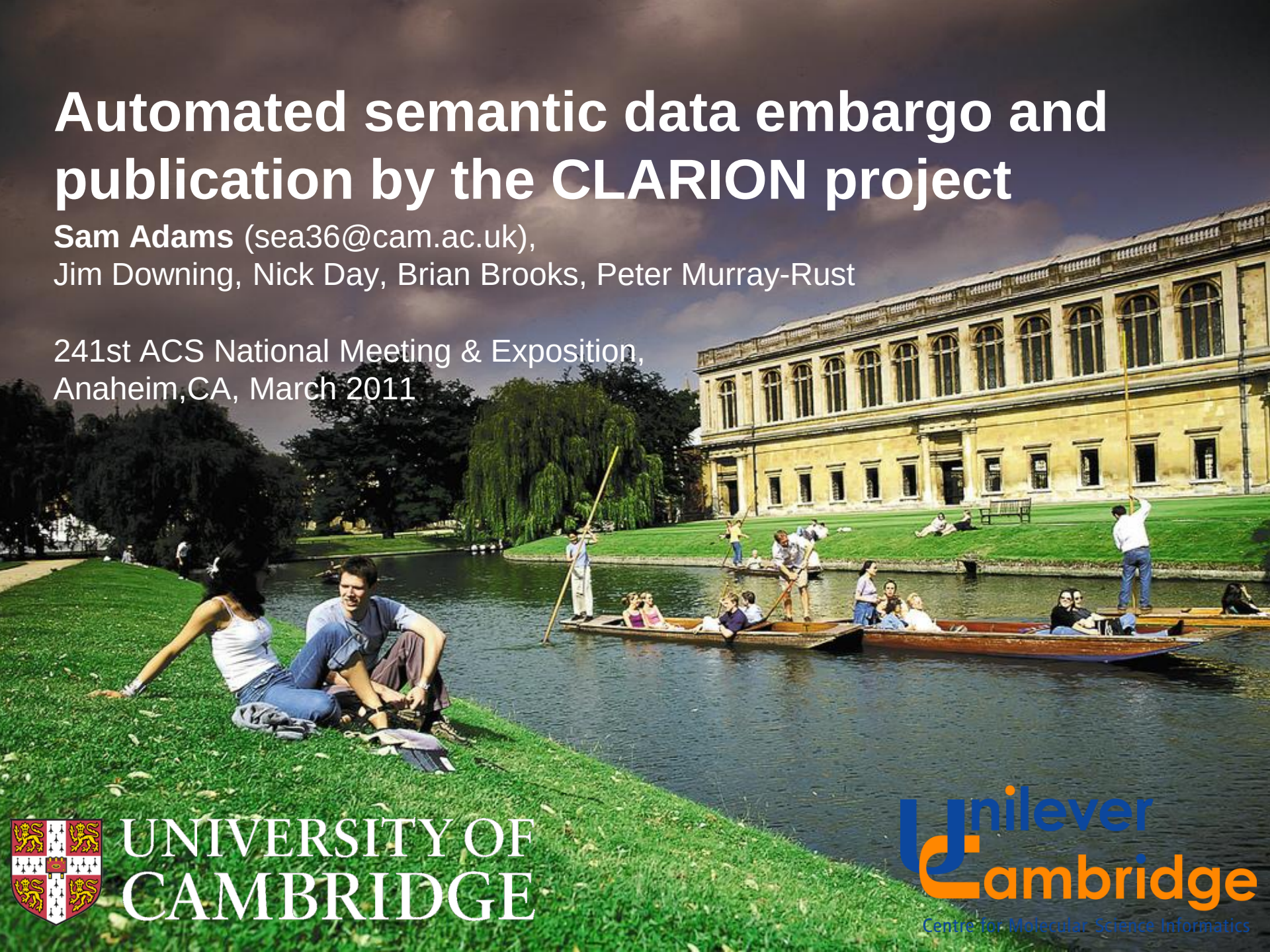
Automated semantic data embargo and publication by the CLARION project

Sam Adams (sea36@cam.ac.uk),
Jim Downing, Nick Day, Brian Brooks, Peter Murray-Rust

241st ACS National Meeting & Exposition,
Anaheim,CA, March 2011



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Science Informatics

Overview

Introduction

Data publication – current practices and issues

Semantic data publication

CLARION project

Conclusions

Data helps spot fraud

(IUCr) Editorial

journals.iucr.org/e/issues/2010/01/00/me0406/me0406bdy.html

editorial

Acta Crystallographica Section E
Structure Reports Online
ISSN 1600-5368

Volume 66
Part 1
Pages e1-e2
January 2010

Online 19 December 2009

cited in
download citation

Editorial

William T. A. Harrison,[a] Jim Simpson[b] and Matthias Weil[c]

[a]Department of Chemistry, University of Aberdeen, Aberdeen AB24 3UE, Scotland, [b]Department c
Analytics, Division of Structural Chemistry, Vienna University of Technology, Getreidemarkt 9/16-

Regrettably, this editorial is to alert readers and authors of *Acta Crystallographica Se* extensive series of scientific frauds involving papers published in the journal, principally every year will continue to reflect results of serious scientific work, the extent of these p acknowledged by the authors as such. Our work is ongoing and it is likely that this figu

These problems were first discovered by Ton Spek during testing of the checking prog of *Acta Crystallographica Sections E* or *C*. Initially, unexplained Hirshfeld rigid-bonc transposed and that more than one structure had been `determined' using identical sets involved.

A program written by Toine Schreurs of Utrecht University that can examine and compare two structure-factor files was then used the program revealed that the data sets used to refine two or more supposedly unique structures were in fact identical, but with the

The falsified structures have many features in common: in each case, a *bona fide* set of intensity data, usually on a compound whos of papers, with the authors changing one or more atoms in the structure to produce what appeared to be a genuine structure determ structures from a single common set of data. There is nothing to suggest that the authors of the original papers describing the real str

Bogus refinements were found for both metal-organic and organic structures. The most common ploy was to acquire a data set for iron(II) or even cobalt(III) produced papers reporting seemingly novel compounds. In order to decrease the risk of detection, chan parameters and also the culling of some reflections from the data sets. The scale of the problems ruled out the possibility of mere inc

Similar procedures with structures containing lanthanide elements offered even greater scope for deception. In addition to changing structures falsely reported.

