Jim Downing, University of Cambridge

Dr. CrystalEye

Or: How I Learned To Stop Worrying And Love The Web

From Desktop to Data Repository

I'm going to tell you a story about a PhD in my research group, and a data system he developed, called CrystalEye. Like all good stories, there's a meta-story between the lines. Since the story teller gets to pick the meta-story, today it's about repositories.

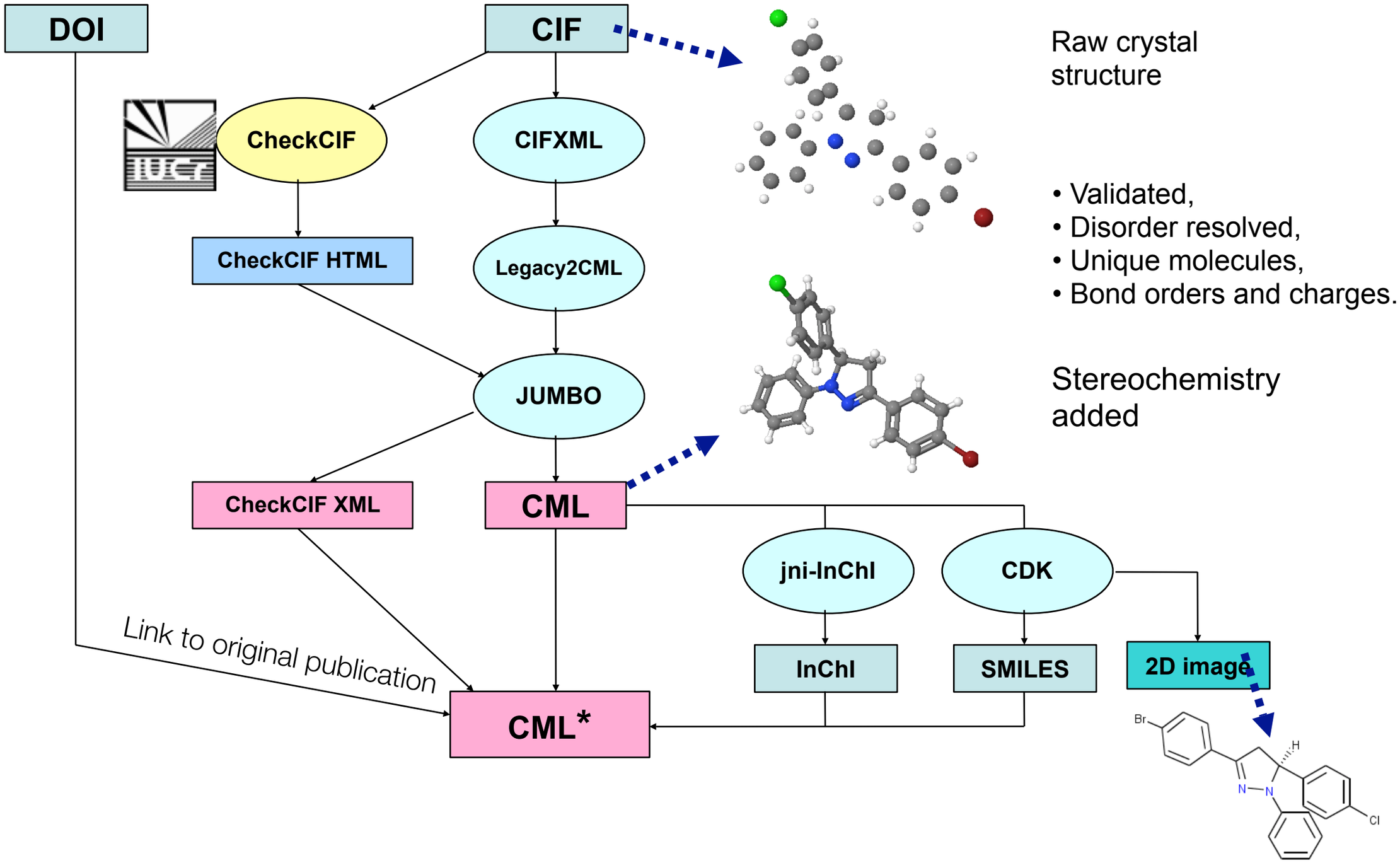
In the beginning...

Nick Day started his PhD under the supervision of Peter Murray-Rust.
Quantum computational programs to calculate molecular structures.
Compare calculated structures to actual structures, as determined by X-Ray crystallography to work out when and why the programs got it wrong.
As things transpired, the collection and publication of the X-Ray crystallography turned out to an interesting area in itself - and it's this side of Nick's PhD that interests us most today.

How to get Open crystallographic structure data (2005)

Nick needed large amounts of open structure information to compare the computational outputs with.
There was no database of open crystallographic data.
There is crystallographic data out there - on journal websites. The information comes from the websites of acta journals (especially the Acta Crystallographica family published by the IUCr) that specialize in reporting X-Ray structure determinations, and from the supporting information of other chemistry journal publications.
Nick wrote a web spider to find it and collect it.

