'the' Diary has now been published (Martin 192–6). Though only a part of the archival holdings, this work does cover most of his famous discoveries (and lots of lesser ones as well). But it is very much a laboratory diary, and thus gives a somewhat misleading picture of the whole. This paper will present a broader perspective, one that sheds light on the entire range of Faraday's records. In this way, we will be able to gain some insights into the uniquely creative mind of a genius.

Why are Faraday's records so extensive? In part, it is because Faraday was meticulous of his own memory (see Williams 1965, pp.47, 491–501; Hare 1974). Faraday more than once repeated an experiment that he had earlier reported as apparently forgotten about, and his use of elaborate memory-retention devices (see below) makes a similar point.

Memory weaknesses aside, however, Faraday was very much part of a cultural tradition of ideas deriving from John Locke (1632–1704) that placed central importance on memory as an essential cognitive process in the acquisition of knowledge. In the Lockean theory of ideas, knowledge is built upon severed sensations and ideas: imagination, reason, and such things as Descartes' 'clear and distinct ideas' are built on the foundation of memory. Locke himself saw that this account had strong implications for the importance of memory aids. If memory was to be the foundation of knowledge, then the weaknesses of human memory—forgetting, distortion, the vagaries of the retrieval process—constituted serious problems. True knowledge needed accurate memories, and thus required the use of accurate records.

Paradox derived his Lockean view of memory from Isaac Watts (1674–1748) whose Improvement of the Mind (1989–1741) was read by Faraday in 1809 and credited by him with 'having taught me to think'. It inspired Faraday's first surviving memory aid, the Philosophical Miscellany (Faraday 1809–10). His use of memory aids evolved subsequently during the course of his career, culminating, after 1813, in the mature recording and retrieval system that is described in the next section. It is worth noting that one important respect in which Faraday was not Lockean was in his reliance on the power of experiment. Neither

own Diary entries or on his speculative thinking. His line of the retrieval aids was a dynamic one.

Looking at the retrieval sheets once suggested a kind of experiment to me. I took one of the sheets and photocopied the relevant sections referred to in the Diary, pasted these up on long sheets, and tried to read the result, to see if I could get insights into what Faraday was after in making the retrieval sheet. Unfortunately, the result made little sense, even when it dealt with a part of the Diary that I was fairly familiar with. The reason is clear: Faraday didn't just need references to the particular facts recorded in the Diary, he needed cues to the entire context of his memories about the incidents in question. In fact, when I abandoned the long paste-ups that I had made and simply read the Diary in the order in which these references were made, the whole made much more sense. My eye could pick up closely associated references in adjacent parts of the Diary, and these were helpful in reconstructing the context. How much richer this would have been for Faraday, for whom these adjacent 'reminders' would have been full of the concrete memories, images, emotions and so on of the original incident.

Figure 2. An example of a retrieval sheet (IEE Archives, Misc. MSS SC2).

Figure 3. A 'paste-up' (IEE Archives, Misc. MSS SC2).

Among the many types of retrieval sheets that survive, some stand out because they don't refer directly to the Diary itself. Instead, they seem to be indexes of indexes, or 'memes' to use a modern computerese term. These are striking because they suggest that Faraday used so many retrieval devices that he needed to organize these as well. Anyone who has lost a file on a hard disk will appreciate the need!

Constructing retrieval aids of the sort described above can be seen as analogous to an encoding process in which a retrieval cue (an index tag, say) is encoded in a physical form with an address to the full diary entry. Once the tag exists, it can be used in one of two ways. First, obviously, it can serve as a finding aid for the retrieval of specific diary entries. Second, and perhaps even more importantly, it can serve as a mnemonic cue useful in the structuring of diary-based knowledge. By sorting slips one can impose one or another organization on diary entries, and vary that organization in the service of other goals.

The development of the memory aids

So far, the discussion has focused on Faraday's use of retrieval aids in their mature form, i.e. after 1832. By looking at earlier records, however, it is possible to trace the evolution of Faraday's use of external memory aids. Faraday's earliest surviving notebook is the Philosophical Miscellany kept in 1809–10, which was used mainly to copy out interesting things Faraday had read. It contains an ordinary alphabetical index evidently prepared after the book was completed; while such an index is useful, note that it can only be prepared after all
Synthesis of 4-substituted methylidene oxindoles
Project E-Lab notebook for the synthesis of the 4-substituted methylidene oxindole from oxindole and their corresponding aromatic aldehydes.

Single Crystal X-ray crystallography
27th June 2012 @ 23:09
Single-crystal X-ray diffraction analyses were performed using a Bruker Apex II CCD diffractometer mounted at the window of a Bruker K8301 rotating anode (MoKα = 0.71073 Å) and equipped with an Oxford Cryosystems cryostream device. Data were processed using the Collect package and unit cell parameters were refined against all data. An empirical absorption correction was carried out using SADABS. The structures were solved by direct methods using SHELXS-97 and refined on F^2 for full-matrix least-squares refinements using SHELXL-97. All non-hydrogen atoms were refined with anisotropic displacement parameters. All hydrogen atoms were added at calculated positions and refined using a riding model with isotropic displacement parameters based on the equivalent isotropic displacement parameter (Ueq) of the parent atom. Figures were produced using Olex2.