Non-metal atom substitutions also generated numerous bogus organic structures. CH_2 groups were replaced by NH or O and *vice* is extensive. The residuals on the resulting fraudulent refinements were generally worse than those of the genuine material but not suf structures arose from these manipulations, and it is a concern and disappointment that these chemical features passed into the literat

The initial set of falsified structures arises from two groups. The correspondence authors are Dr H. Zhong and Professor T. Liu, bot from Jinggangshan University together with authors from different institutions in China. Both these correspondence authors and all c Professor Liu. Details of these retractions appear elsewhere in this issue of the journal. Having found these problems with articles fr

.com/news/2009/091222/full/462970a.html

nature news

nature news home | news archive | specials | opinion | features | news blog | nat

comments on this story

Stories by subject
- Cell and molecular biology
- Health and medicine
- Lab life

Stories by keywords
- Protein
- Structure
- Crystallography
- Misconduct
- Fraud
- Dengue

This article elsewhere
Blogs linking to this article
Add to Connotea
Add to Digg
Add to Facebook
Add to Newsvine
Add to Del.icio.us
Add to Twitter

Published online 22 December 2009 | *Nature* **462**, 970 (2009) | doi:10.1038/462970a

News

Fraud rocks protein community

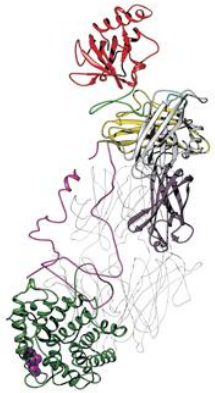
University finds that researcher falsified data supporting 11 protein structures.

Brendan Borrell

The finding by a university misconduct investigation that a crystallographer "more likely than not" faked almost a dozen protein structures has left the field in shock. The fraud is the largest ever in protein crystallography. The disputed structures had important implications for discovering drugs against dengue virus and for understanding the human immune system.

"It's massive," protein crystallographer Wayne Hendrickson of Columbia University in New York says of the investigation's conclusion. "It's the worst possible thing."

In a report released earlier this month, the University of Alabama at Birmingham concluded that H. M. Krishna Murthy acted alone in fabricating and falsifying results that appeared in ten papers[1,2,3,4,5,6,7,8,9,10] published during the past decade. The disputed papers have been cited more than 450 times



The first of the protein structures to be disputed, that for human C3b. *Ref. 10*

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge

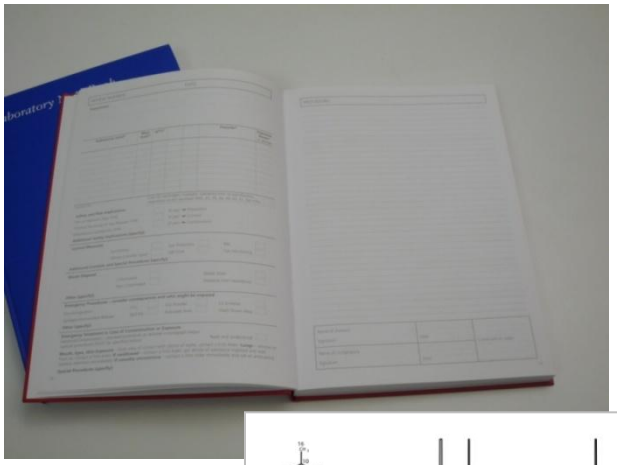
Science can be wrong

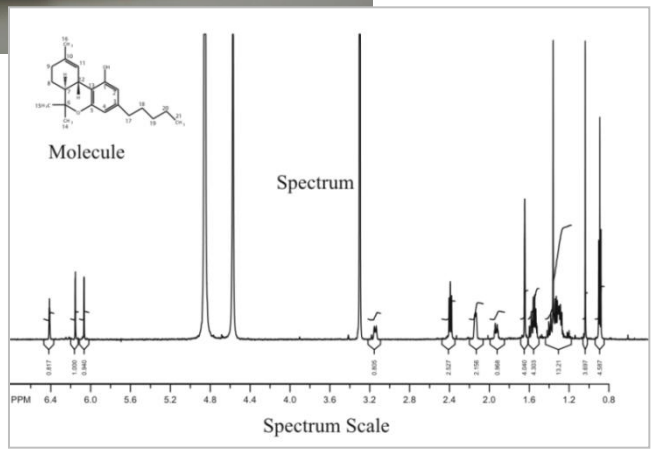
"*We have recently attempted to perform a diagnostic meta-analysis, and found that most of the relevant papers reported only frequency distributions. Some papers reported individual patient data in scatterplots, from which we attempted to derive the original datasets by a computer-aided method.* ***To our surprise, nearly half the papers showed a different number of data points compared to the stated number of included patients.*** *As a result, we were unable to aggregate the data.*"

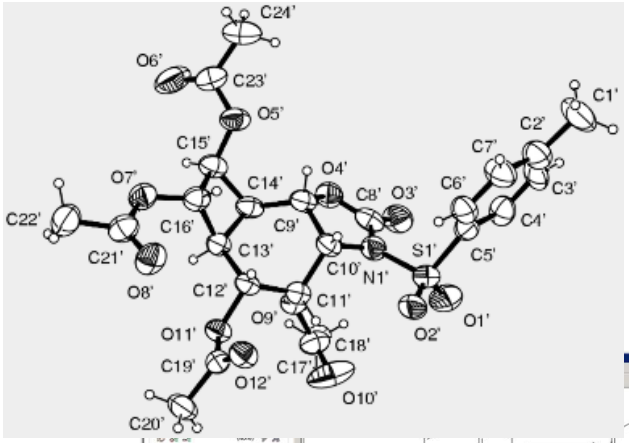
– Gustav Nilsonne, Karolinska Institutet