CrystalEye Data Processing

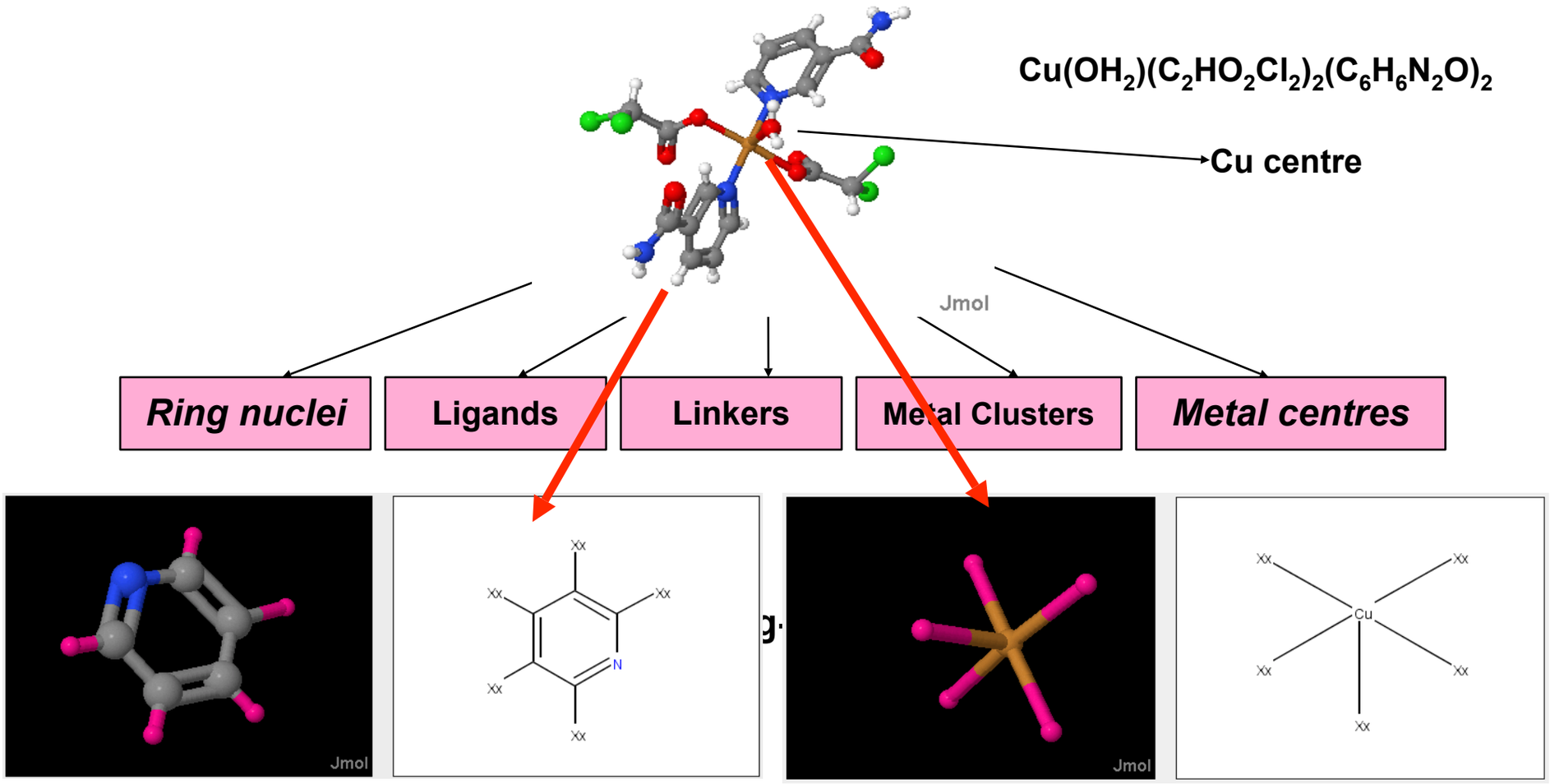


Nick needed to convert his data into Chemical Markup Language, an XML for chemical data. Being a good geek, he didn't miss the opportunity follow some interesting side lines, including developing heuristic approaches to fix and enhance missing data.

(This slide shows the processing of the crystallographic data once it has been collected by the spider. The detail isn't necessary to follow the story)

CrystalEye Data Processing 2

- Automatic generation of fragments



... and break apart molecules to form a fragment library (these can be used for an empirical approach to predicting 3D molecule structures). Nick also had web authoring skills, so he created HTML pages for the data he was collecting as part of the processing. To help himself keep tabs on the growing collection, he created a small but growing website on his desktop machine.

UNIVERSITY OF CAMBRIDGE

Crystal

- Home
- Search
- Browse Issues
- RSS feeds
- Bond Lengths
- Greasemonkey
- FAQ

Acta Crystallographica - Section E:

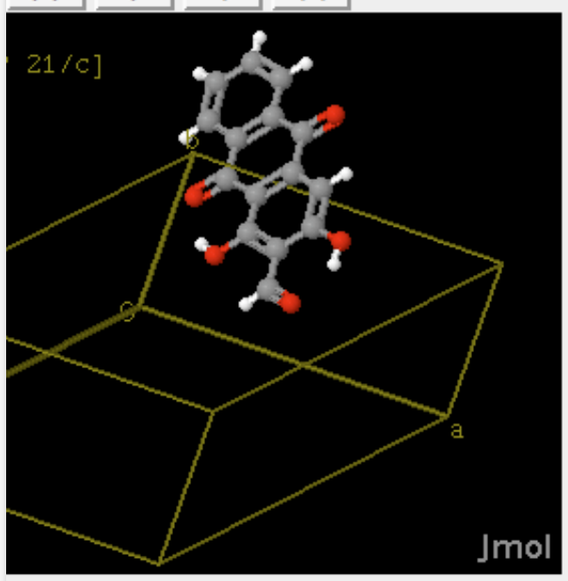
- 2008
  - Issue 03-00
  - Issue 02-00
  - Issue 01-00
- 2007
  - Issue 12-00
  - Issue 11-00
  - Issue 10-00
  - Issue 09-00
  - Issue 08-00
  - Issue 07-00
  - Issue 06-00
  - Issue 05-00
  - Issue 04-00
  - Issue 03-00

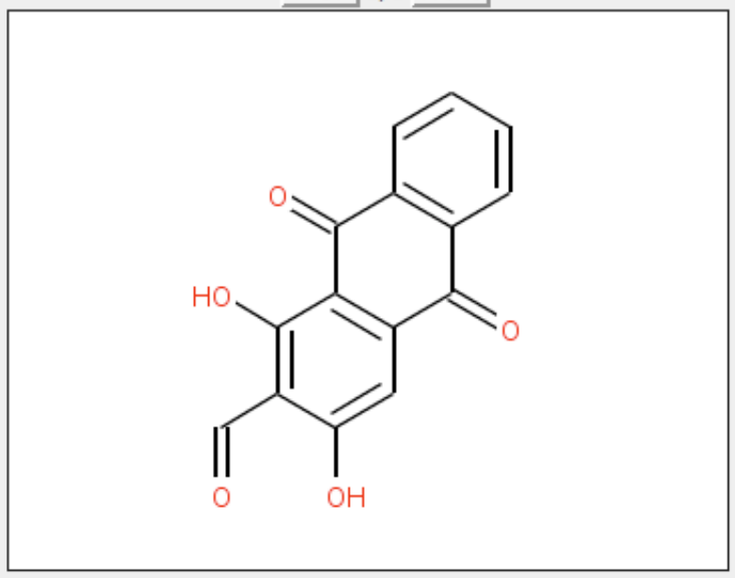
Acta Crystallographica Section E, 2008, issue 03-00

Organic Structures

Published Formula (clickable)	Article	Summary
$C_{15}H_8O_5$	view	view
$C_{14}H_{21}N_3O_2S$	view	view
$C_{19}H_{21}NOS$	view	view
$C_{16}H_{15}F_2NO_3$	view	view
$C_{18}H_{16}N_2O_3$	view	view
$C_{17}H_{28}O_5$	view	view
$C_9H_{11}FN_2O_3$	view	view
$C_9H_8N_2O_2$	view	view
$C_{12}H_{13}NO_5$	view	view
$C_{16}H_{18}N_4O_7S$	view	view
$C_{19}H_{17}ClN_4O_2S$	view	view

<<- <- -> ->> Prev 1/1 Next





Browse

Drill Down By Journal, By Issue, and then by structural sub-components

1,3-Dihydroxy-9,10-dioxo-9,10-dihydroanthracene-2-carbaldehyde

OPEN DATA

<< Table of Contents

Publisher: Acta Crystallographica
Journal: Section E
Year/Issue: 2008/03-00

Article (via DOI): 10.1107/S1600536808004169
Compound Class:organic
Date Recorded:2008-01-28

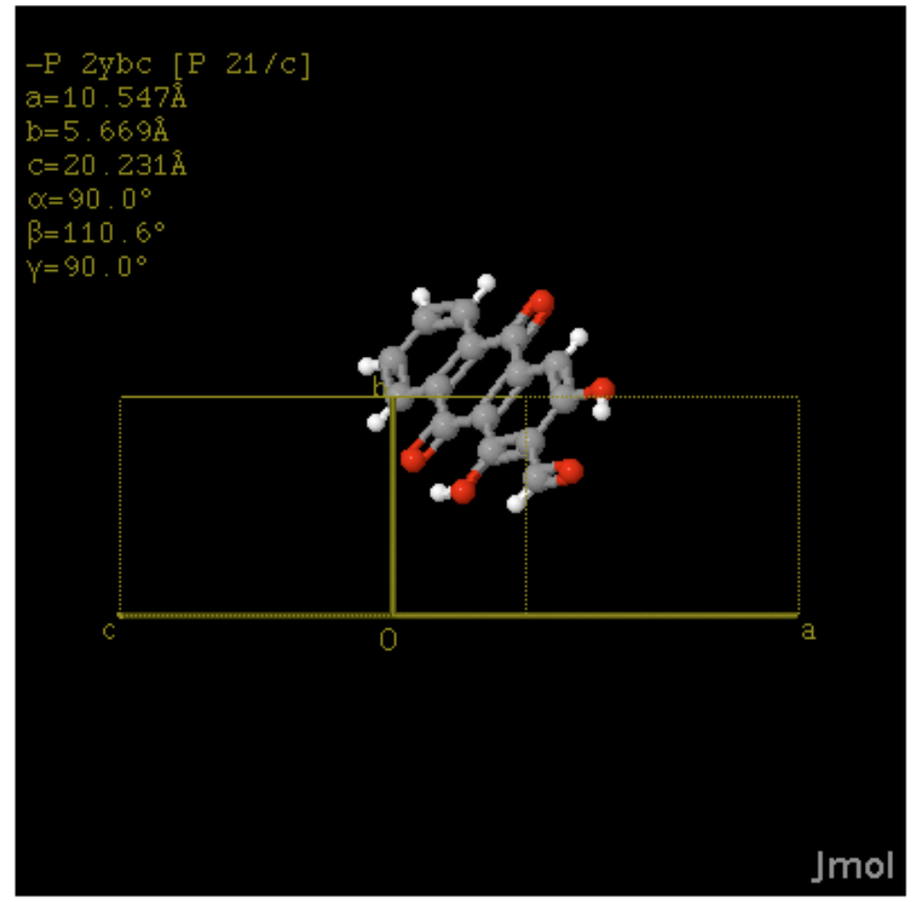
Contact Author: Retailleau, Pascal
e-mail: pascal.retailleau@icsn.cnrs-gif.fr

