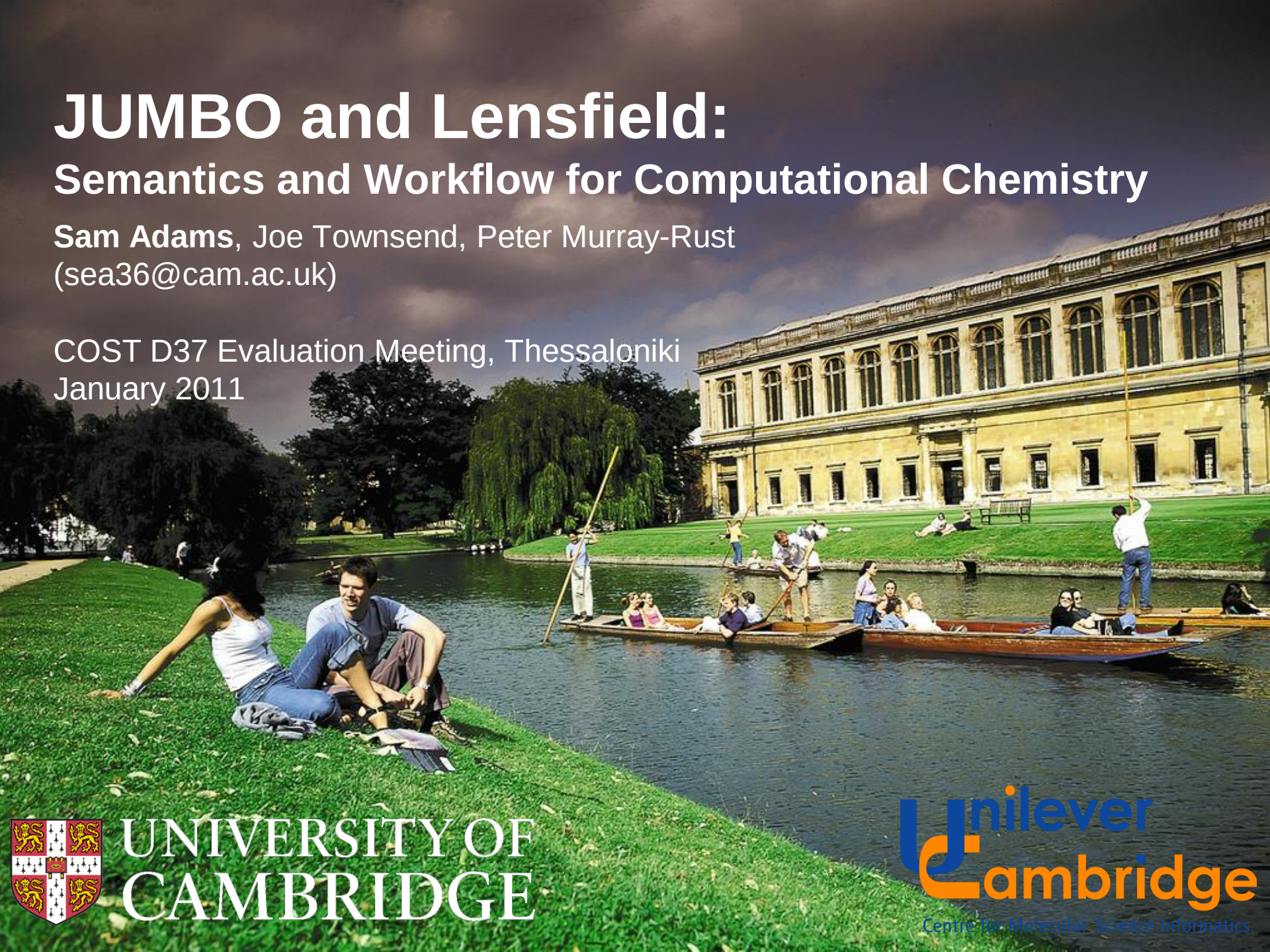
JUMBO and Lensfield:

Semantics and Workflow for Computational Chemistry

Sam Adams, Joe Townsend, Peter Murray-Rust
(sea36@cam.ac.uk)

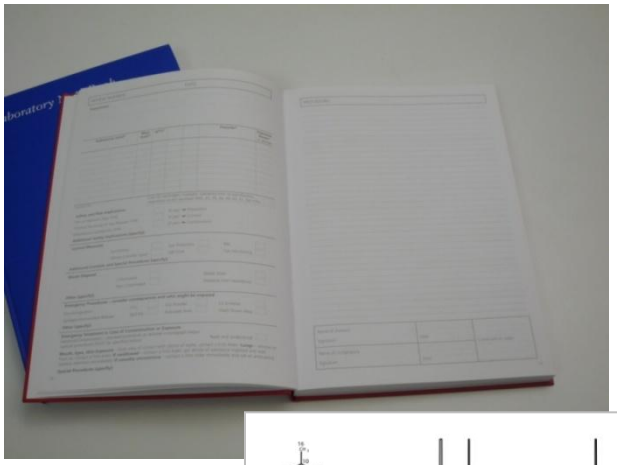
COST D37 Evaluation Meeting, Thessaloniki
January 2011

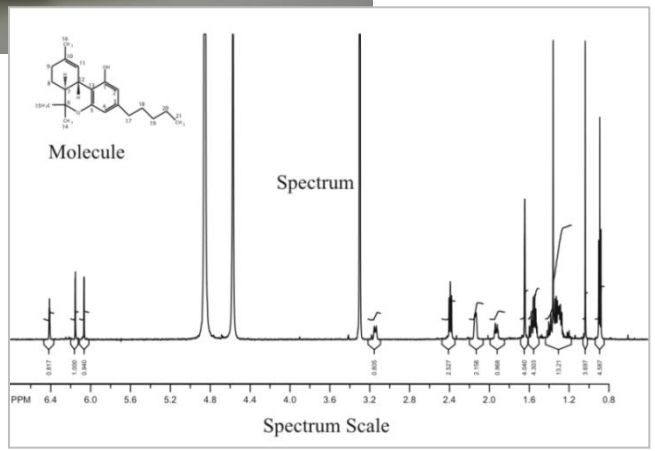


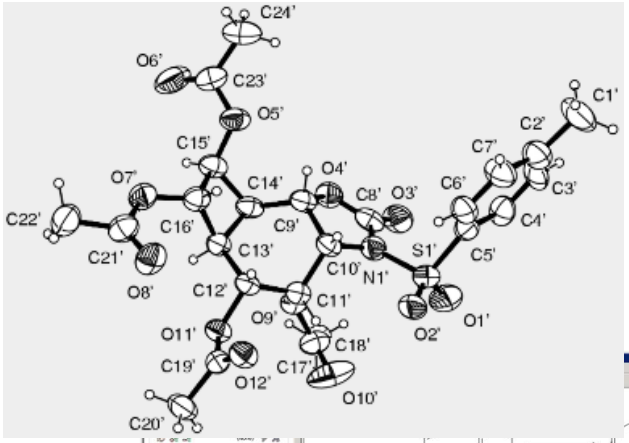
UNIVERSITY OF CAMBRIDGE

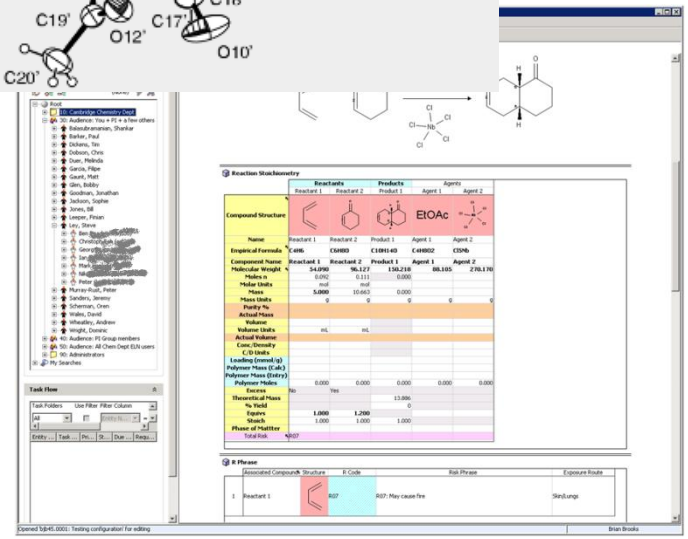
Unilever Cambridge
Centre for Molecular Science Informatics