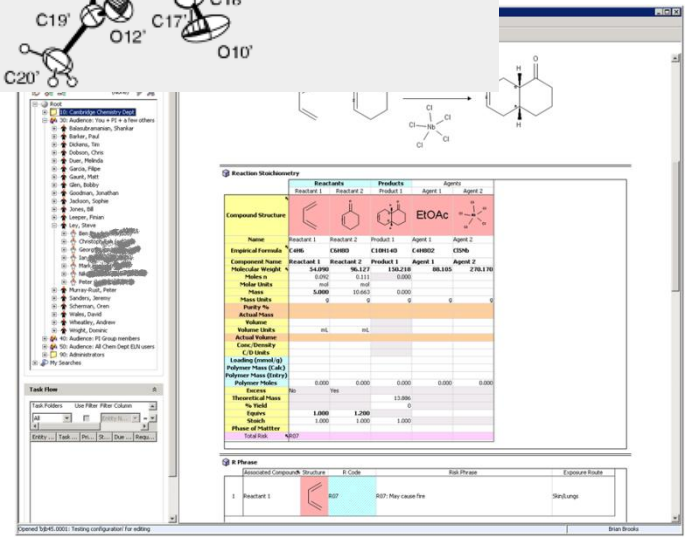
http://blogs.openaccesscentral.com/blogs/bmcblog/entry/join_the_data_debate_draft#comment-1284221254522

Most data is never published

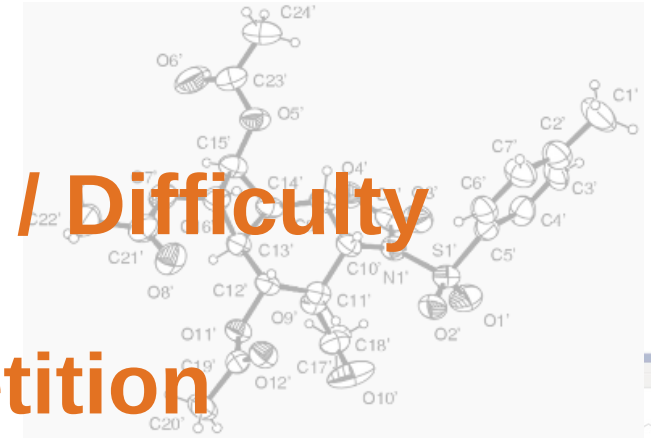








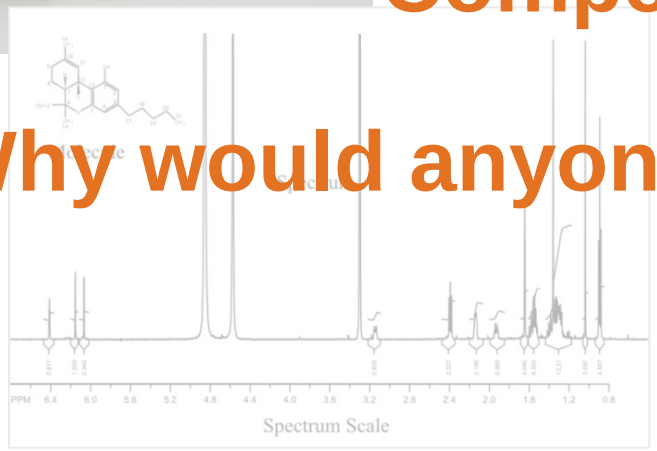
Most data is never published

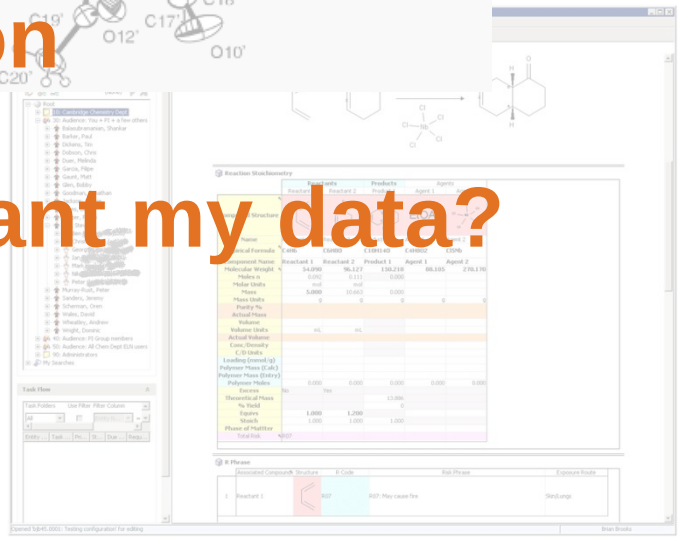


Time / Effort / Difficulty

Competition

Why would anyone want my data?





Lots of data does get published

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

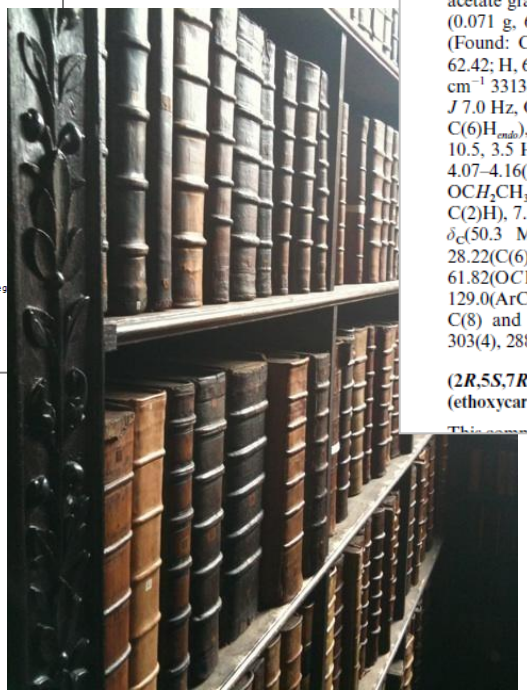
DOCTOR OF PHILOSOPHY

OF THE

UNIVERSITY OF CAMBRIDGE

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

Wolfson College
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

This compound was prepared from amine 13a on a 0.067 g scale

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), 3.98–4.01(2H, m, NCHPh, and C(2)H), 4.08–4.20(2H, m, OCH_2