Data collection parameters

Chemical formula sum	$C_{15}H_8O_5$
Chemical formula moiety	$C_{15}H_8O_5$
Crystal system	monoclinic
Space group H-M	P 21/c
Space group Hall	-P 2ybc
Data collection temperature	293.0

Refinement results

R Factor (Obs)	0.051
R Factor (All)	0.081
Weighted R Factor (Obs)	0.132
Weighted R Factor (All)	0.15



Show no. of unit cells along axis:

a: 1
b: 1
c: 1

enter

Enter Jmol script:

```
load./bg2162supl_I.complete.cml.xml
```

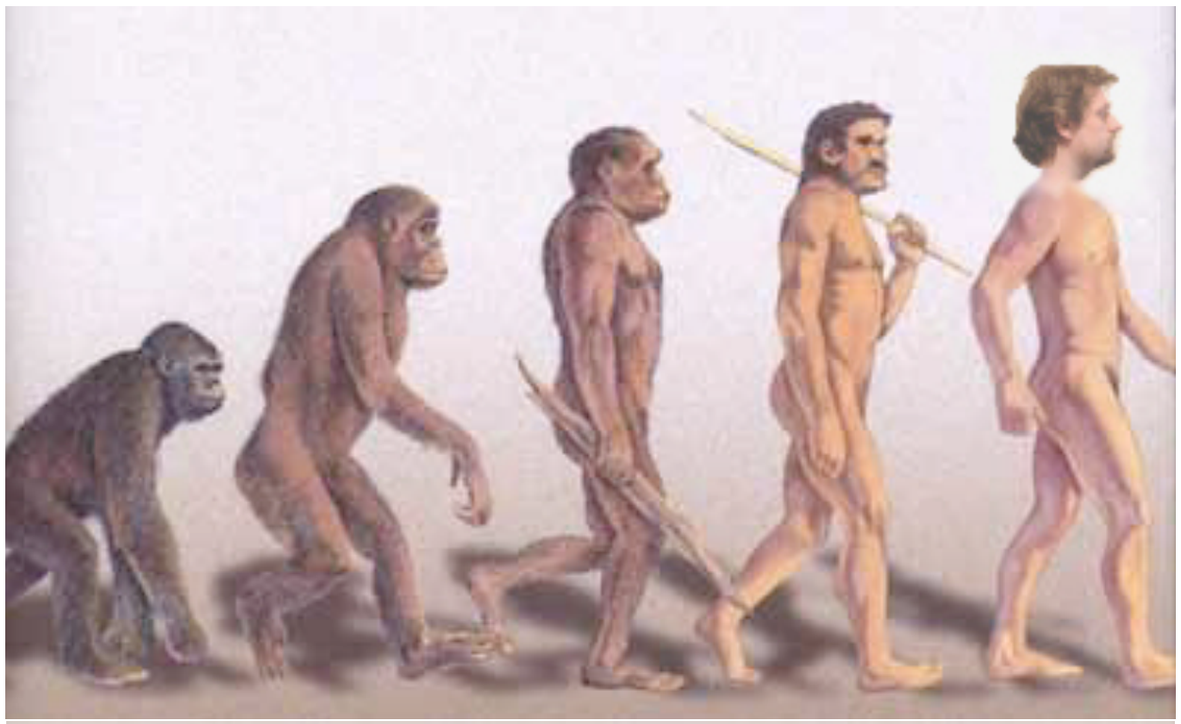
View Record

Metadata specific to data domain
The applet controls or the script loading input can be used to rotate, zoom and otherwise alter and manipulate the visualisation.

.. and so Nick had built himself a web site of his data on his desktop.

Open Data

So Nick had a growing pile of result data, and a visualisation system of Java classes that generated CML and HTML web pages. Peter Murray-Rust and Nick decided to share his results with the community. For myself as part-time system administrator this was no problem - deployment just meant a couple of cron jobs, a large file system partition and line or two of apache config.



Evolution

Sharing your data hopefully means getting users. And having users usually means getting good ideas about what to develop next. As the data set grew, Nick evolved CrystalEye, adding features to make this data collection more useful.

Feeds

By journal, compound class, atom and bond.

Once Nick got going by creating one main feed, he created per-journal feeds, feeds for each compound class, for each atom type and for each pair of bonded atom-types. Some of these are CMLRSS feeds - feeds with CML data embedded inline. If you have a chemistry enabled feed reader (e.g. the Bioclipse platform) you can browse crystaleye data directly in your reader.

CrystalEye (beta

Search:

1: Search for substructures in CrystalEye using a SMILES string.

Enter SMILES:

c1cccc1

2: Search for crystals by cell paramet

Enter cell parameters:

a ± Å

b ± Å

c ± Å

α ± °

β ± °

γ ± °

Search

CrystalEye (beta)

Searching...

Search results for SMILES string "**c1cccc1**" (in reverse chronological order) :

