Crystallography and the Semantic Web

Peter Murray-Rust
University of Cambridge
Open Knowledge Foundation

IUCr COMCIFs workshop, Warwick, 2013-08-25
Themes

• What is OPEN?
• Governments and Funder mandates for Open Data.
• The Semantic Web.
• Crystallography Open Database and Crystaleye
• Recommendations for Open Crystallography

Funding includes JISC, Unilever, EPSRC.
“[we] owe a huge debt to the International Union of Crystallography (IUCr) and Brian McMahon... Quite simply they are the best semantic scientific publishers of the current century. They also have the best community-base for scientific publishing that I know. ... ...they have steadily invested in the information infrastructure (ontology) of their discipline. And it’s been a community effort...

Murray-Rust and Rzepa: Journal of Cheminformatics 2012, 4:14 http://www.jcheminf.com/content/4/1/14
Semantic authoring IUCr

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Linked **Open** Data – the world’s knowledge

RDF triples

**Music, Art, Literature**

**Social**

Knowledge bases

DBPedia

**GOV.**

GOV.uk

**Ontologies**

BIO

very little physical science 😞

“Which Rivers flow into the Rhine and are longer than 50 kilometers?” or “Which Skyscrapers in China have more than 50 floors and have been constructed before the year 2000?”

“Which countries where tropical diseases are endemic have published structures of chiral natural products?”

Open Crystallography?
The scientist’s amanuensis

- "The bane of my life is doing things I know computers could do for me" (Dan Connolly, W3C)

Example: A semantic amanuensis could
- Give me a **daily digest of zeolite** papers
- **Extract all the crystal structures** from them
- Compute **physical properties** with GULP and NWChem
- Compare the results **statistically**
- Preserve and **distribute** the complete operation
- Prepare the results **for publication**

*The semantic web is having a personal amanuensis*
Semantics for Crystal Materials Informatics

- 1993 COMCIFS
- 1994 1st WWW Conference, Chemical MIME
- 1994 Chemical Markup Language (Henry Rzepa, PMR)
- 2001 UK eScience programme, eMinerals
- 2005 Materials Grid (Martin Dove group)
- 2006 Blue Obelisk (Open Source chemistry)
- 2006 Polymer Informatics (Unilever, Nico Adams)
- 2009 Chem4Word, OREChem (Microsoft Research)
- 2011 PNNL (US) meetings and visit
- 2012 Semantic Physical Science (Cambridge)
- 2013 CSIRO Materials Science Informatics
Open Definition

• “A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.”

<table>
<thead>
<tr>
<th>OPEN</th>
<th>NOT OPEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>CCDC, ICSD</td>
</tr>
<tr>
<td>COD, Crystaleye</td>
<td></td>
</tr>
<tr>
<td>RSC/ACS/IUCr CIFs</td>
<td>Elsevier/Wiley/Springer CIFs</td>
</tr>
<tr>
<td>Acta Cryst E</td>
<td>Acta Cryst ABCD (default)</td>
</tr>
<tr>
<td>CIF dictionaries</td>
<td></td>
</tr>
</tbody>
</table>
[at Research Data Alliance], we are entering a new “era of open science”, which will be “good for citizens, good for scientists and good for society”. She explicitly highlighted the transformative potential of open access, open data, open software and open educational resources – mentioning the EU’s policy requiring open access to all publications and data resulting from EU funded research.

http://blog.okfn.org/2013/03/21/we-are-entering-an-era-of-open-science-says-eu-vp-neelie-kroes/#sthash.3SWDXDE6.dpuf
Current scientific information flow... is broken for data-rich science

Non-semantic data

Human input

PDF

Lineprinter output

Text files

Data extraction difficult and incomplete

Human input

Current scientific information flow... is broken for data-rich science

Non-semantic data

Human input

PDF

Lineprinter output

Text files

Data extraction difficult and incomplete

Human input

Current scientific information flow... is broken for data-rich science

Non-semantic data

Human input

PDF

Lineprinter output

Text files

Data extraction difficult and incomplete

Human input

Current scientific information flow... is broken for data-rich science

Non-semantic data

Human input

PDF

Lineprinter output

Text files

Data extraction difficult and incomplete

Human input
Semantic network closes the loop

Data available for e-science and re-use

Data mined from document
The network grows autonomously

Human-machine

Human-human

Machine-human

Machine-machine
TimBerners-Lee’s Open data

http://5stardata.info

★
make your stuff available on the Web (whatever format) under an **OPEN** license

★★
make it available as **structured data** (i.e. NOT PDF)

★★★
use **non-proprietary formats** (e.g., CSV)

★★★★
use **URIs** to denote things, so that people can point at your stuff

★★★★★
link your data to other data to provide context
Creating semantic content

1. Authoring tools for humans
2. Program output
3. Chemical databases
4. Content mining and Natural Language Processing (Text) (NLP)
5. Community
New: Now works on Atmospheric Chemistry Abstracts (Beta)

ChemicalTagger is an open-source tool that uses OSCAR4 and NLP techniques for tagging and parsing the chemistry literature.