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

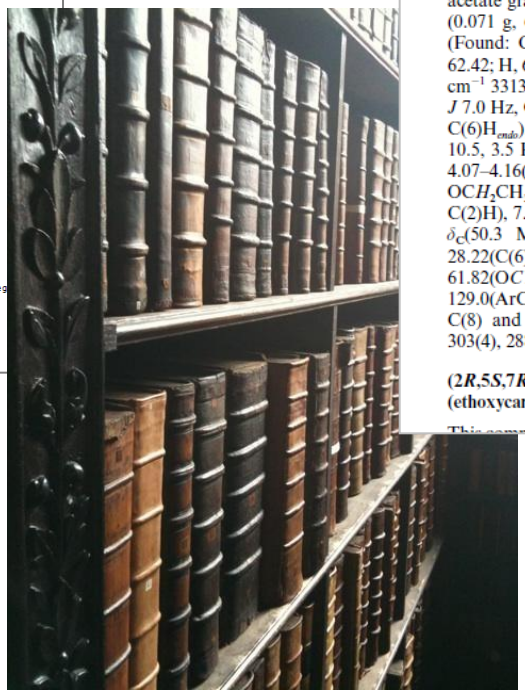
DOCTOR OF PHILOSOPHY

OF THE

UNIVERSITY OF CAMBRIDGE

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

Wolfson College
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

This compound was prepared from amine **13a** on a 0.067 g scale

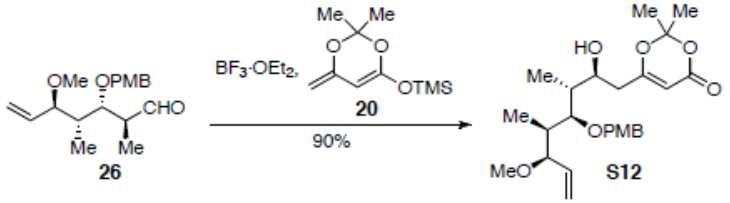
acetate) to give the product **14** as a colourless oil (40 mg, 61% over 2 steps): R_f 0.12 (1 : 6 petrol–ethyl acetate); ν_{max} (thin film)/cm^{-1} 2924(br m), 1737(s), 1700(s), 1667(s); δ_H(200 MHz, $CDCl_3$) 1.14(3H, t, *J* 7.0 Hz, OCH_2CH_3), 2.03(3H, s, $CH_3C(O)$), 2.12–2.24(1H, m, C(6)H_{endo}), 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), 3.98–4.01(2H, m, NCHPh and C(2)H), 4.08–4.20(2H, m, OCH_2

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.

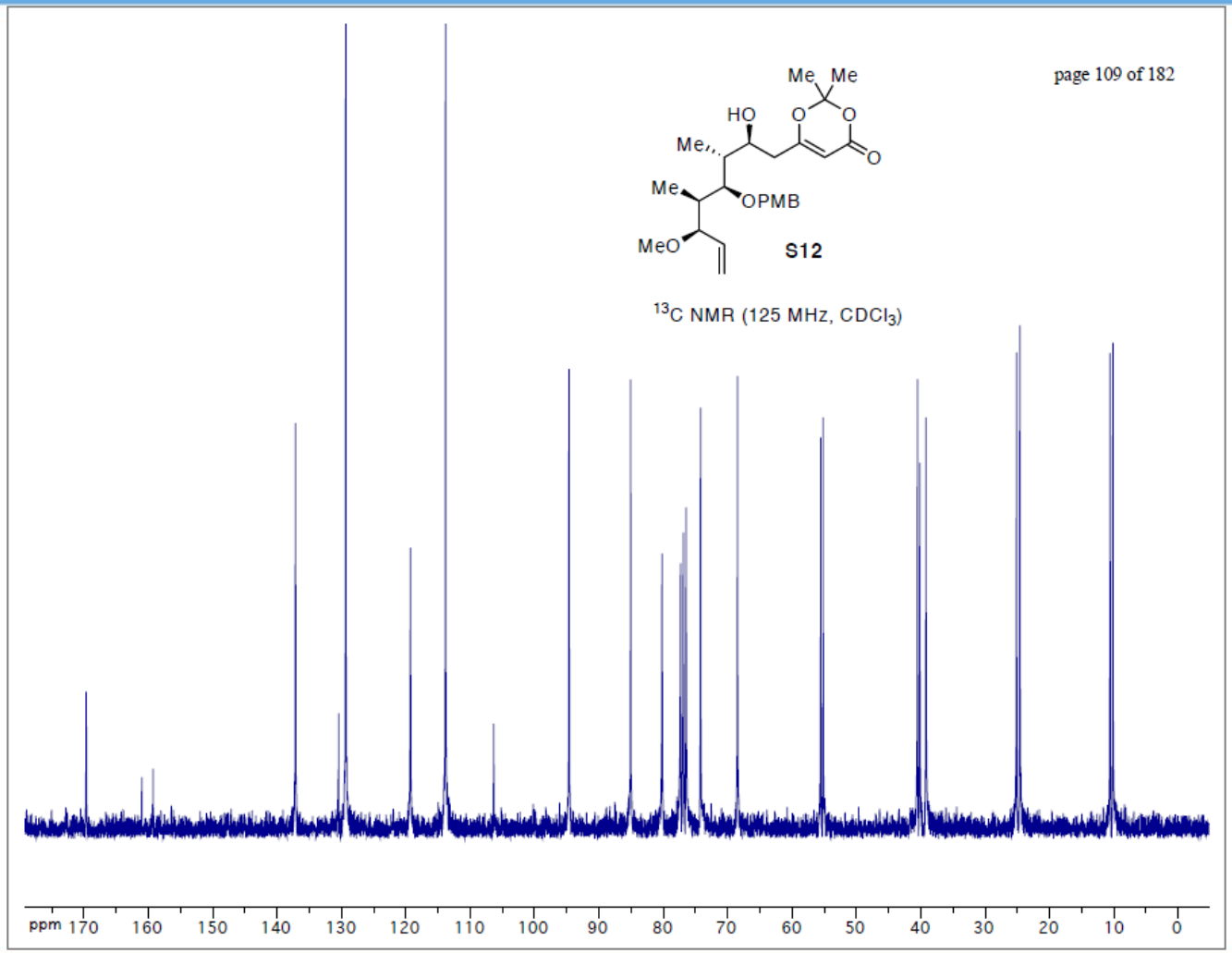
IR (neat) 3470, 2974, 2934, 1728, 1634, 1514, and 1249 cm^{-1}.