Most data is never published

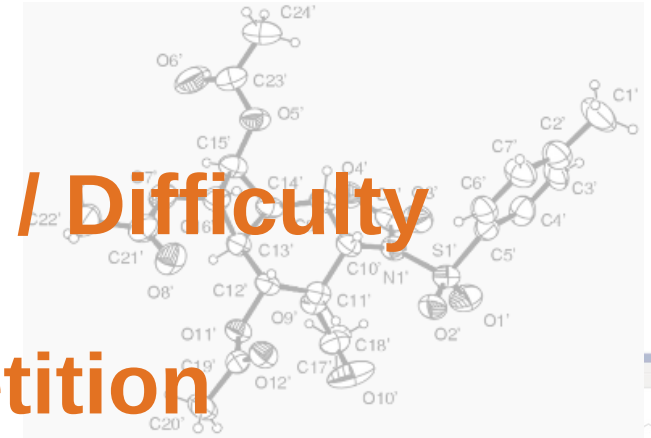








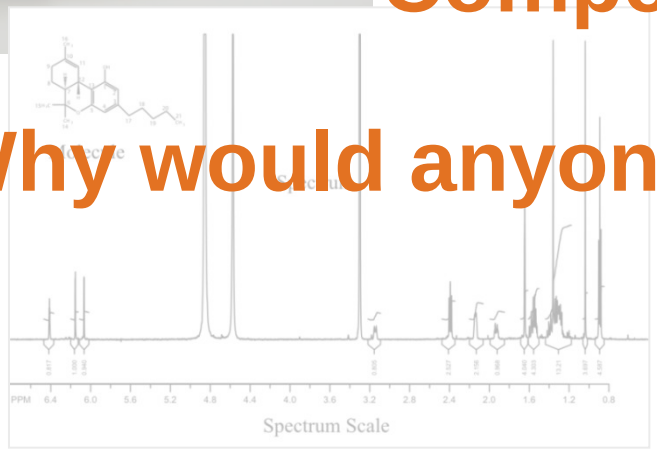
Most data is never published

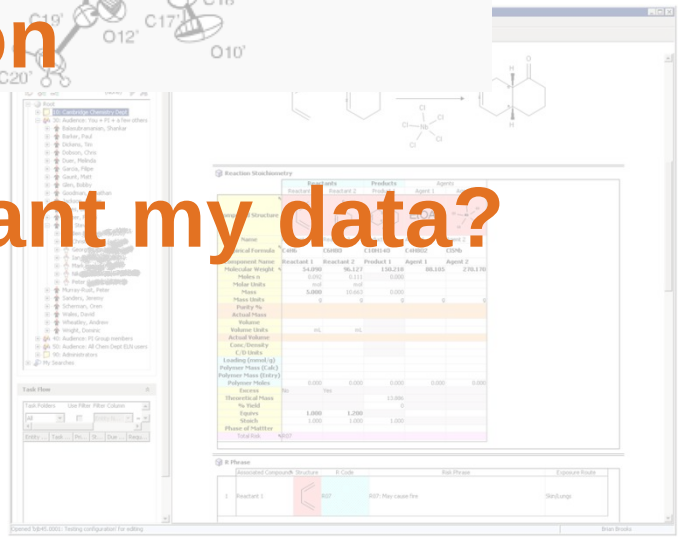


Time / Effort / Difficulty

Competition

Why would anyone want my data?





Most published data is unusable

π-Allyltricarbonyliron lactone complexes: versatile tools for asymmetric synthesis

A thesis presented by

Jürgen Harter

In partial fulfilment of the requirements for the award of the degree of

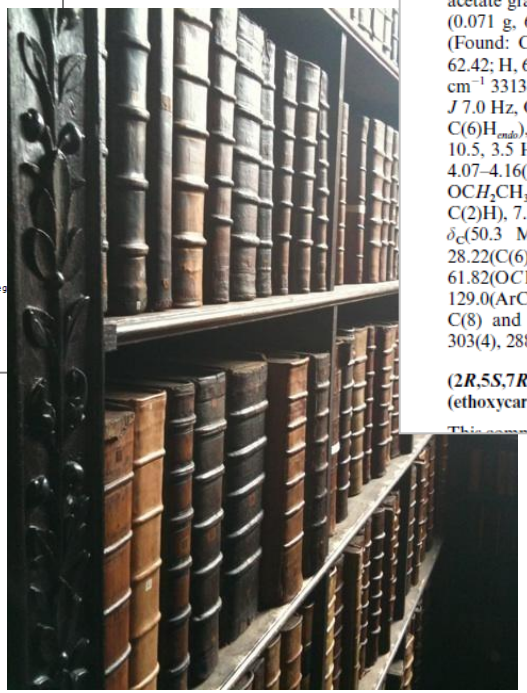
DOCTOR OF PHILOSOPHY

OF THE

UNIVERSITY OF CAMBRIDGE

B.P. Whiffen Laboratory,
Department of Chemistry,
University of Cambridge,
Lensfield Road,
Cambridge, CB2 1EW

Wolfson Colleg
August 2002



288(100), 242(14), 231(19), 204(94).

(2*R*,5*S*,7*R*,1′*R*)-1-Aza-3-oxa-8-oxo-2-phenyl-7-[*N*-acetylamino-(ethoxycarbonyl)methyl]bicyclo[3.3.0]octane 12b

At −5 °C, acetic anhydride (0.042 g, 0.41 mmol) was added to a solution of amine **12a** (0.10 g, 0.33 mmol) and triethylamine (0.067 g, 0.66 mmol) in chloroform (9 ml). The mixture was stirred at −5 °C for 10 minutes and then at 0 °C for a further 4 hours. Following washing with citric acid solution (10% in H_2O; 3 × 8 ml) and drying over magnesium sulfate, the solvent was evaporated. The resulting yellow oil was purified by flash column chromatography on silica (1 : 1 petroleum ether–ethyl acetate gradient to 1 : 3) to give the product, a pale yellow oil (0.071 g, 62%): R_f 0.14 (1 : 3 petroleum ether–ethyl acetate); (Found: C, 62.26; H, 6.84; N, 7.68. $C_{18}H_{22}N_2O_5$ requires C, 62.42; H, 6.40; N, 8.09%); $[\alpha]_D^{25}$ +120 (*c* 0.20, $CHCl_3$); ν_{max} (film)/cm^{-1} 3313, 1739, 1703, 1690; δ_H(500 MHz, $CDCl_3$) 1.28(3H, t, *J* 7.0 Hz, OCH_2CH_3), 2.01(3H, s, $CH_3C(O)$), 2.14–2.20(1H, m, C(6)H_{endo}), 2.54–2.60(1H, m, C(6)H_{exo}), 3.27(1H, ddd, *J* 10.5, 10.5, 3.5 Hz, C(7)H_{exo}), 3.66(1H, dd, *J* 8.0, 8.0 Hz, C(4)H_{endo}), 4.07–4.16(1H, m, C(5)H), 4.17–4.27(3H, m, C(4)H_{exo} and OCH_2CH_3), 4.84(1H, dd, *J* 8.5, 3.5 Hz, C(1′)H), 6.23(1H, s, C(2)H), 7.11(1H, br d, *J* 8.5 Hz, NH), 7.29–7.42(5H, m, ArH); δ_C(50.3 MHz, $CDCl_3$) 13.98(OCH_2CH_3), 22.94($H_3CC(O)$), 28.22(C(6)), 48.19 and 51.51(C(7) and C(1′)), 57.07(C(5)), 61.82(OCH_2CH_3), 72.26(C(4)), 86.89(C(2)), 126.2, 128.7, 129.0(ArC), 138.6(4° ArC), 169.9, 170.7 and 176.9($CH_3C(O)N$, C(8) and CO_2Et); *m/e* (probe CI, NH_3) 347(MH^+, 100%), 303(4), 288(4), 273(7), 231(14), 211(8), 202(26).

(2*R*,5*S*,7*R*,1′*S*)-1-Aza-3-oxa-8-oxo-2-phenyl-7-[*N*-acetylamino-(ethoxycarbonyl)methyl]bicyclo[3.3.0]octane 13b

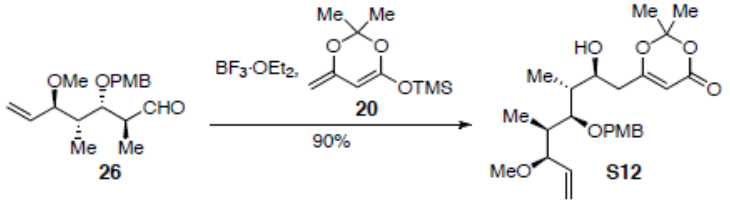
acetate) to give the product **14** as a colourless oil (40 mg, 61% over 2 steps): R_f 0.12 (1 : 6 petrol–ethyl acetate); ν_{max} (thin film)/cm^{-1} 2924(br m), 1737(s), 1700(s), 1667(s); δ_H(200 MHz, $CDCl_3$) 1.14(3H, t, *J* 7.0 Hz, OCH_2CH_3), 2.03(3H, s, $CH_3C(O)$), 2.12–2.24(1H, m, C(6)H_{endo}), 2.38–2.51(1H, m, C(6)H_{exo}), 3.01–3.10(1H, m, C(7)H), 3.42(1H, dd, *J* 8.5, 8.5 Hz, C(4)H_{endo}), 4.00–4.28(4H, m, C(5)H, C(4)H_{exo} and OCH_2CH_3), 4.90(1H, dd, *J* 5.0, 8.5 Hz, C(1′)H), 6.28(1H, s, C(2)H), 6.81(1H, br d, *J* 8.5 Hz, NH), 7.37–7.39(5H, m, ArH); δ_C(50.3 MHz, $CDCl_3$) 13.85(OCH_2CH_3), 23.04($H_3CC(O)$), 25.48(C(6)), 47.73 and 53.08(C(7) and C(1′)), 57.37(C(5)), 61.93(OCH_2CH_3), 71.49(C(4)), 86.90(C(2)), 125.7, 128.4 and 128.6(ArC), 138.4(4° C), 169.9(2 × CO), 176.3(CO); *m/e*($APCI^+$) 347(MH^+, 100%), HRMS(CI^+) 347.1607, MH^+ requires 347.1606.