Showing results 1 to 25 of 1793

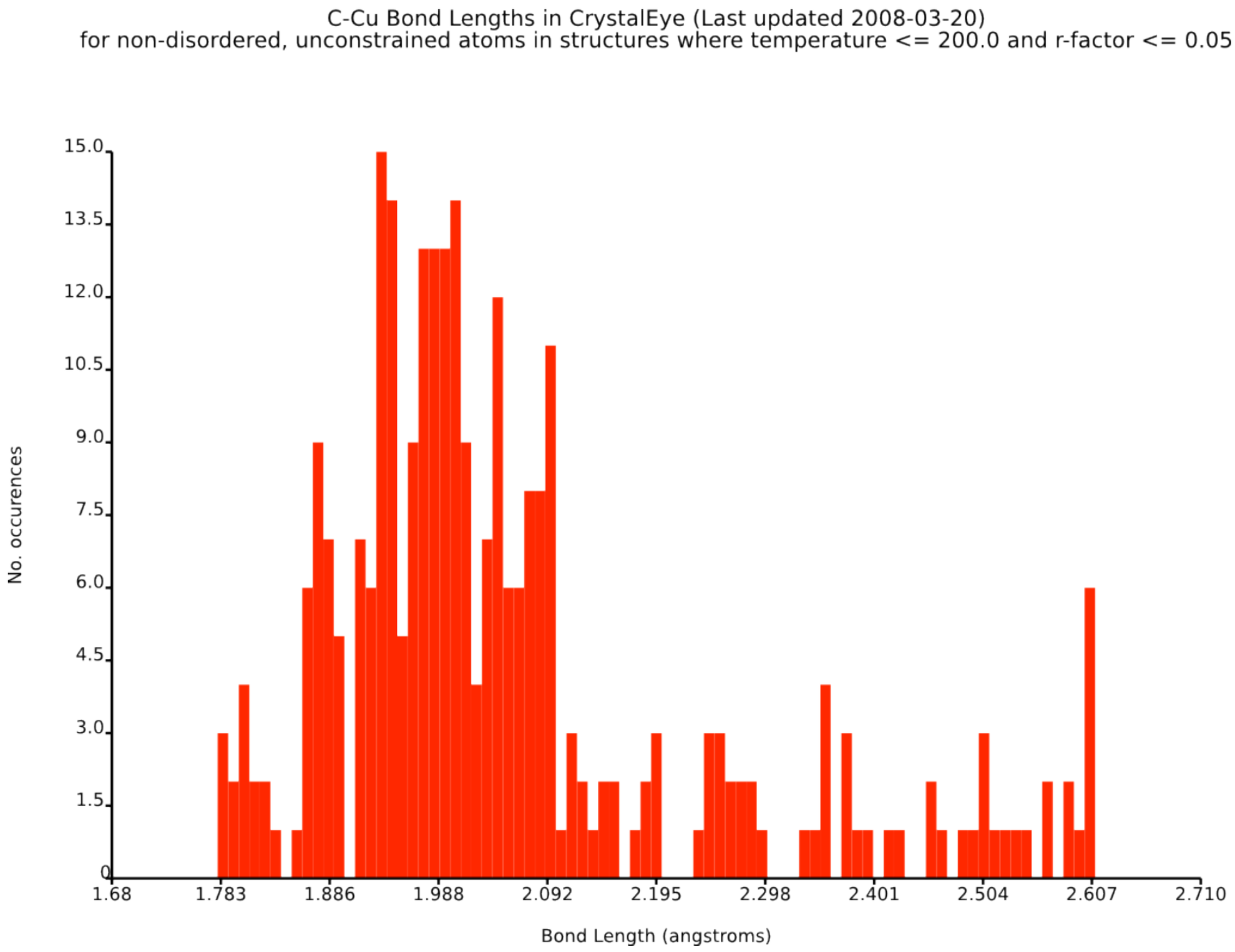
1: c1ccc(cc1)P(c1ccccc1)CCB1(C2CCCC1CCC2)=P(c1ccccc1)(c1ccccc1)CCB1(C2CCCC1C
(c1ccccc1)CCBC12C3(CCC1)C2CC3
The American Chemical Society, Organometallics, 2008, 5, article om7010886, databl

2: C=C1C2(C3CC4C(C(=CC3C14)C(C)C)=O)C1COC(C1CC2)=O.C=C1C2(C3CC4C(C(=CC
(C)C)=O)C1COC(C1CC2)=O
Royal Society of Chemistry, Chemical Communications, 2008, 10, article b718754h, d

3:
OC[C@H](CC(C(=O)C)=C1[C@@H]([C@H]2([C@@]3(CC[C@]4[C@@]5(CC[C@H](C[C@@
([C@@H]3(C[C@@H]2(O1)))))O)C)C))([CH2]))C
Acta Crystallographica, Section E, 2008, 03-00, article rz2197, datablock I

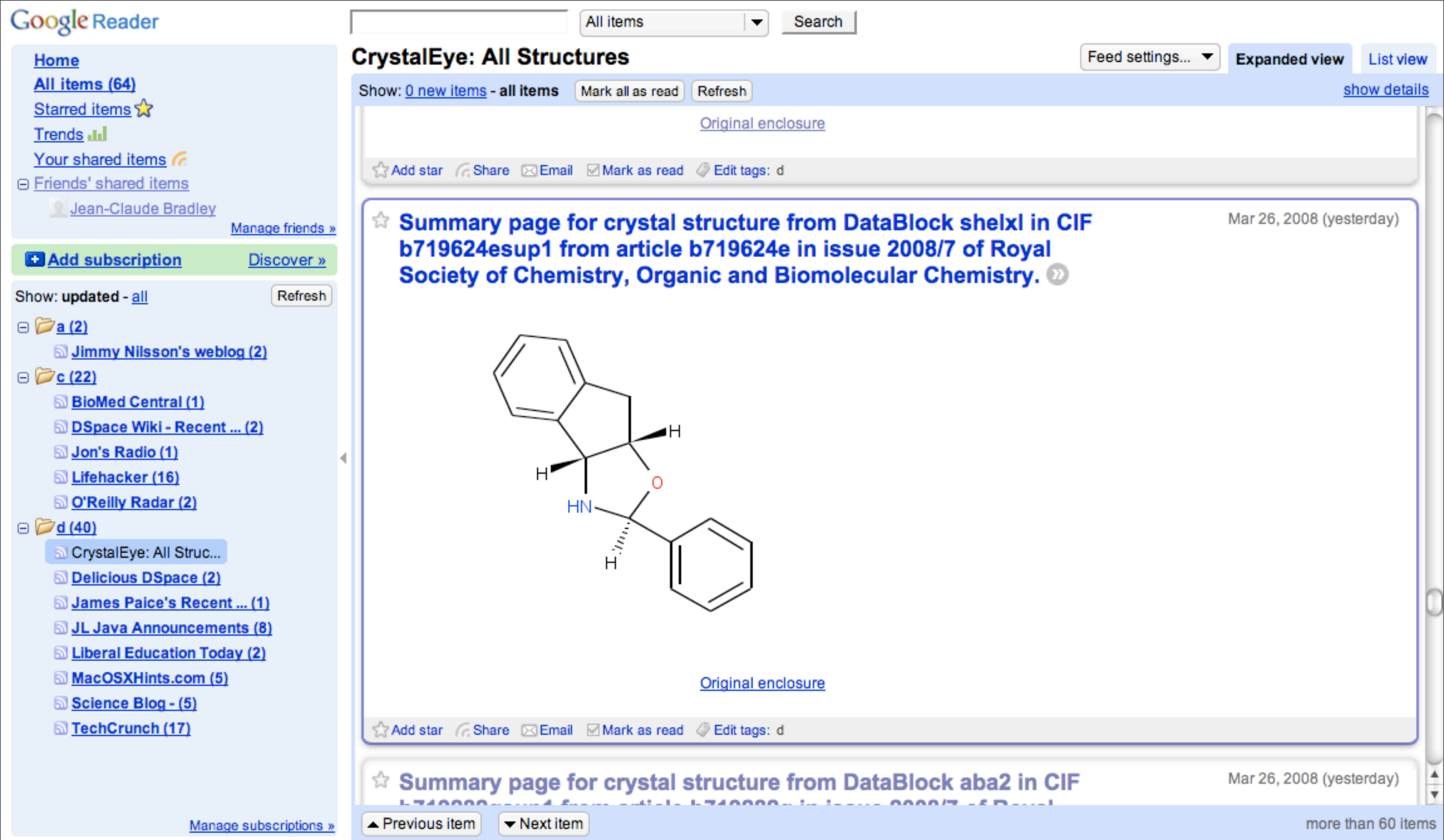
Search

The search tool was the first dynamic part of the application, and runs as a simple set of Java servlets.
N.B. that the search is chemistry specific – the first search method might be used for any collection of molecule data, the second is specific to crystallographic data.



Bond Length Histograms

Bond lengths histograms are a good example of the kind of processing and data checking that can only be done with a large collection of data.
Here we see the distribution of bond lengths between copper and carbon atoms in structures in CrystalEye. Nick generates these histograms for every type of bond.
Outliers: bad experimentation, also helped spot edge cases where computational codes failed to take account of some effect.