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).

Supporting information



We should be publishing the raw data

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

... and some are very bad

Supplementary Material (ESI) for Chemical Communications

Supplementary Material

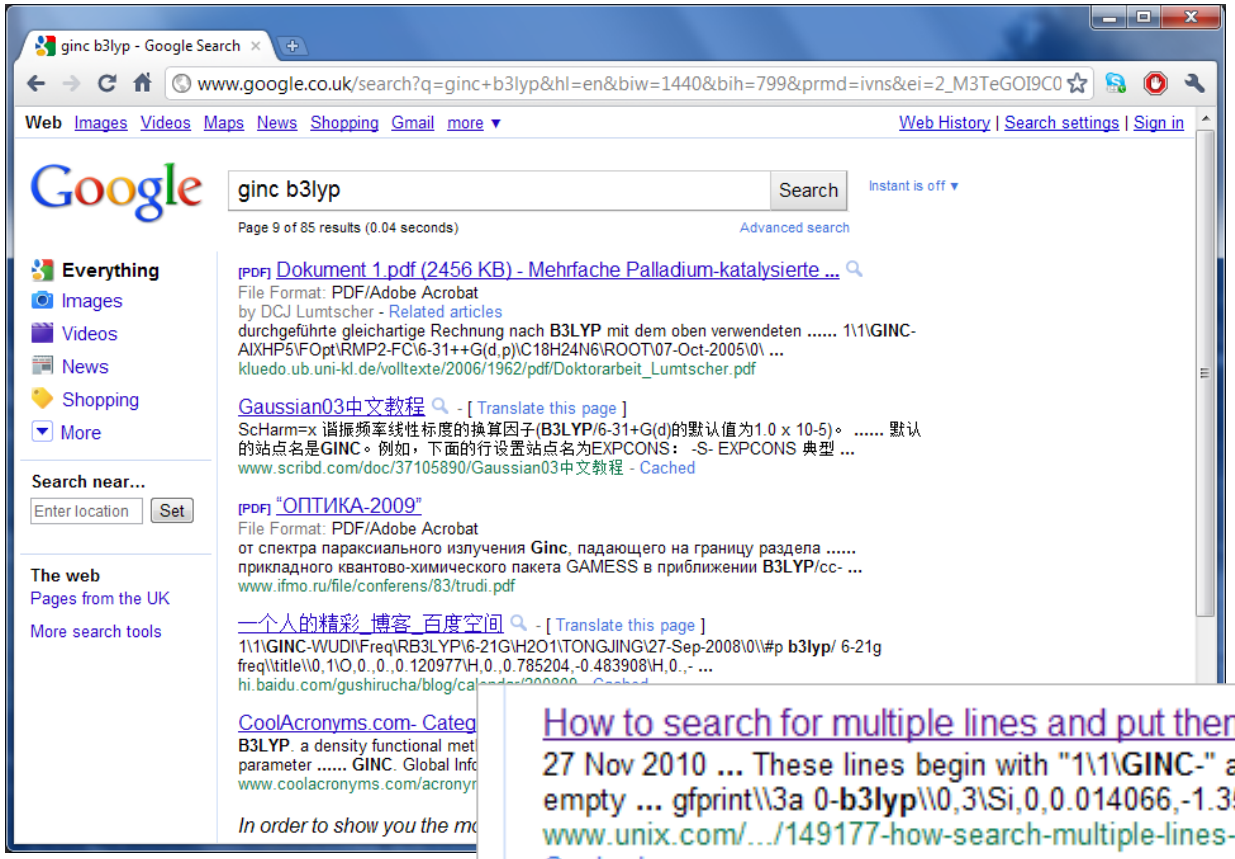
Unexpected dual orbital effects in radical addition reactions involving acyl, silyl and related radicals

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

MP2/6-311G**

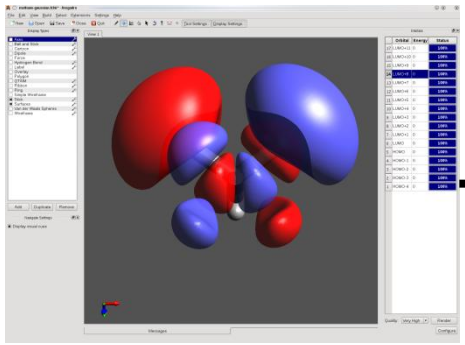
1\1\ CHEMISTRY CLUSTER KIRKLAND-KNET5\FTS\UMP2-FC\6-311G(d,p)\C3H6N1O1(2)\HIROSHI\16-May-2005\1\\#MP2/6-311G** SCF=DIRECT OPT=(TS,EF,CALCHFFC,MAXCYCLE=100) NOSYMM FREQ=NORAMAN\\TS for addition to nitorgen of imine\\0,2\C\O,1,r2\C,1,r3,2,a3\N,1,r4,2,a4,3,d4,0\C,4,r5,1,a5,2,d5,0\H,3,r6,1,a6,2,d6,0\H,3,r7,1,a7,2,d7,0\H,3,r8,1,a8,2,d8,0\H,4,r9,1,a9,2,d9,0\H,5,r10,4,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)]\\@

Virtually no data published at all



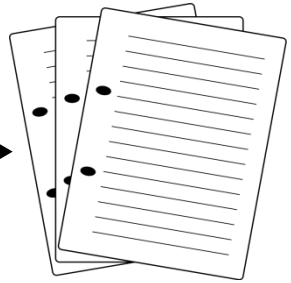
Waste of resources

- 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.