(2*S*,4*S*)-*N*-Benzyl-2-methoxycarbonyl-4-[*N*-acetylamino-(ethoxycarbonyl)methyl]-5-oxopyrrolidine 15

Lactam **14** (50 mg, 0.14 mmol) was hydrogenated to yield the crude alcohol product (40 mg): ν_{max} (film)/cm^{-1} 3286(br m, OH, NH), 1738(s, ester CO), 1672(s, lactam CO); *m/e* ($APCI^+$) 349 (MH^+, 100%). This was immediately oxidized according to the Sharpless protocol[69] to give a white solid (12 mg) LRMS ($APCI^+$) *m/e* 363 (MH^+, 100%), which was in turn immediately treated with diazomethane in ether. The solvent was removed *in vacuo* to give a pale yellow oil which was purified by flash column chromatography on silica (ethyl acetate). The product was obtained as a mixture of C-1′ diastereomers in a ratio of 1 : 2 (12 mg, 23% over 3 steps): R_f 0.31, 0.24 (EtOAc); ν_{max} (film)/cm^{-1} 3320(br m), 1742(s), 1695(s); δ_H(500 MHz, $CDCl_3$) (major diastereomer) 1.21(3H, t, *J* 7.0 Hz, OCH_2CH_3), 2.06(3H, s, $CH_3C(O)$), 2.27–2.33(1H, m, C(3)H), 2.46–2.51(1H, m, C(3)H), 2.98–3.03(1H, m, C(4)H), 3.68(3H, s, OCH_3), 3.98–

Virtually unreadable
Totally undiscoverable

Supporting information can require massive effort



(+)-6-{[2*S*-(2*R,3*S**,4*S**,5*S**,6*S**)]-2-Hydroxy-4-[(4-methoxyphenyl)methoxy]-6-methoxy-3,5-dimethyloct-7-enyl}-2,2-dimethyl-4*H*-1,3-dioxin-4-one (S12)** To aldehyde **26** (12 mg, 0.039 mmol) and ketene acetal **20** (84 mg, 0.39 mmol) in DCM (2.0 mL) was added $BF_3 \cdot OEt_2$ (10 µL, 0.078 mmol) at -78 °C. The mixture was stirred 45 min at this temperature before being warmed to rt and quenched with aqueous $NaHCO_3$. The resulting mixture was diluted with H_2O, extracted with DCM, dried over Na_2SO_4, and concentrated. Flash chromatography (hexanes:EtOAc = 7:3 to 1:1) gave **S12** (13 mg, 80%).

^{1}H NMR (500 MHz, $CDCl_3$) δ 7.28 (d, *J* = 8.5 Hz, 2H), 6.95 (d, *J* = 8.5 Hz, 2H), 5.57 (ddd, *J* = 8.4, 10.2, and 17.0 Hz, 1H), 5.33 (dd, *J* = 1.8 and 10.1 Hz, 1H), 5.27 (s, 1H), 5.22 (dd, *J* = 1.6 and 17.1 Hz, 1H), 4.61 (d, *J* = 11.0 Hz, 1H), 4.47 (d, *J* = 11.0 Hz, 1H), 4.19 (m, 1H), 3.84 (dd, *J* = 2.9 and 11.0 Hz, 1H), 3.81 (s, 3H), 3.39 (dd, *J* = 8.4 and 8.4 Hz, 1H), 3.27 (s, 3H), 2.4 (d, *J* = 4.4 Hz, 1H), 2.39 (dd, *J* = 9.0 and 14.3 Hz, 1H), 2.23 (dd, 4.4 and 14.3 Hz, 1H), 1.80 (ddq, *J* = 2.9, 8.4, and 7.1 Hz, 1H), 1.70 (ddq, 1.8, 11.0, and 7.1 Hz, 1H), 1.65 (s, 6H), 0.89 (d, *J* = 7.1 Hz, 3H), and 0.88 (d, *J* = 7.0 Hz, 3H).

^{13}C NMR (125 MHz, $CDCl_3$) δ 169.6, 161.0, 159.2, 137.1, 130.4, 129.3, 119.2, 113.8, 106.3, 94.6, 85.0, 80.2, 74.2, 68.5, 55.5, 55.2, 40.5, 40.2, 39.2, 25.1, 24.6, 10.6, and 10.2.

IR (neat) 3470, 2974, 2934, 1728, 1634, 1514, and 1249 cm^{-1}.

HRMS (FAB) Calcd for $(C_{25}H_{36}O_7 + Na)^+$: 471.2353. Found: 471.2359.

TLC R_f = 0.3, hexanes:EtOAc = 1:1.

$[\alpha]^{RT}$ +5.59° (c = 1.18, DCM).

Some disciplines are better than others...

Authors are required to provide crystallographic data in the crystallographic information file (CIF) format *at the time of manuscript submission*. Details on the preparation, validation, and submission of this material are available from the Journal's Web site

RCSB PDB
PROTEIN DATA BANK

... but not CompChem

Supplementary Material

Unexpected dual orbital effects in radical addition reactions involving acyl, silyl and related radicals

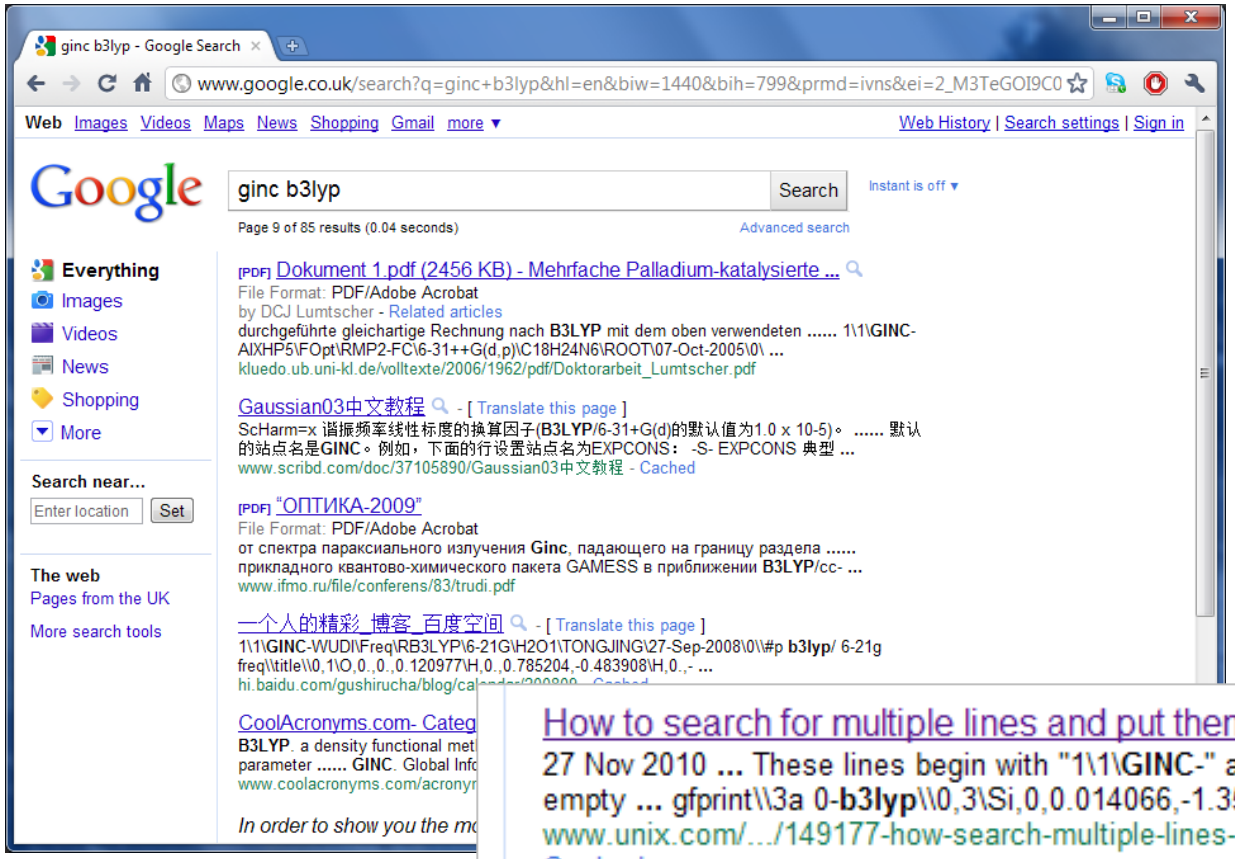
Carl H Schiesser,*[a,b] Hiroshi Matsubara,*[c] Ina Ritsner[a] and Uta Wille*[a,b]