Feeds for all

Development of Atom feeds allowed embedded HTML and CML as enclosures

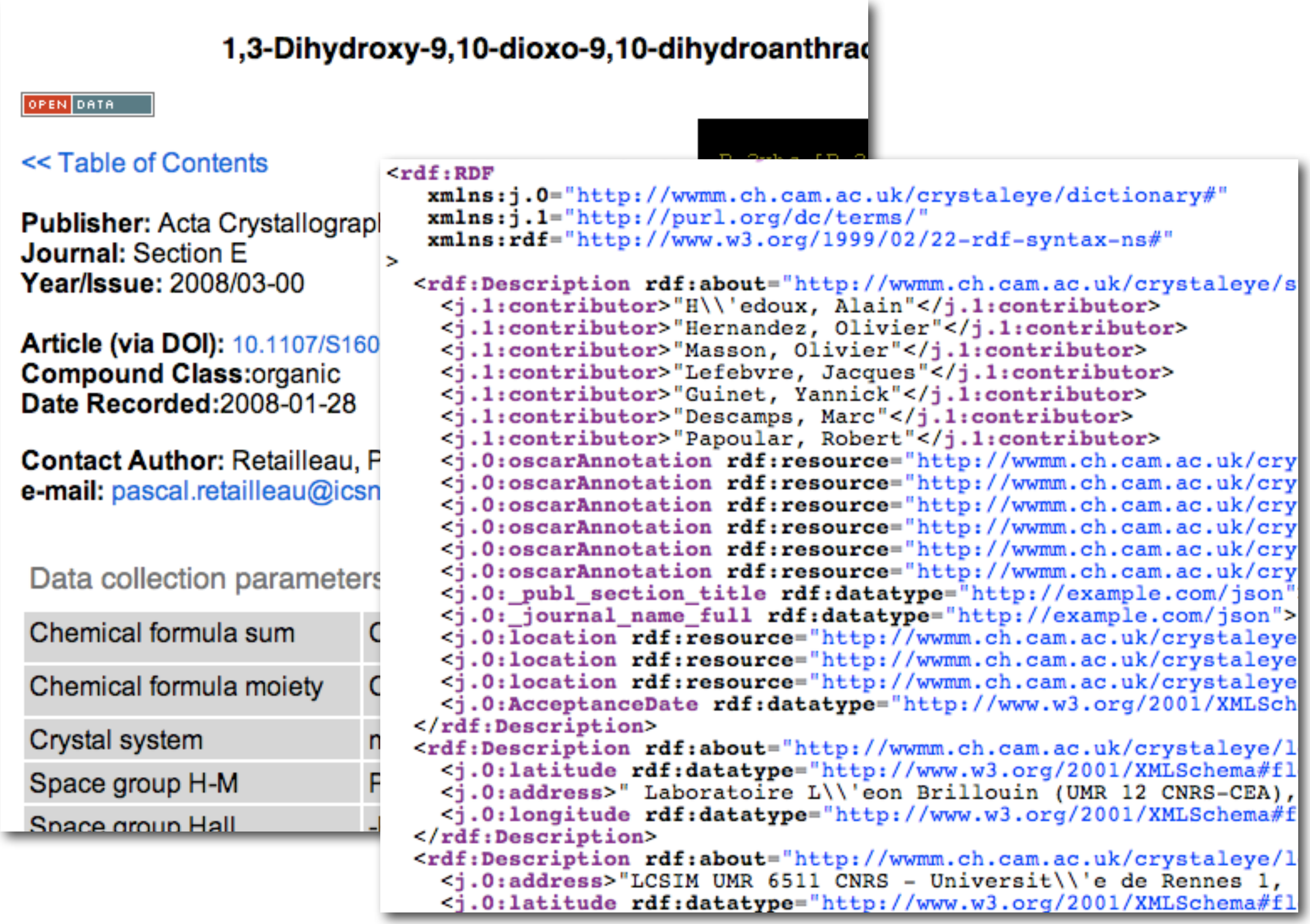
There were some problems with the existing CMLRSS feeds – including the CML inline was a big performance hit, and couldn't be done with standard XML libraries because the CML was too large. We started to use Atom, and link to the CML data using enclosures. Using Atom also allowed us to link to images in the entries, which meant the feeds looked good in common-or-garden feed readers.

Harvesting the Atom

RFC 5005
Feed paging and archiving

OAI-PMH was on the development road map, but users wanted to harvest the data using RSS feeds. RSS feeds aren't effective for CrystalEye harvesting, the data tends to arrive in big clump updates rather than a nice steady trickle, and there was no standard way for harvesters to discover or recover when they missed items. RFC 5005, which was published in its final form last September, extends Atom to fix these problems through a few special elements, and requirements on how the server must maintain and publish archive feed documents. We implemented RFC5005 in our main Atom feed, and published a simple harvester client application.

RDF metadata



1,3-Dihydroxy-9,10-dioxo-9,10-dihydroanthrac

OPEN DATA

<< Table of Contents

Publisher: Acta Crystallograp
Journal: Section E
Year/Issue: 2008/03-00

Article (via DOI): 10.1107/S160
Compound Class:organic
Date Recorded:2008-01-28

Contact Author: Retailleau, P
e-mail: pascal.retailleau@icsn

Data collection parameters

Chemical formula sum	C
Chemical formula moiety	C
Crystal system	m
Space group H-M	P
Space group Hall	-

```
<rdf:RDF
   xmlns:j.0="http://wwmm.ch.cam.ac.uk/crystaleye/dictionary#"
   xmlns:j.1="http://purl.org/dc/terms/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:about="http://wwmm.ch.cam.ac.uk/crystaleye/s
    <j.1:contributor>"H\\'edoux, Alain"</j.1:contributor>
    <j.1:contributor>"Hernandez, Olivier"</j.1:contributor>
    <j.1:contributor>"Masson, Olivier"</j.1:contributor>
    <j.1:contributor>"Lefebvre, Jacques"</j.1:contributor>
    <j.1:contributor>"Guinet, Yannick"</j.1:contributor>
    <j.1:contributor>"Descamps, Marc"</j.1:contributor>
    <j.1:contributor>"Papoular, Robert"</j.1:contributor>
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:oscarAnnotation rdf:resource="http://wwmm.ch.cam.ac.uk/cry
    <j.0:_publ_section_title rdf:datatype="http://example.com/json"
    <j.0:_journal_name_full rdf:datatype="http://example.com/json">
    <j.0:location rdf:resource="http://wwmm.ch.cam.ac.uk/crystaleye
    <j.0:location rdf:resource="http://wwmm.ch.cam.ac.uk/crystaleye
    <j.0:location rdf:resource="http://wwmm.ch.cam.ac.uk/crystaleye
    <j.0:AcceptanceDate rdf:datatype="http://www.w3.org/2001/XMLSch
  </rdf:Description>
  <rdf:Description rdf:about="http://wwmm.ch.cam.ac.uk/crystaleye/l
    <j.0:latitude rdf:datatype="http://www.w3.org/2001/XMLSchema#fl
    <j.0:address>" Laboratoire L\\'eon Brillouin (UMR 12 CNRS-CEA),
    <j.0:longitude rdf:datatype="http://www.w3.org/2001/XMLSchema#f
  </rdf:Description>
  <rdf:Description rdf:about="http://wwmm.ch.cam.ac.uk/crystaleye/l
    <j.0:address>"LCSIM UMR 6511 CNRS - Universit\\'e de Rennes 1,
    <j.0:latitude rdf:datatype="http://www.w3.org/2001/XMLSchema#fl
```

Using a link element in the head of the HTML pages, machine clients can obtain machine readable metadata about the structure in RDF format.

CrystalEye query for: Redfern S A T

Get this page as: Google Earth/Maps format

An in situ neutron diffraction study of cation disordering in synthetic qandilite Mg2TiO4 at high temperature Sample: First series, T = 1020 C (raw CML)

published in *American Mineralogist*

Authors:

- Short S M
- O'Neill H St C
- Redfern S A T
- Kesson S

Chemical named entities

- Mg2TiO4

Chemical ontology terms

- cation
- neutron
- An

An in situ neutron diffraction study of cation disordering in synthetic qandilite Mg2TiO4 at high temperature Sample: First series, T = 1054 C (raw CML)

published in *American Mineralogist*

Authors:

- Short S M
- O'Neill H St C
- Redfern S A T
- Kesson S

Chemical named entities

- Mg2TiO4

Chemical ontology terms

- cation

Mashing up RDF data

CrystalEye RDF data, concepts extracted, reindexed reindexed by by author, chemical name etc.