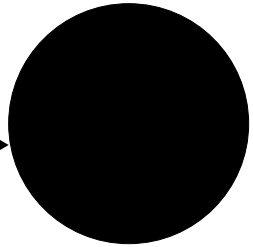


Generate data from scratch

Expensive computation



Cumbersome data format



Black hole

Data publication options are growing

Proteome Commons https://proteomecommons.org/

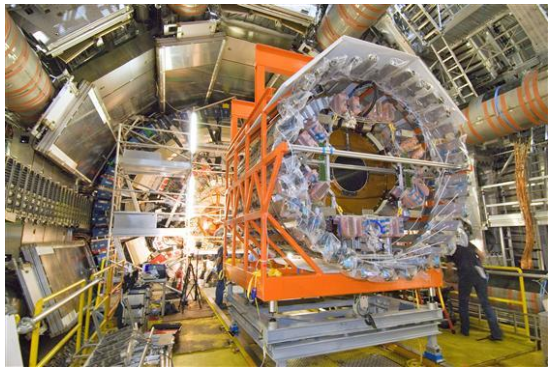
Tranche Project https://trancheproject.org/

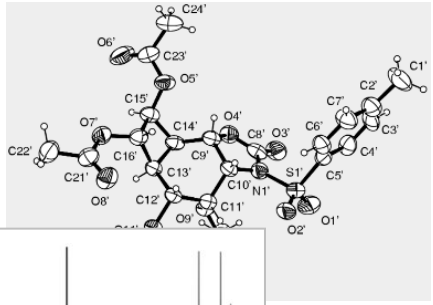
DRYAD http://datadryad.org/

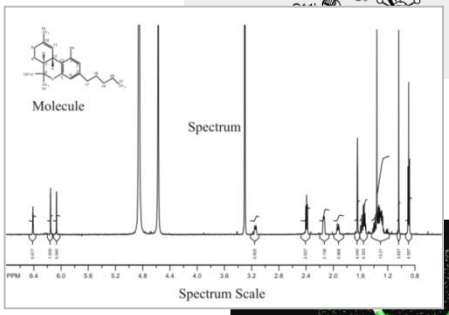
DSpace @Cambridge http://www.dspace.cam.ac.uk/

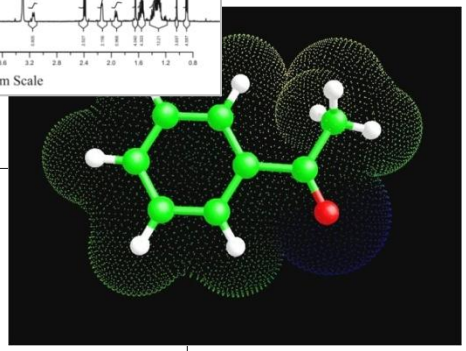
[n]² Talis Connected Commons
http://blogs.talis.com/n2/cc/