To use this demo, select the type of chemistry you would like to analyse (below), enter some chemical text, and click the 'Process Text' button.

- Organic
- Atmospheric

Typical chemical synthesis

To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0°C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol). The reaction mixture was then warmed to room temperature and stirred overnight. After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate. The drying agent was filtered off and the filtrate concentrated under reduced pressure. The residue obtained was purified by flash chromatography (silica gel, 0-5% methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).
To a stirred solution of 4-hydroxypiperidine (0.97 g, 9.60 mmol) in anhydrous dimethylformamide (20 mL) at 0 °C was added 1-(bromomethyl)-4-methoxybenzene (1.93 g, 9.60 mmol) and triethylamine (2.16 g, 21.4 mmol). The reaction mixture was then warmed to room temperature and stirred overnight. After this time the mixture was concentrated under reduced pressure and the resulting residue was dissolved in ethyl acetate (40 mL), washed with water (20 mL) and brine (20 mL) before being dried over sodium sulfate. The drying agent was filtered off and the filtrate concentrated under reduced pressure. The residue obtained was purified by flash chromatography (silica gel, 0-5 % methanol/methylene chloride) to afford 1-(4-methoxybenzyl)piperidin-4-ol as a brown oil (1.70 g, 80%).

Actions:
- Filter
- Yield
- Heat
- Wash
- Concentrate
- Purify

Conditions:
- TempPhrase
- TimePhrase
- Molecules:

Could be used for analytical, crystallization, etc.
Crystallography Open Database
A grass-root initiative

From Saulius Grazulis

http://www.crystallography.net/

- Total ≈ 230,000 records
- 236 registered users
- 31 depositors (deposited at least one structure)
- In year 2012:
  - > 56,000 new structures uploaded
    (26,000 more than in 2011)
  - 24 active depositors
    (who deposited at least one structure in 2012)
Crystaleye

• A database of 200,000 crystal structures scraped from publications CIF supplemental information
• CML molecules and name-value pairs
• Re-usable as fragment base

Nick Day, Jim Downing, Sam Adams, N. W. England and Peter Murray-Rust*


http://wwmm.ch.cam.ac.uk/crystaleye
Supplemental Information (CIFs) harvested from Publications

No. of CIFs aggregated per year

Year


ACS
IUCr
RSC
ELS
As-Cl Bond lengths

Short

Long
## Structures Containing As-Cl Bonds between 2.16-2.17 Å

<table>
<thead>
<tr>
<th>Published Formula (clickable)</th>
<th>Article</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>AsCl₃</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>(C₁₆H₁₆As₂Cl₂S₄)(AsCl₃)₂</td>
<td>view</td>
<td>view</td>
</tr>
</tbody>
</table>

**Below is a diagram and some structural data:**

**P 2ac 2ab**  
P 2 1 2 1  
*a* = 9.475 Å  
*b* = 11.331 Å  
*c* = 4.296 Å  
*α* = 90.0°  
*β* = 90.0°  
*γ* = 90.0°  

The As-Cl bond length is 0.216 nm.
### Organic Structures

<table>
<thead>
<tr>
<th>Chemical Structure</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁₆H₁₅F₃N₂O₃C₁₆H₁₃F₃N₂O₃</td>
<td>view</td>
</tr>
<tr>
<td>C₁₅H₁₅F₃N₂O₄C₁₅H₁₅F₃N₂O₄</td>
<td>view</td>
</tr>
<tr>
<td>C₂₄H₂₀N₄S</td>
<td>view</td>
</tr>
</tbody>
</table>

**Notes:**
- **P 1 [P -1]**
- a = 7.800 Å
- b = 8.487 Å
- c = 11.860 Å
- α = 72.4°
- β = 80.5°
- γ = 72.5°

**Images:**
- Jmol
- Chemical diagram
Total \(\approx 230,000\) records

236 registered users

31 depositors (deposited at least one structure)

In year 2012:

>56,000 new structures uploaded

(26,000 more than in 2011)

24 active depositors

(who deposited at least one structure in 2012)
COD Letter to Editors 2012

[We] have become aware of **growing concerns** regarding the publication, preservation and quality maintenance of **crystallographic data**. However, we believe that **completely open deposition** of data and multiple checks can ensure the quality and wide availability of scientific data.

[Please] recommend to your authors that, **they also deposit their supplementary crystallographic data into the COD** when they submit scientific papers to your journals.

Being **open by its design**, the COD enables the creation of multiple mirrors and backup copies. It provides, thus, **archival storage** of scientific data with adequate reliability. **services for reviewers and editors** to facilitate the peer-review. Since our database follows the Open Access model, all material deposited into the COD is available to other databases. The COD team actually encourages the use of our data collection for any possible scientific or industrial application by putting the **database into the public domain**.
Recommendations for Open Crystallography

- Require Open Crystal Data for all publications
- Deposition of Open Data in COD
- Integrate CIF dictionaries as RDF into Linked Open Data
- Integrate COD into Linked Open Data Cloud
- CCDC/ICSD to publish RAW author CIFs Openly