Graham Tizzard | View Source | Analytical Procedures | Comments (0)

MS Spectrum of (3E)-3-(4-Nitrobenzylidene)-1,3-dihydro-2H-indol-2-one
6th May 2017 @ 18:16
spectroscopic Method: MS-ESI
Substituent: Nore
MS Spectrum of (3E)-3-(4-Nitrobenzylidene)-1,3-dihydro-2H-indol-2-one:

The mass spectrum of (3E)-3-(4-Nitrobenzylidene)-1,3-dihydro-2H-indol-2-one has been obtained by negative electrospray ionization (ESI). The peak at m/z = 265.2 confirms the molecular mass of this compound as the molecular ion gains a proton.

Interpretation of MS Spectrum of (3E)-3-(4-Nitrobenzylidene)-1,3-dihydro-2H-indol-2-one:

<table>
<thead>
<tr>
<th>Peak Position</th>
<th>Diff. between molecular mass</th>
<th>Suspected molecules or ion</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>265.2</td>
<td>265 - 264 = 1</td>
<td>(M+H)+</td>
<td>Parent compound plus a proton</td>
</tr>
</tbody>
</table>

Nicola Knight | View Source | Spectroscopic Data | Comments (0)

Pictet–Spengler route to Praziquantel
Synthesis of intermediate and derivatives of PQ

Hydrolysis of MNR11–18 to give MNR26–7
23rd April 2013 @ 01:09
Mnr: 21–30
As for
Hydrolysis of MNR11–16 to MNR26–4
Hydrolysis of MNR11–17 to MNR26–5
Starting material from Synthesis of MNR11–18

Hazard and Risk Assessment:

<table>
<thead>
<tr>
<th>Compound</th>
<th>MW</th>
<th>Density (g/ml)</th>
<th>Viscosity (mPa s)</th>
<th>Volume (ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNR11–18</td>
<td>371.46</td>
<td>0.87</td>
<td>100.02</td>
<td>3.96</td>
</tr>
<tr>
<td>MNR26–7</td>
<td>362.30</td>
<td>0.87</td>
<td>100.02</td>
<td>3.96</td>
</tr>
</tbody>
</table>

Procedure:
MNR11–18 (3.66 g, 10.63 mmol) was dissolved in DCM (21 mL) and HCI (1M, 110 mL) and heated to reflux for 2.5 hours. The solution was allowed to cool to room temperature then cooled in an ice bath. Washed with NaOH pellets (approx 8 g, pellets used to minimise volume of aqueous material) to pH 12–13 and extracted with DCM (4 x 100 mL). The organic fractions were combined, dried over magnesium sulphate, filtered and concentrated under reduced pressure to give an orange syrup solid (1.08 g).

TLC
Reaction mixture after 7.5 hours run in 100% DCM.

2013-04-28
13:15:20.png
Imposing structure – Planning & Enactment Ontology

• Plan (Prospective provenance)

• Enactment (Retrospective provenance)

• Realisation
oreChem Plan for eCrystals

- Machine-readable representation of methodology
- Describes requirements for software and data products
Bringing it all together.
Structures in context of compound, property, etc data

- RDF (ChemAxiom)
- ChemSpider ID: 75956
- Molecular Formula: C_{41}H_{29}O_{7}
- Average Mass: 688.305603 Da
- Monoisotopic Mass: 688.131073 Da
- Systematic Name: 2,3,5,6,8,9,10,11,12-Octahydro-1,4,7,10,13-benzopentaacoxypentadecine
- SMILES and InChI
- Cite this record

Associated Hyperlink: http://dx.doi.org/10.3737/ocrystals.chem.oxon.ac.uk/145

Comments: Structure already known, but accurately redetermined for a local research project. Citation: Rousay, Es Simon J. and Hursthouse, Michael B. (2004) University of Southampton, Crystal Structure Report Archive.

Approved No. Submitted by: sharped
Structures in the context of ‘traditional’ publishing

- ELN providing Electronic Supplementary Information for conventional publication (Chemistry Central: accepted)
DOI’s for Data

- DataCite Institutional registration of DOI’s
- Promotes rigorous ‘self publishing’ of data
- Soton exemplars
- eCrystals (Repository/structured data)
DOI’s for Data

- LabTrove (unstructured data)
- Locked down HTML export of selected records
Sample Information Management System

• A standard/format for crystallographic sample and experiment data management and archival

• Supported by CrystalClear and NCS Portal, providing interaction between facility, instruments and CIF, ImgCIF etc
• 3 layer metadata model for description, export & packaging

• This is the first (information) layer – leads into knowledge

• Published through Dial-a-Molecule at [http://wp.me/p2JoQ6-xF](http://wp.me/p2JoQ6-xF) & submitted to J. ChemInf
Thanks