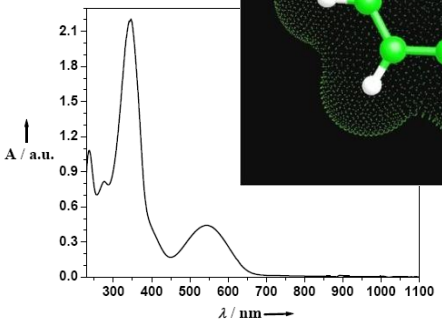
Different Scales of Data

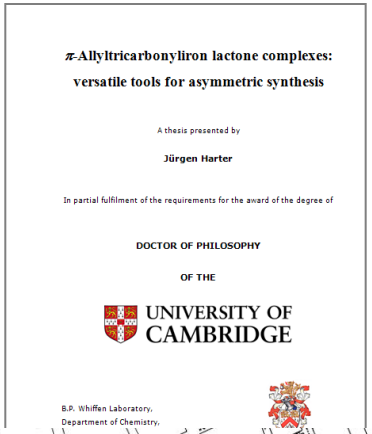


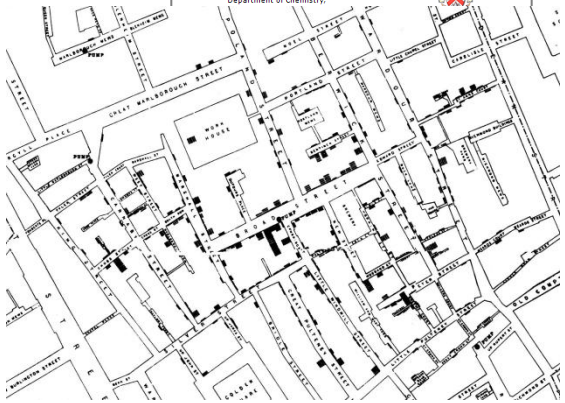




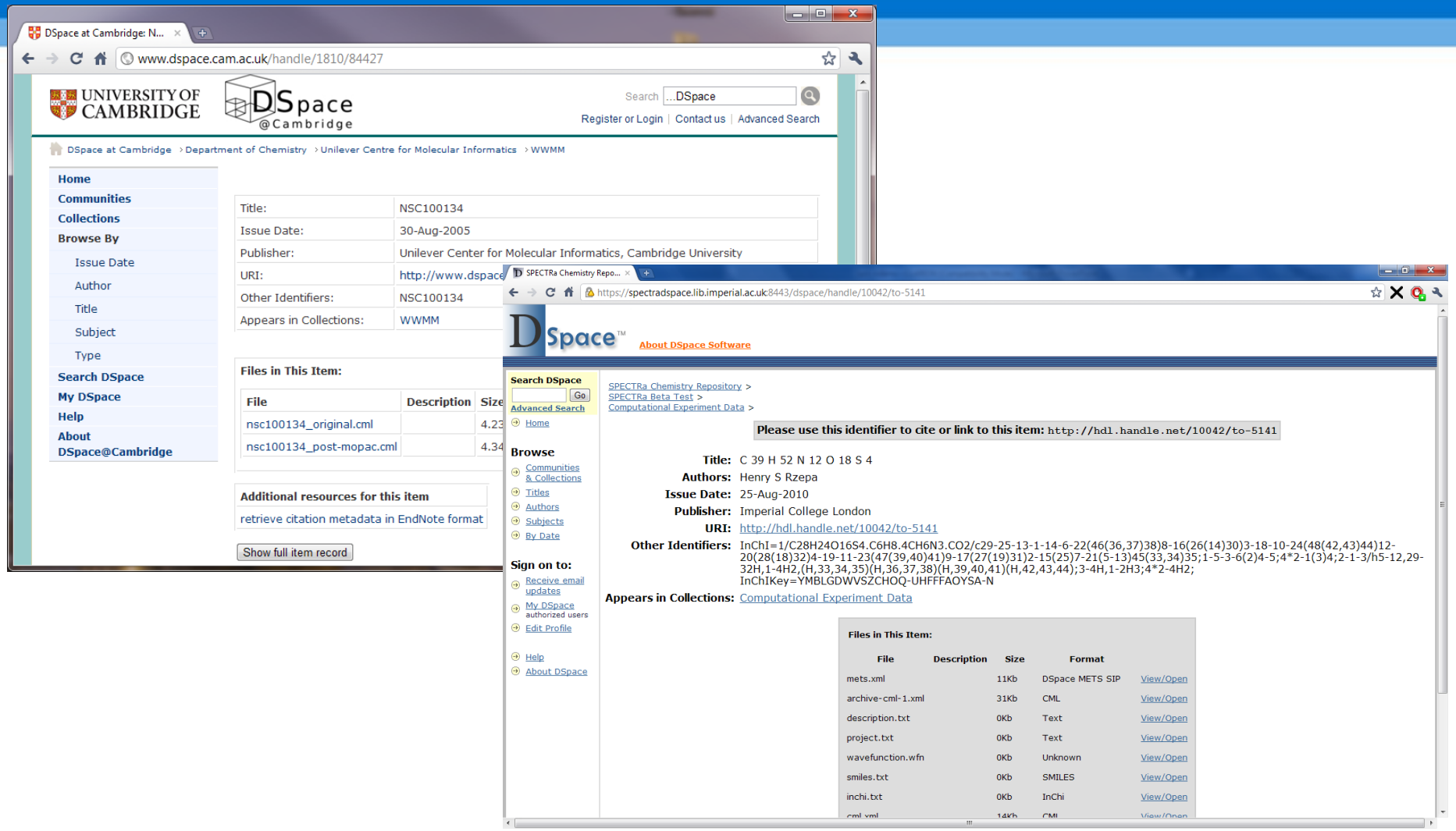








More than archiving: domain knowledge required



It needs to be simple

DC Field	Value	Language
dc.creator	US National Cancer Institute	en_GB
dc.date.accessioned	2005-08-30T09:49:52Z	-
dc.date.available	2005-08-30T09:49:52Z	-
dc.date.created	2003-02-01	en_GB
dc.date.issued	2005-08-30T09:49:52Z	-
dc.identifier	NSC100134	en_GB
dc.identifier.uri	http://www.dspace.cam.ac.uk/handle/1810/84427	-
dc.format.extent	4331 bytes	-
dc.format.extent	4440 bytes	-
dc.format.mimetype	chemical/x-cml	-
dc.format.mimetype	chemical/x-cml	-
dc.language.iso	en_GB	-
dc.publisher	Unilever Center for Molecular Informatics, Cambridge University	en_GB
dc.title	NSC100134	en_GB
dc.type	Other	en_GB
dc.identifier.ichi	C12H8ClN5O2,1H3-7-4H-8(2H-3H-9(7)18(19)20)17-6H-16-10-11(13)14-5H-15-12(10)17	en_GB
Appears in Collections:	WWMM	

Files in This Item:

File	Description	Size	Format	Checksum
nsc100134_original.cml		4.23 kB	CML	7052b48e3a9cd95bf7f459712

Login

Search Data

Sporulation Rate Data

dc:contributor.author	Hill, Jessica A.	
dc:contributor.author	Otto, Sarah P.	
dc:date.accessioned	2007-12-14T22:32:19Z	
dc:date.available	2007-12-14T22:32:19Z	
dc:date.issued	2007-12-14T22:32:19Z	
dc:identifier	doi:10.5061/dryad.18/3	*
dc:identifier.uri	http://hdl.handle.net/10255/dryad.19	
dc:relation.ispartof	http://hdl.handle.net/10255/dryad.18	
dc:title	Sporulation Rate Data	en_US
dc:type	Dataset	en_US
dryad:status	scanned	

Files in this item

Files	Size	Format	View
sporedata.xls	174.0Kb	Microsoft Excel	View/ Open

04src0553 - C37H60ClFe...

ecrystals.chem.soton.ac.uk/654/

eCrystals

University of Southampton

Home | About | Browse by Year | Browse by People

Login | Create Account

Search

04src0553 - $C_{37}H_{60}ClFeO_4P_2Rh$

Sample Originator: I. R. Butler[b].

Data Collection: Simon J. Coles[a]

Structure Determination: Simon J. Coles[a] and Michael B. Hursthouse[a].