Andrew Walkingshaw took the CrystalEye RDF data, added some data of his own using the OSCAR3 chemistry natural language processing application, and mashed it back together to provide views and indices into the CrystalEye data we hadn't even thought of creating. Outside In-novation! It gets much, much sexier than this screenshot, but I can't show you too much as Andrew is presenting his work at XTech next month!

The global distribution of crystallography: papers in CrystalEye, 2000-2007

Dr Andrew Walkingshaw
Unilever Centre for Molecular Science Informatics

for Open Repositories 2008
28th March, 2008

http://www.lexical.org uk/
http://wwmm.ch.cam ac uk/blogs/walkingshaw/

Summary of Development

- Crystallographic data is collected daily from publisher websites, processed and enhanced.
- Web resources are constructed as part of data processing and published as static files by Apache httpd
- Browse & View
- Feeds for viewing and for harvesting
- Linked Machine Readable Metadata
- Data resources indexed for Search (implemented as Java servlets)
- Aggregate processing and results publication

The Future

Development on and around CrystalEye continues...

Planned work

- Refactoring!
  - Enhanced, archived atom feeds replacing CMLRSS throughout
  - Decouple the spider, processing and publishing parts
  - More auto-discovery
- Departmental crystallography repository, using CrystalEye and the JISC SPECTRa tools.
- New features
  - Dumpfile download (through S3)
  - SWORD support
  - ORE support (Microsoft ORE project)
  - OAI-PMH support (for e-Crystals federation)

Nearly there!

Don’t worry, we’re in the final furlong, and I’m about to start talking about repositories...

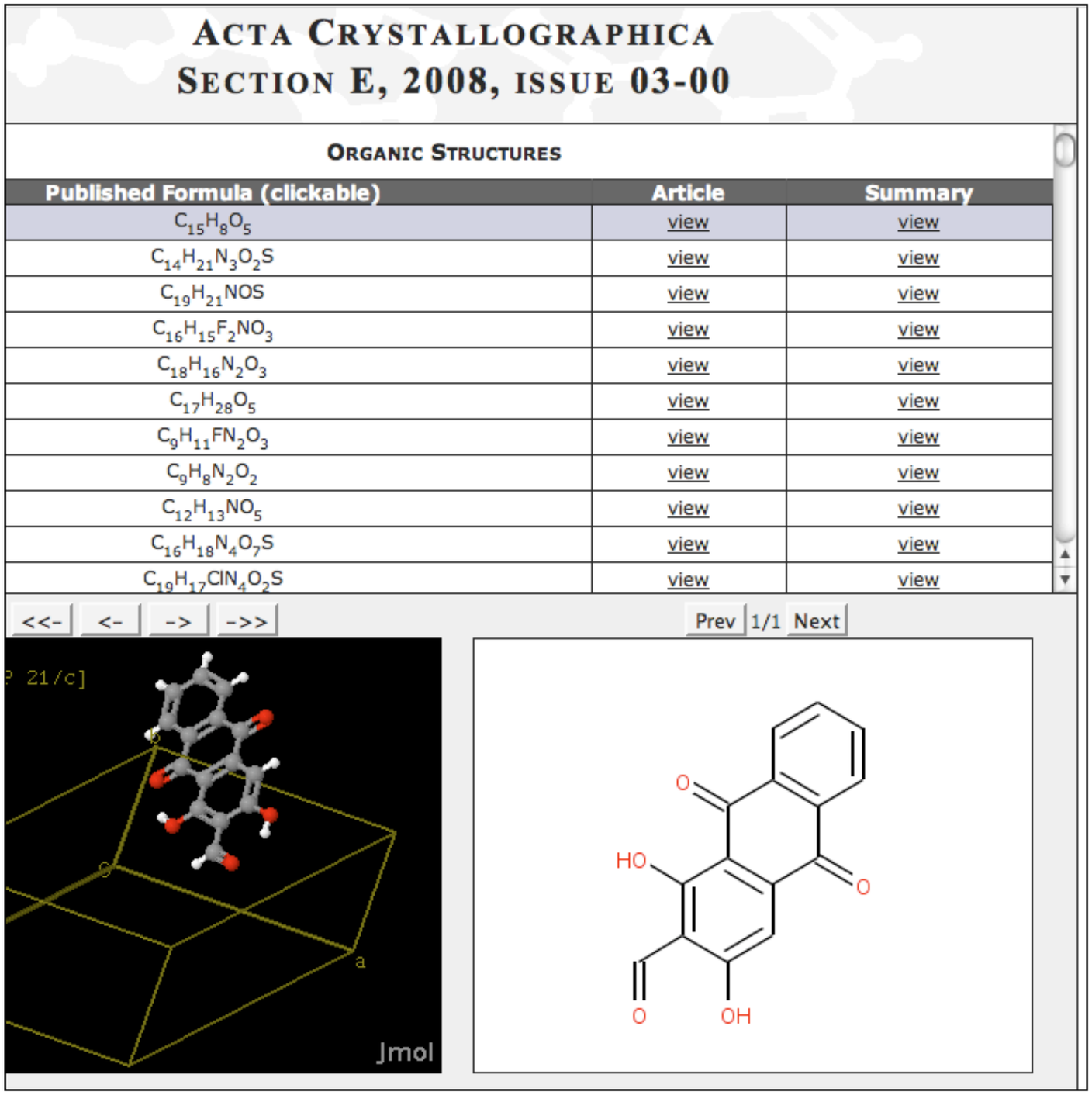
That’s the end of my narrative of the CrystalEye story so far, so I want to move on to talk about the meta-story, which is how CrystalEye is relevant to repositories...

Is CrystalEye a Repository?

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

Is CrystalEye a Repository?



Browse

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

Is CrystalEye a Repository?

1,3-Dihydroxy-9,10-dioxo-9,10-dihydroanthracene-2-carbaldehyd

OPEN DATA

<< Table of Contents

Publisher: Acta Crystallographica
Journal: Section E
Year/Issue: 2008/03-00

Article (via DOI): 10.1107/S1600536808004169
Compound Class:organic
Date Recorded:2008-01-28

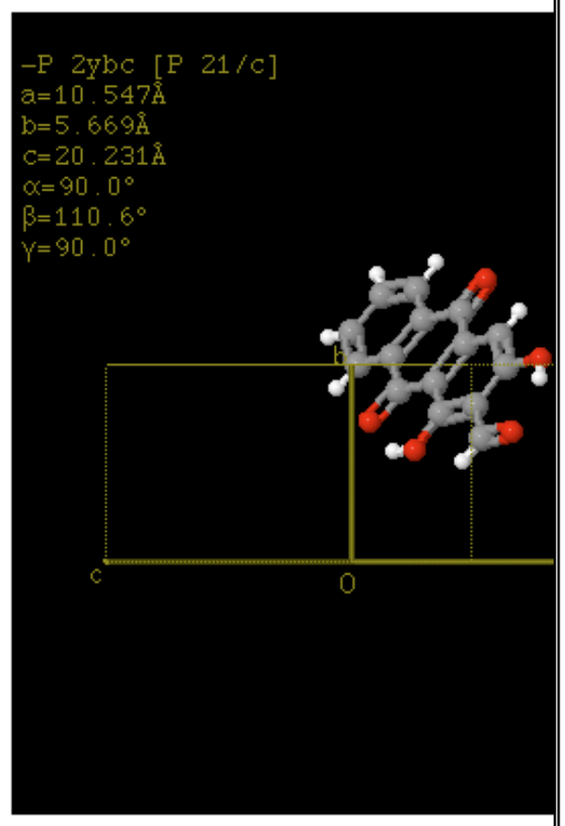
Contact Author: Retailleau, Pascal
e-mail: pascal.retailleau@icsn.cnrs-gif.fr