MP2/6-311G**

1\1\ CHEMISTRY CLUSTER KIRKLAND-KNET5\FTS\UMP2-FC\6-311G(d,p)\C3H6N1O1(2)\HIROSHI\16-May-2005\1\\#MP2/6-311G** SCF=DIRECT OPT=(TS,EF,CALCHFFC,MAXCYCLE=100) NOSYMM FREQ=NORAMAN\\TS for addition to nitorgen of imine\\0,2\C\O,1,r2\C,1,r3,2,a3\N,1,r4,2,a4,3,d4,0\C,4,r5,1,a5,2,d5,0\H,3,r6,1,a6,2,d6,0\H,3,r7,1,a7,2,d7,0\H,3,r8,1,a8,2,d8,0\H,4,r9,1,a9,2,d9,0\H,5,r10,4,a10,1,d10,0\H,5,r11,4,a11,1,d11,0\\r2=1.2223954\r3=1.51801246\a3=124.65042078\r4=1.76588121\a4=109.24841904\d4=124.33673948\r5=1.25097908\a5=116.47411713\d5=7.63510449\r6=1.0927799\a6=111.71725127\d6=173.79797951\r7=1.09886014\a7=110.64005704\d7=51.47921365\r8=1.09269408\a8=107.9458183\d8=293.3479316\r9=1.0179727\a9=124.02108618\d9=173.41276736\r10=1.0881745\a10=123.70838963\d10=167.50178173\r11=1.09199551\a11=115.54893759\d11=-10.92919791\\Version=x86-Linux-G03RevB.04\HF=-246.3672269\MP2=-247.1570942\PUHF=-246.3750591\PMP2-0=-247.1635975\S2=0.831467\S2-1=0.803382\S2A=0.752174\RMSD=3.031e-09\RMSF=9.917e-05\Dipole=-0.8269389,-1.3467298,-1.0521902\PG=C01 [X(C3H6N1O1)]\\@

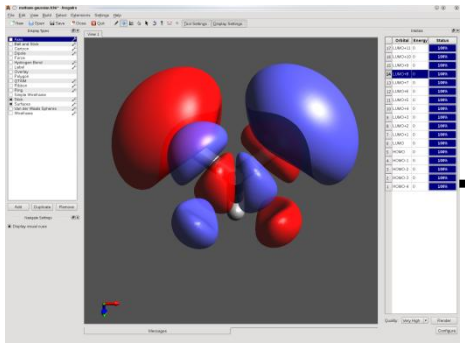
PDF destroys data

Virtually no data published at all



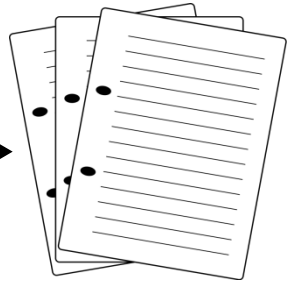
The Problem

- No standard way to archive or search the data from CompChem calculations; valuable data festers on disk.
- There isn't even a standard data format (despite the data being rigorously defined) so each computational chemistry code needs specialised tools to understand its output.

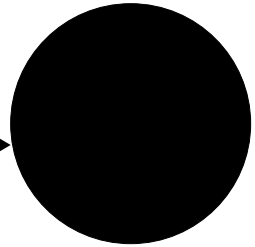


Generate data from scratch

Expensive computation



Cumbersome data format



Black hole

The Quixote Project

- **An Open Source, Open Data international collaboration, based entirely in the internet started following a CECAM meeting in Zaragoza:**
  http://neptuno.unizar.es/events/qcdatabases2010/
- **Three key ideas:**
  - scientific data (and ideally codes too) should be "open"
  - a standard data model/format is a Very Good Thing
  - universally accessible and open databases of the results of calculations are scientifically highly valuable.
- **Aim:**
  - create a useful infrastructure and consolidate the model around the tools

The Quixote Project

Countless CPU-hours are spent calculating data most of which is "archived" (discarded?) ... never to be found again

Data is the bread-and-butter of computational chemistry, yet most of it is never seen by the community

- Lots of people have tried to solve the problem of QC interoperability, but no one can agree on a solution.
- Quixote is… an internet-based, international community of scientists, passionate about open-source and open-data, looking to solve the problem in a bottom-up, pragmatic way.

...

Build it, and they will come.

The Quixote Project

- **Standard Data Model:**
  - different codes can interoperate to create complex workflows.
  - tools (e.g. GUI's) can operate on the input and output of any code.
- **Open results databases:**
  - codes can be easily validated and benchmarked
  - are essential for the development of new methods
  - avoid costly duplication of results
  - provide a valuable resource for data mining
  - a straightforward way of archiving / publishing

The Quixote Project

- Mailing lists + Wiki
- Skype meetings, Collaborative authoring: EtherPad
- No funding...
  Just enthusiasm
- Engaging with existing Open Source projects
- Daresbury workshop late March 2011

Roger Martin	QuantumBio, Inc, PA
Tamás Beke-Somfai	Chalmers University, Göteborg
Sergio Maffioletti	Zurich
Jens Thomas	Daresbury
Peter Murray-Rust	Cambridge
Pablo Echenique	Zaragoza
Noel O'Boyle	University College Cork
Weerapong Phadungsukanan	Cambridge
Marcus Hanwell	Kitware, Inc, NY
Jorge Estrada	Zaragoza
Egon Willighagen	Karolinska Institutet
Sam Adams	Cambridge

Chemical Markup Language

eXtensible Markup Language (XML) vocabulary for Chemistry

One of the earliest XML dialects... over the past fifteen years has become the *de facto* XML standard for chemistry – accepted by publishers and the open source and commercial software communities. Supported by well over 1 million lines of Open Source code.

Example CML supporting tools:

- JUMBO / JUMBOConverters
- ChemDraw / ChemBioOffice
- Open Babel
- CDK
- Microsoft Word

```
<molecule xmlns="http://www.xml-cml.org/schema">
    <atomArray>
        <atom id="a1" elementType="C" formalCharge="0" />
        <atom id="a2" elementType="O" formalCharge="0" />
        <atom id="a3" elementType="O" formalCharge="0" />
    </atomArray>
    <bondArray>
        <bond id="b1" atomRefs2="a1 a2" order="2" />
        <bond id="b2" atomRefs2="a1 a3" order="2" />
    </bondArray>
</molecule>
```

Dictionaries and Conventions

- **Conventions:**
  - molecular
  - cmlcomp
  - crystallographic
- **Dictionaries:**
  - units
  - properties
  - data types

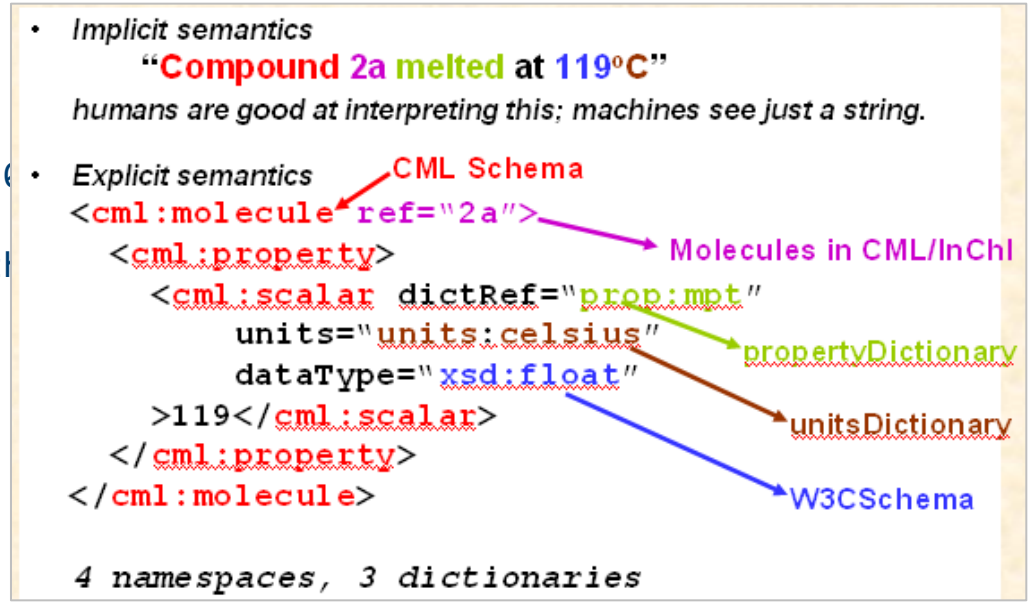
Documentation:
http://www.xml-cml.org/

Validator Service:
http://validator.xml-cml.org/

Impose data models and semantics on sections of CML documents

Dictionaries

```
<entry id="counterpoiseEnergy" cmlx:name="counterpoiseEnergy" cmlx:type="xsd:float"
  cmlx:definition="energy calculated by the Counterpoise method; differentiate from Counterpoise
keyword which takes an integer"
  cmlx:description="Counterpoise method resultant energy" cmlx:superclass="property">
  <h:p class="manual">
    See <h:a href="http://www.gaussian.com/g_tech/g_ur/k_counterpoise.htm">Gaussian09 online manual</h:a>
  </h:p>
  <h:p class="notes"><h:pre>
  Example:
   Counterpoise: corrected energy =
   Counterpoise: BSSE energy =
  </h:pre>
  Units are not specified, we guess th
  </h:p>
</entry>
```



JUMBO-Converters

CIF

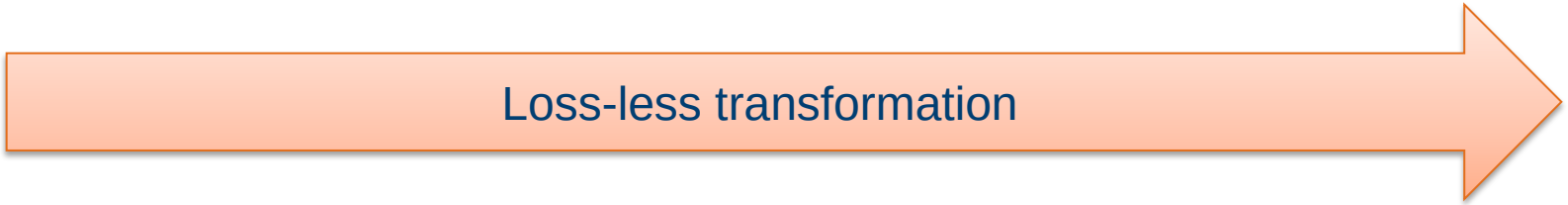
GAMESS-UK
GAMESS-US
GAUSSIAN
NWCHEM

Reflects structure of legacy input format

Properties and parameters are referenced against application specific dictionaries