Data collection parameters

Chemical formula sum	$C_{15}H_8O_5$
Chemical formula moiety	$C_{15}H_8O_5$
Crystal system	monoclinic
Space group H-M	P 21/c
Space group Hall	-P 2ybc
Data collection temperature	293.0

Refinement results

R Factor (Obs)	0.051
R Factor (All)	0.081
Weighted R Factor (Obs)	0.132
Weighted R Factor (All)	0.15



Show no. of unit cells along axis:
a: 1
b: 1
c: 1
enter

Enter Jmol script:
load./bg2162supl_I.complete.cml.x

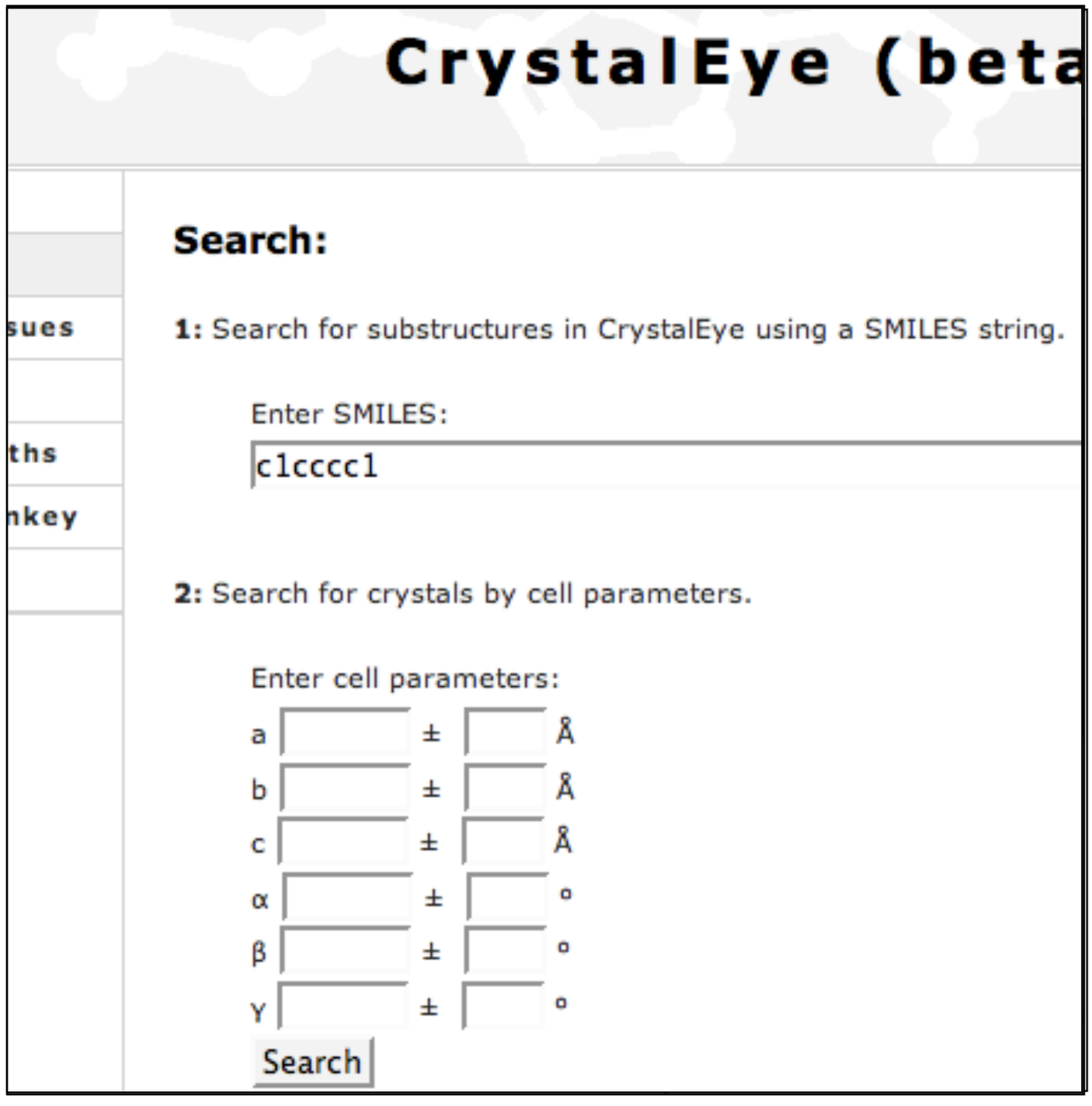
Browse

View

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

Is CrystalEye a Repository?



Browse

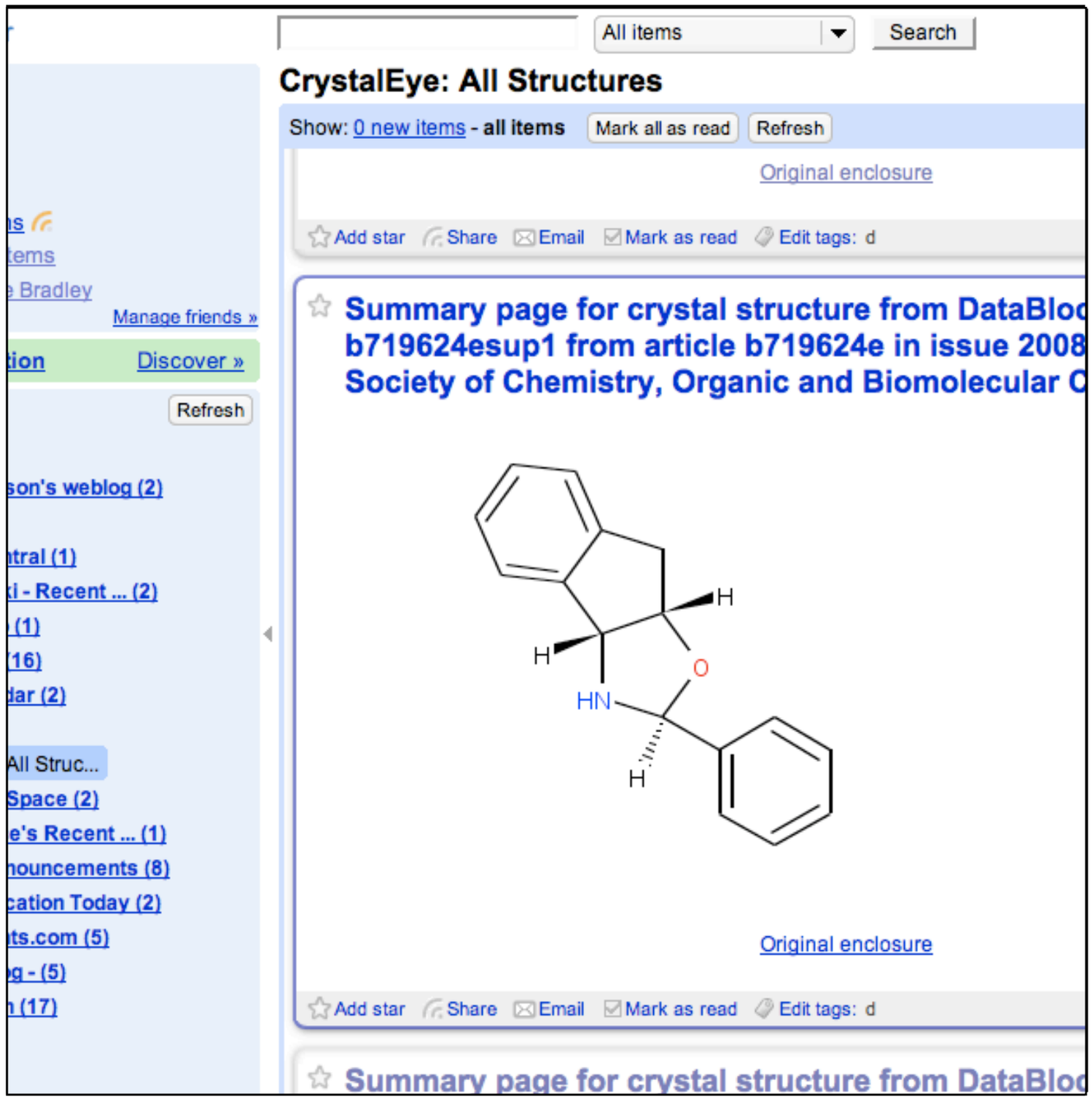
View

Search

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

Is CrystalEye a Repository?