Data re-factored to fit standard data model

Where possible properties and parameters referenced against standard dictionary



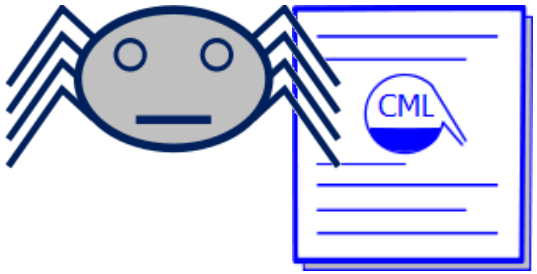
CrystalEye

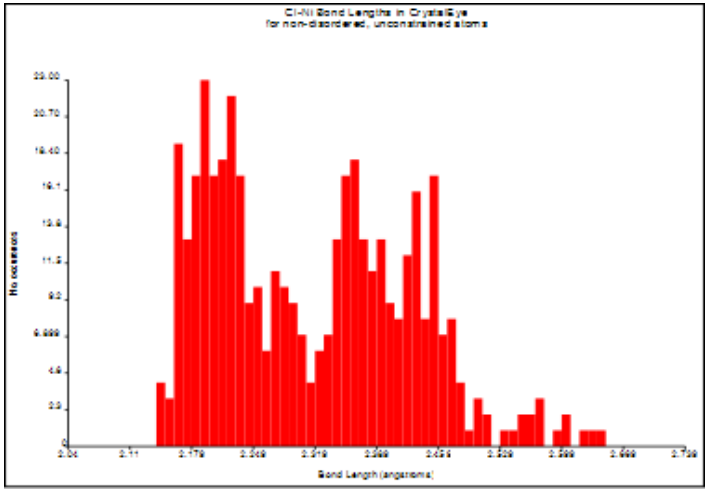
ACS

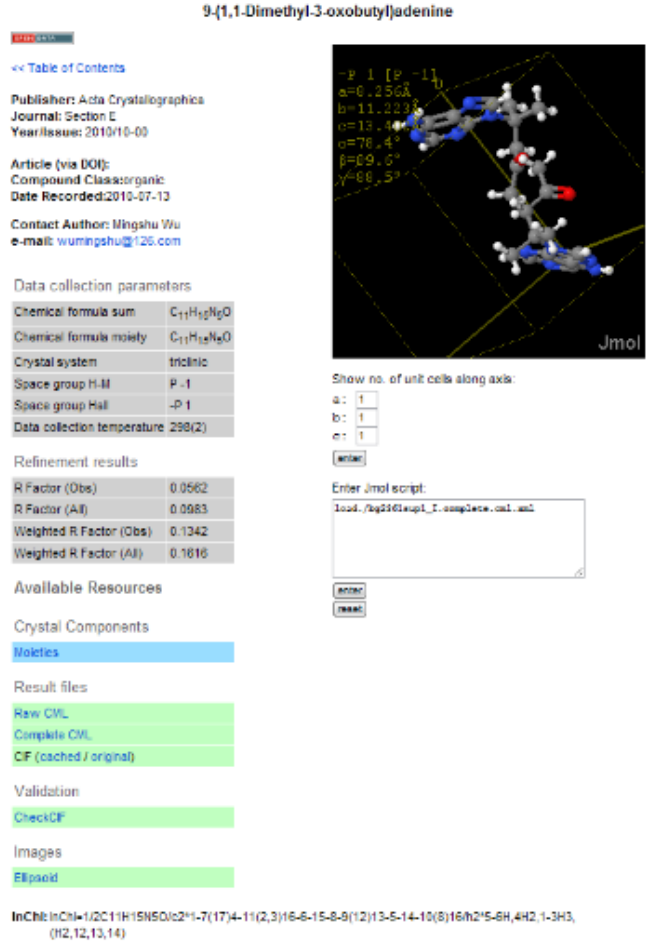
Nature

RSC

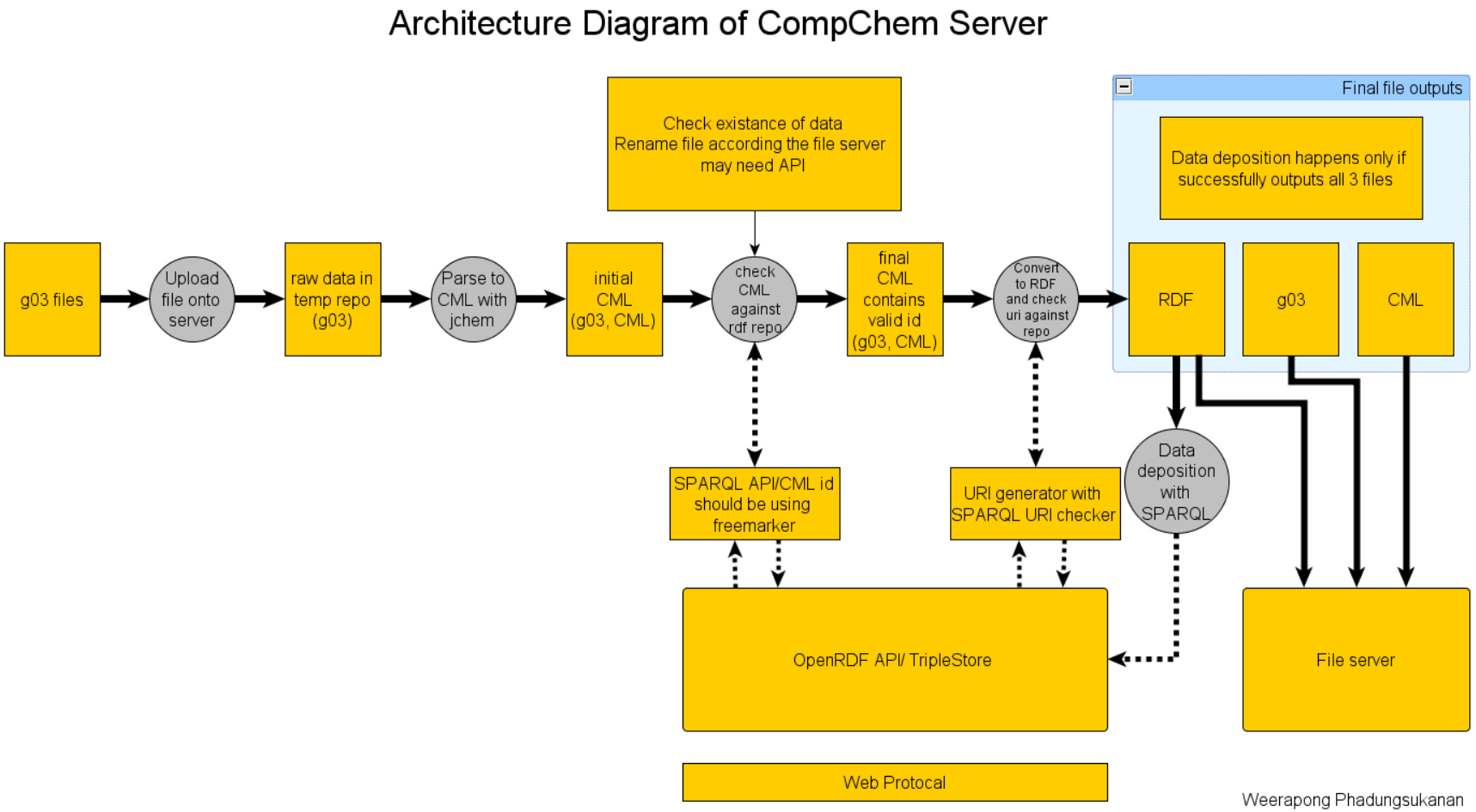
IUCr



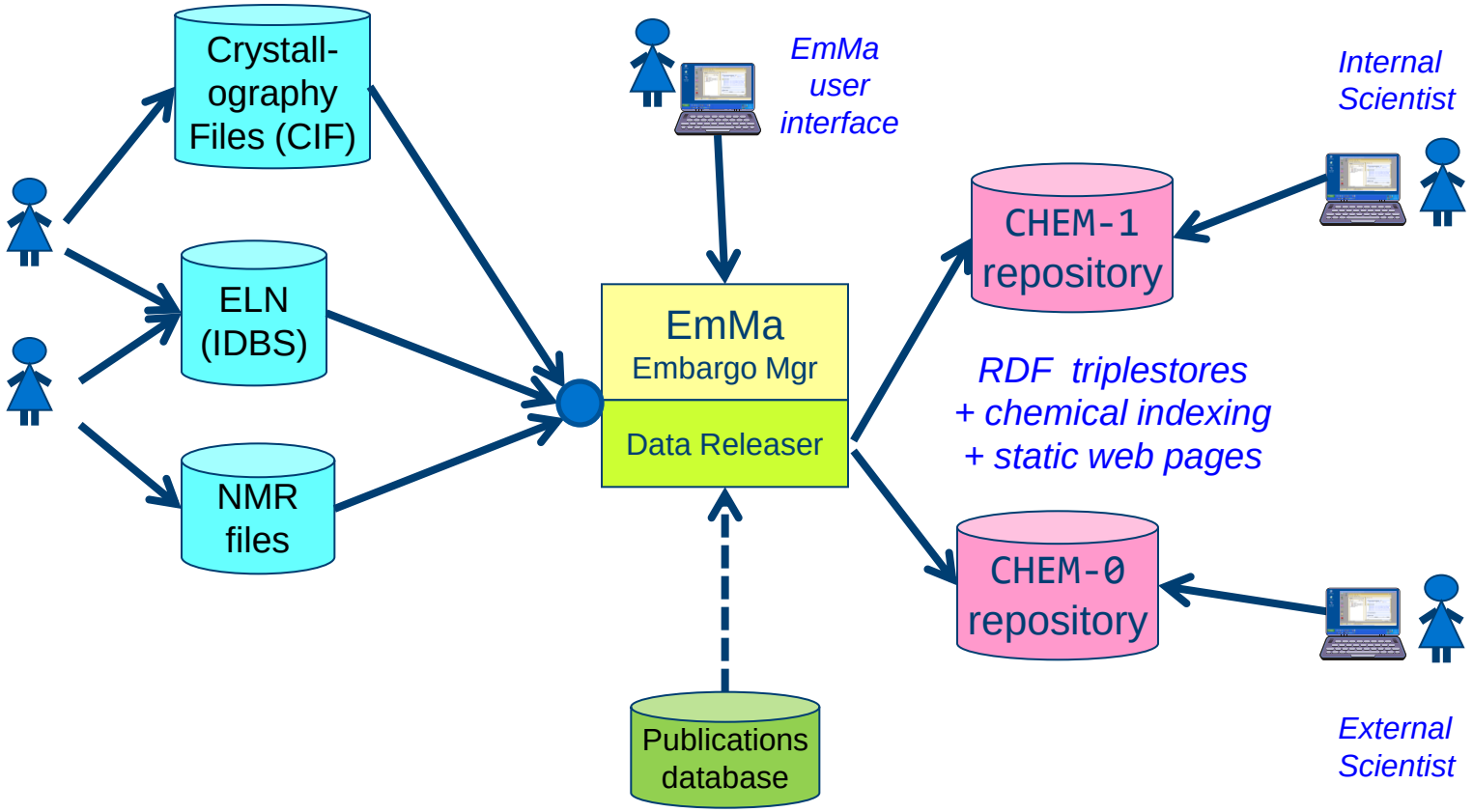




Lensfield


Architecture Diagram of CompChem Server
Check existance of data
Rename file according the file server
may need API
Final file outputs
Data deposition happens only if
successfully outputs all 3 files
g03 files
Upload
file onto
server
raw data in
temp repo
(g03)
Parse to
CML with
jchem
initial
CML
(g03, CML)
check
CML
against
rdf repo
final
CML
contains
valid id
(g03, CML)
Convert
to RDF
and check
uri against
repo
RDF
g03
CML
SPARQL API/CML id
should be using
freemarker
URI generator with
SPARQL URI checker
Data
deposition
with
SPARQL
OpenRDF API/ TripleStore
File server
Web Protocal
Weerapong Phadungsukanan

CLARION


Crystall-ography Files (CIF)
ELN (IDBS)
NMR files
EmMa user interface
EmMa
Embargo Mgr
Data Releaser
Publications database
CHEM-1 repository
CHEM-0 repository
RDF triplestores
+ chemical indexing
+ static web pages
Internal Scientist
External Scientist

Tools

Lensfield2

‘make’ for data

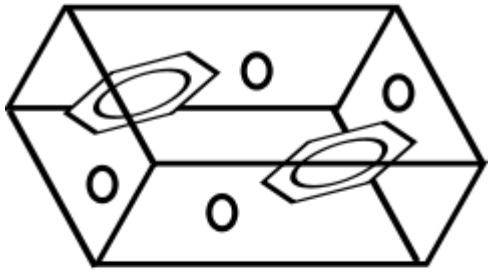
handles both file *and* code versioning

Java oriented

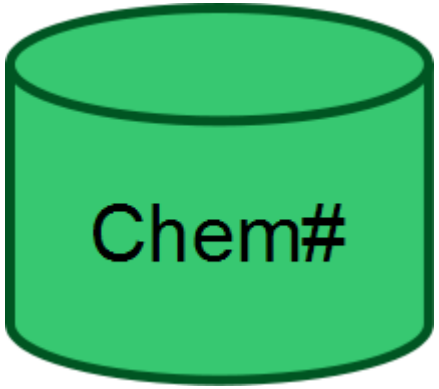
Chempound (chem#)