Browse

View

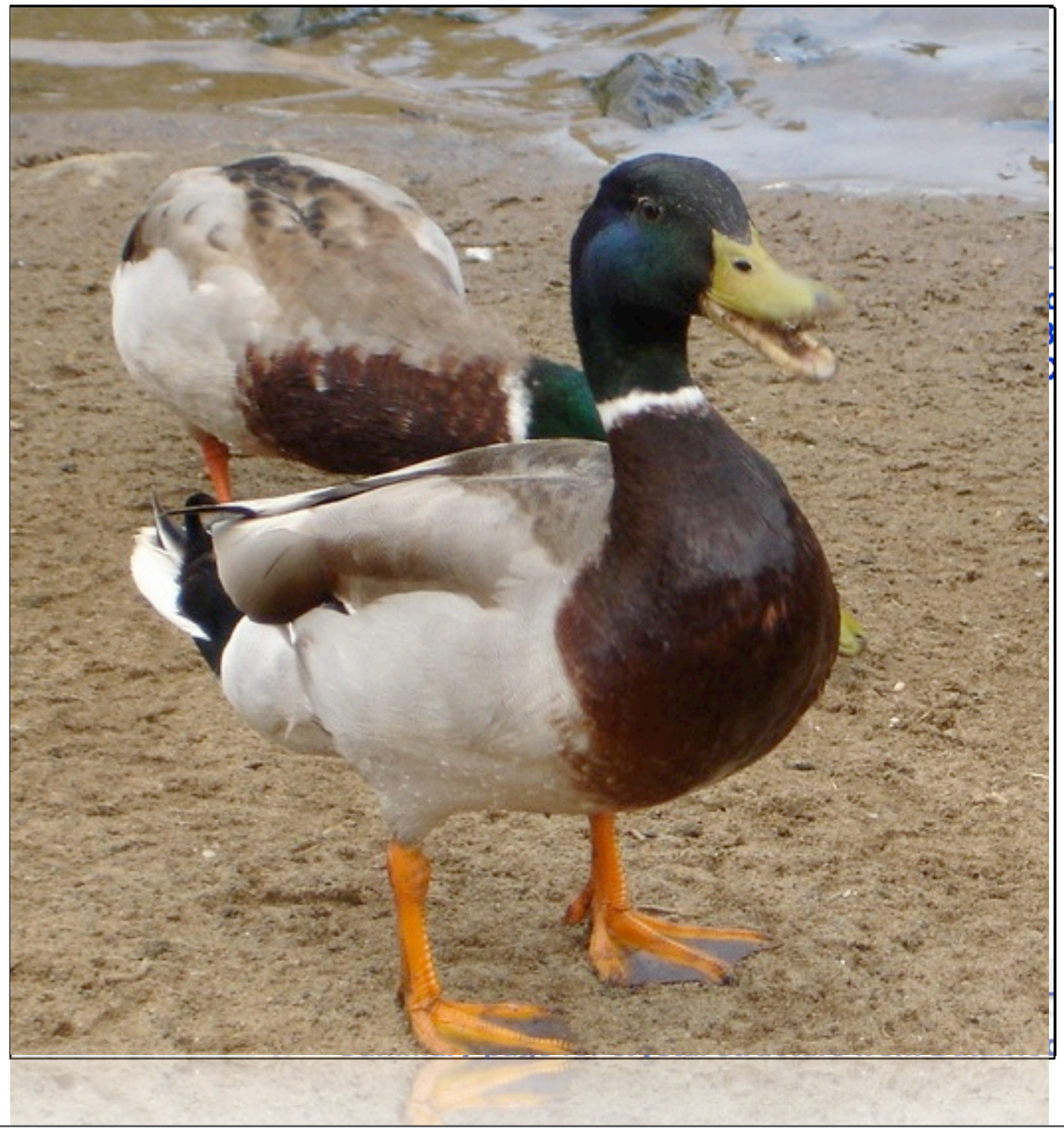
Search

Harvest

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

Is CrystalEye a Repository?



Browse

View

Search

Harvest

It walks like one, quacks like one, swims like one and goes well with orange sauce like one...

CrystalEye is a “just” a web site, but has many of the features you’d expect from a repository.

CrystalEye and the Subject Repository

e-Crystals Federation

Simon Coles has presented on e-Crystals already this morning. Over the course of e-Crystals we'll be working out how CrystalEye can work as part of a federated pan-institutional subject repository.

CrystalEye and the Institutional Repository

How could / should the University of Cambridge institutional repository interact with CrystalEye?

CrystalEye is a data repository, or at the very least an amply-featured data collection.

The way it's been most usually done in the past has been to move or copy the data over to a centralized repository. We don't believe this would work for most long-tail science applications; there's too much domain expertise involved in curating the data, and it would make managing the data more difficult.

Clifford Lynch's original definition of the Institutional Repository:

IR Services: Some Ideas

Technological

- Additional backup
- Preservation
- URL redirection and management
- Promote access through portals
- Equipment

Organisational

- Create data management roles in departments
- Data management training for academics
- “Transferrable skills” training for PhD students?

These are just ideas from our perspective.

I’d like to emphasize how useful data training could be - Nick lost plenty of time due to data management mistakes, particular around identifiers and discarding intermediate data.

Conclusions

Short version:
* Keep it simple
* Share your data to let the community drive functionality
* Focus on the web

Long version:
So why did CrystalEye teach me to love the web?
Because we never had to stop and think “now we have to stop and get this data into a

Copyright and Notices

Flickr images:

PMR: borazivkovic (BY)

Crystal: wabberjocky (BY-NC-SA)

Spider: dincordera (BY-NC-SA)

Car: snellgrove (BY-NC-SA)

Ascent of man : Kaptain Kobold (BY-NC-SA)

Open/Close: mag3737 (BY-NC-SA)

Spices: estherase (BY-NC-SA)

Combine harvester: tricky ™ (BY-NC-SA)

Sunset: nexus6 (BY-NC-SA)

Horse race: Chris Breeze (BY)

Duck: Frogzone 1 (BY-NC)

Web/Pylon: Orcoo (BY-NC-SA)

Others:

Atom icon: Mozilla Foundation (MPL)

Credit also due to the excellent www.compfight.com for making collecting these images so much easier.

Credits and Links

- Thanks to Nick Day and Peter Murray-Rust for all their work on CrystalEye
- Thanks to Andrew Walkingshaw for his work on the CrystalEye RDF and to Talis for the use of their Platform Store
- CrystalEye: http://wwmm.ch.cam.ac.uk/crystaleye/
- Coverage on some of the features mentioned here on my blog: http://wwmm.ch.cam.ac.uk/blogs/downing/
- Coverage on CrystalEye and much, much more on PM-R's blog: http://wwmm.ch.cam.ac.uk/blogs/murrayrust/