Semantic Chemical Data Repository

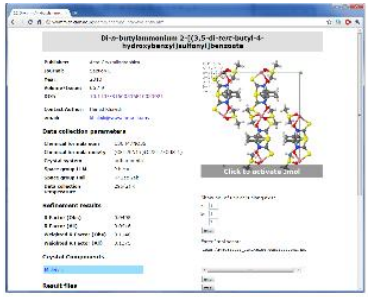
Chempound



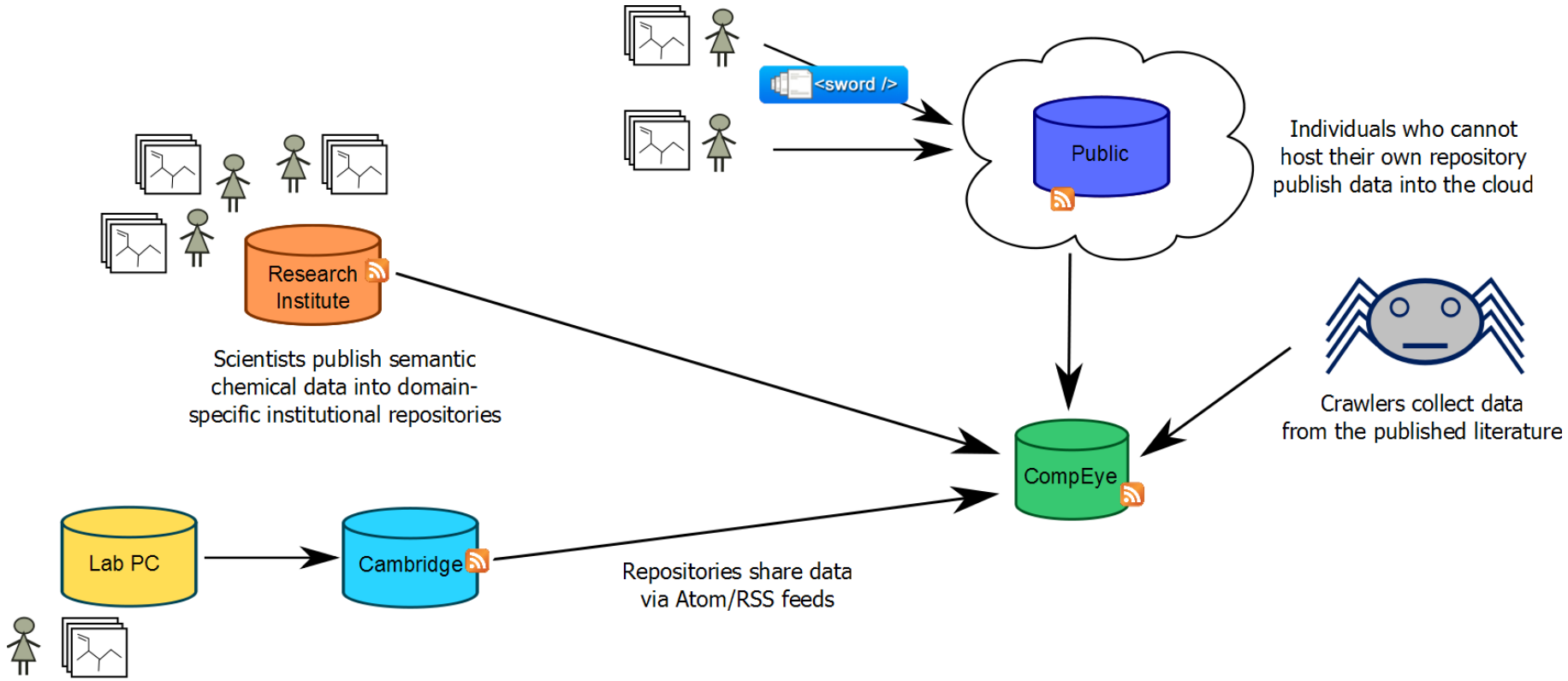
linked open data:
the chemical semantic web



Chempound stores
legacy and semantic files
indexed using RDF



Vision



Public

Individuals who cannot host their own repository publish data into the cloud

Research Institute

Scientists publish semantic chemical data into domain-specific institutional repositories

Crawlers collect data from the published literature

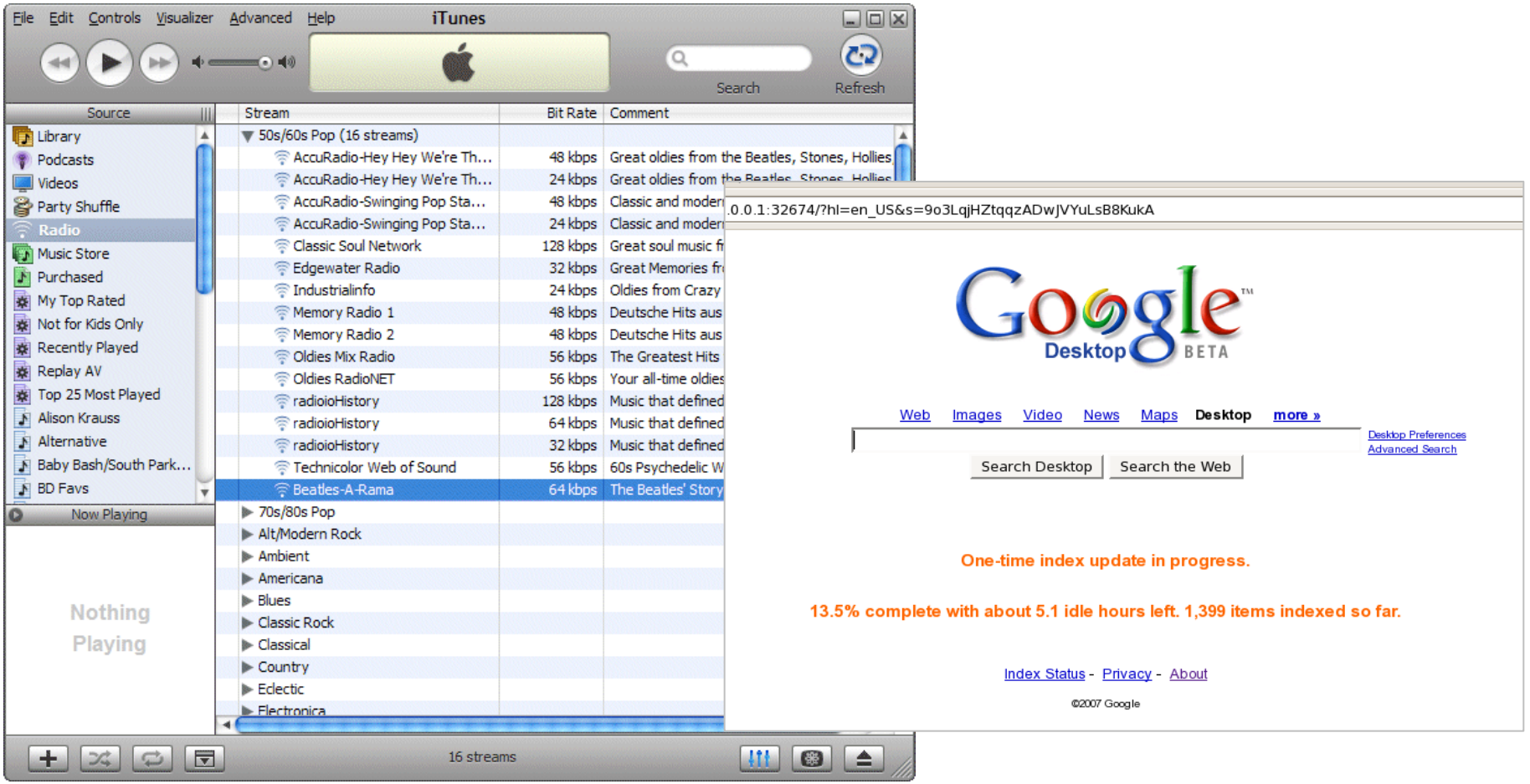
CompEye

Lab PC

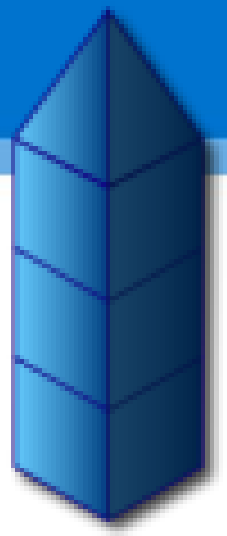
Cambridge

Repositories share data via Atom/RSS feeds

Fit existing workflows



Blue Obelisk



- An internet group dedicated to creating open interoperable resources for chemistry. Some of the software is the *de facto* approach in specific sub-domains.

Open Data, Open Standards, Open Source

Reproducible science

Avogadro

Bioclipse

cinfony

CDK

NMRShiftDB

Summary

- Data drives science, but masses of scientific data is currently lost
- Publication needs to be easy – fit into scientist's existing workflows
- Lightweight, modular technologies

Faster Science, New Discoveries, Avoid Duplication, Improve Repeatability, Advertise Work, Better (communal) Tools, Funder Mandates, Improved Data Management

"rough consensus and running code"
– IETF http://www.ietf.org/tao.html

methane-gaussian.fchk* - Avogadro

File Edit View Build Select Extensions Settings Help

New Open Save Close Quit | Tool Settings... Display Settings...

Display Types

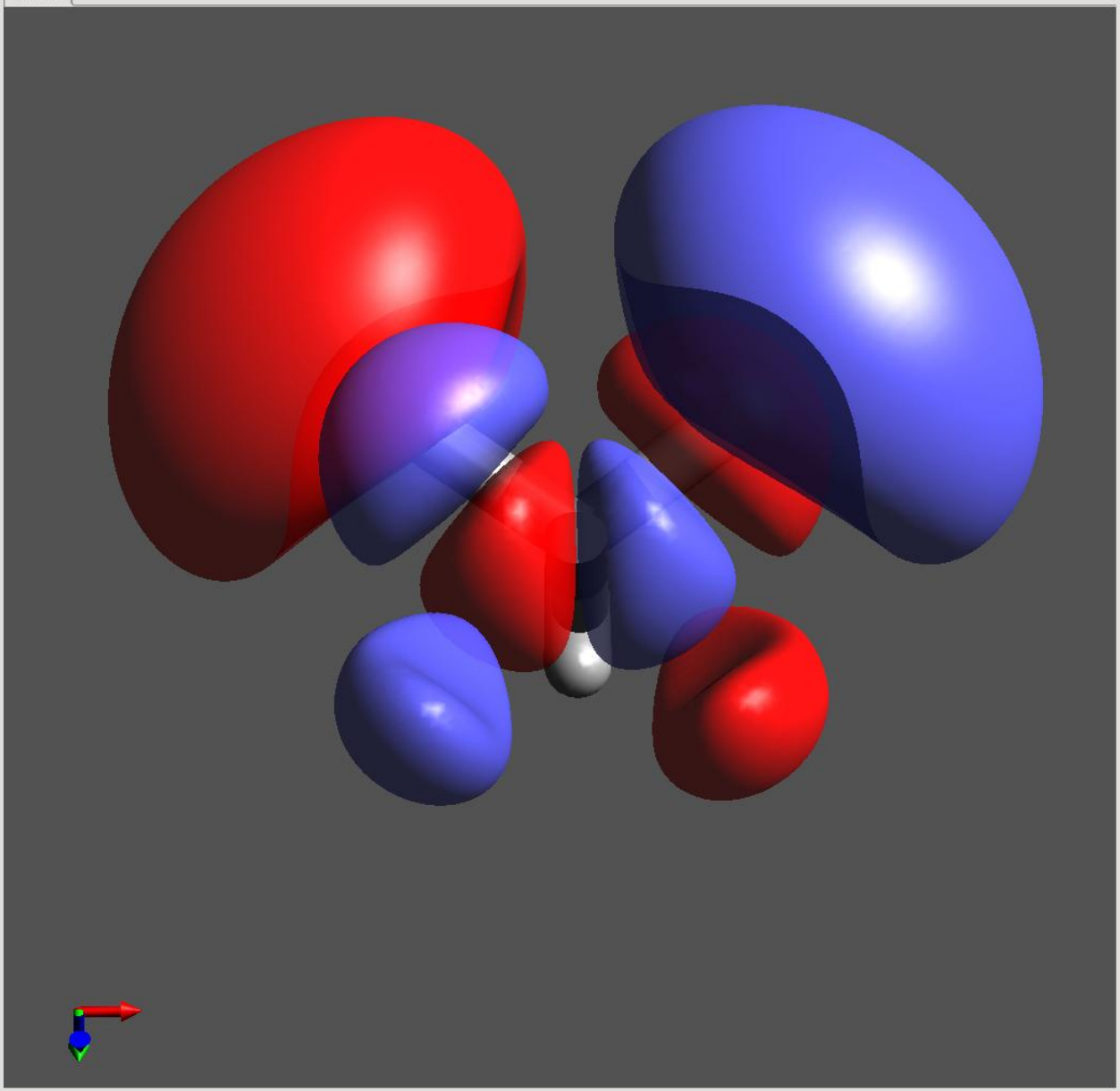
- [ ] Axes
- [ ] Ball and Stick
- [ ] Cartoon
- [ ] Dipole
- [ ] Force
- [ ] Hydrogen Bond
- [ ] Label
- [ ] Overlay
- [ ] Polygon
- [ ] QTAIM
- [ ] Ribbon
- [ ] Ring
- [ ] Simple Wireframe
- [x] Stick
- [x] Surfaces
- [ ] Van der Waals Spheres
- [ ] Wireframe

Add Duplicate Remove

Navigate Settings

- [x] Display visual cues

View 1



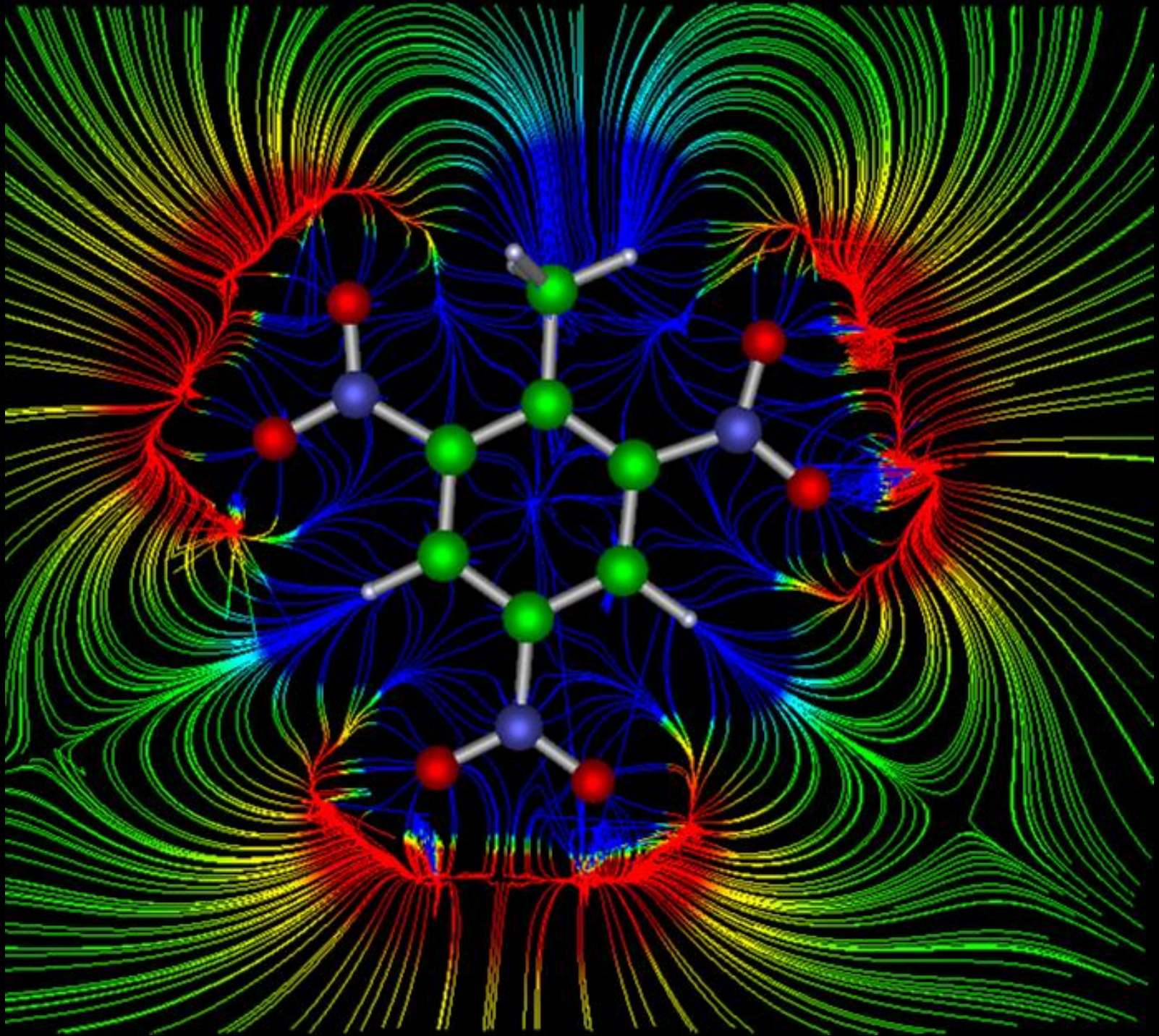
Messages

Orbitals

	Orbital	Energy	Status
17	LUMO+11	0	100%
16	LUMO+10	0	100%
15	LUMO+9	0	100%
14	LUMO+8	0	100%
13	LUMO+7	0	100%
12	LUMO+6	0	100%
11	LUMO+5	0	100%
10	LUMO+4	0	100%
9	LUMO+3	0	100%
8	LUMO+2	0	100%
7	LUMO+1	0	100%
6	LUMO	0	100%
5	HOMO	0	100%
4	HOMO-1	0	100%
3	HOMO-2	0	100%
2	HOMO-3	0	100%
1	HOMO-4	0	100%

Quality: Very High Render

Configure



Acknowledgements

Cambridge:
Peter Murray-Rust
Joe Townsend
Jim Downing
Nick Day
Weerapong Phadungsukanan

Visualisations:
Jens Thomas

Quixote Project

Thank you for listening

sea36@cam.ac.uk

The Quixote Project

http://groups.google.com/group/quixote-qcdb

http://quixote.wikispot.org/