University of Southampton[a]
University of Wales, Bangor[b]

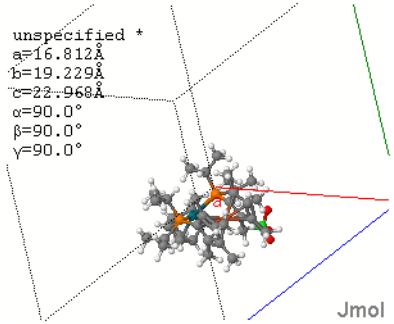
$C_{37}H_{60}ClFeO_4P_2Rh$

InChI=1/C15H28P.C15H26P.C7H10.ClHO.Fe.3H2O.Rh/c2*1-9(2)16(10(3)4)15-13(7)11(5)12(6)14(15)8;1-2-7-4-3-6(1)5-7;1-2;;;;;/h9-10,13-14H,1-8H3;9-10H,1-8H3;1,3,6-7H,2,4-5H2;2H;;3*1H2;/t;;6-,7+;;;;;;

Identification Number:	10.3737/ecrystals.chem.soton.ac.uk/654
Date Created:	02 March 2009
Deposited On:	02 Mar 2009 18:19
Deposited By:	Dr Simon J Coles

Data collection parameters

Chemical formula	C37 H60 Cl Fe O4 P2 Rh
Crystal morphology	Block
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	16.8121(12)
Cell length b	19.229(2)
Cell length c	22.968(3)



Available Files

Final Result

04src0553.cif	27k
04src0553.cml	14k
04src0553.fcf	415k

Validation

04src0553_checkcif.htm	9k

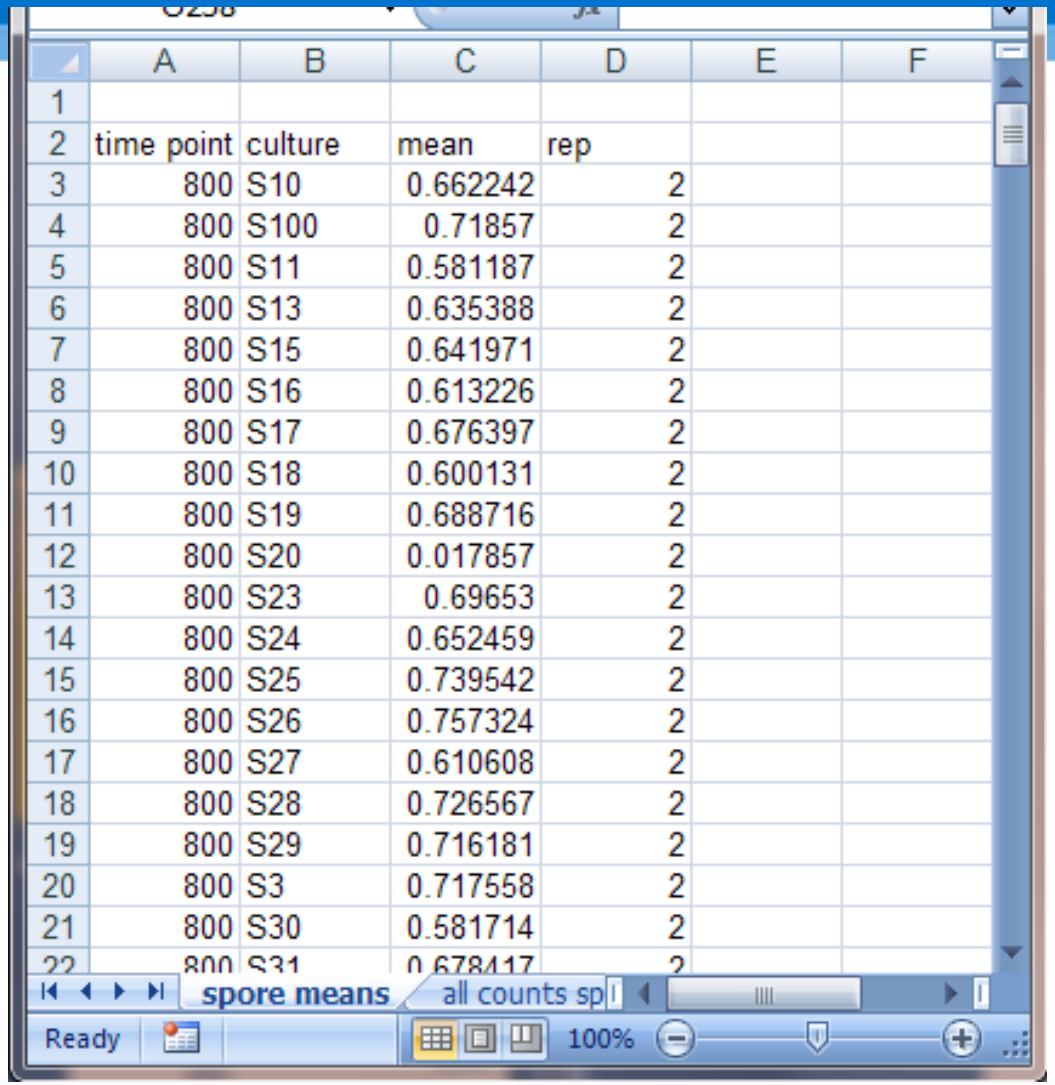
Refinement

04src0553.res	12k
04src0553_xl.lst	66k

Solution

04src0553.prp	5k

Semantics and provenance are vital

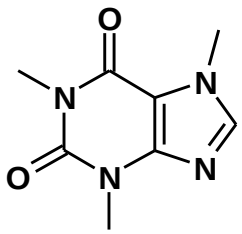


	A	B	C	D	E	F
1						
2	time point	culture	mean	rep		
3	800	S10	0.662242	2		
4	800	S100	0.71857	2		
5	800	S11	0.581187	2		
6	800	S13	0.635388	2		
7	800	S15	0.641971	2		
8	800	S16	0.613226	2		
9	800	S17	0.676397	2		
10	800	S18	0.600131	2		
11	800	S19	0.688716	2		
12	800	S20	0.017857	2		
13	800	S23	0.69653	2		
14	800	S24	0.652459	2		
15	800	S25	0.739542	2		
16	800	S26	0.757324	2		
17	800	S27	0.610608	2		
18	800	S28	0.726567	2		
19	800	S29	0.716181	2		
20	800	S3	0.717558	2		
21	800	S30	0.581714	2		
22	800	S31	0.678417	2		

spore means | all counts sp

Ready | 100%

Semantics and provenance are vital



Temperature	Solubility g/l	Year
25	2.132	1926 [1]
25	896.2	1985 [2]
25	21.0	2002 [3]
25	49.79	Merck Index
25	18.67	2005 [4]
25	21.6	SRC PhysProp Database

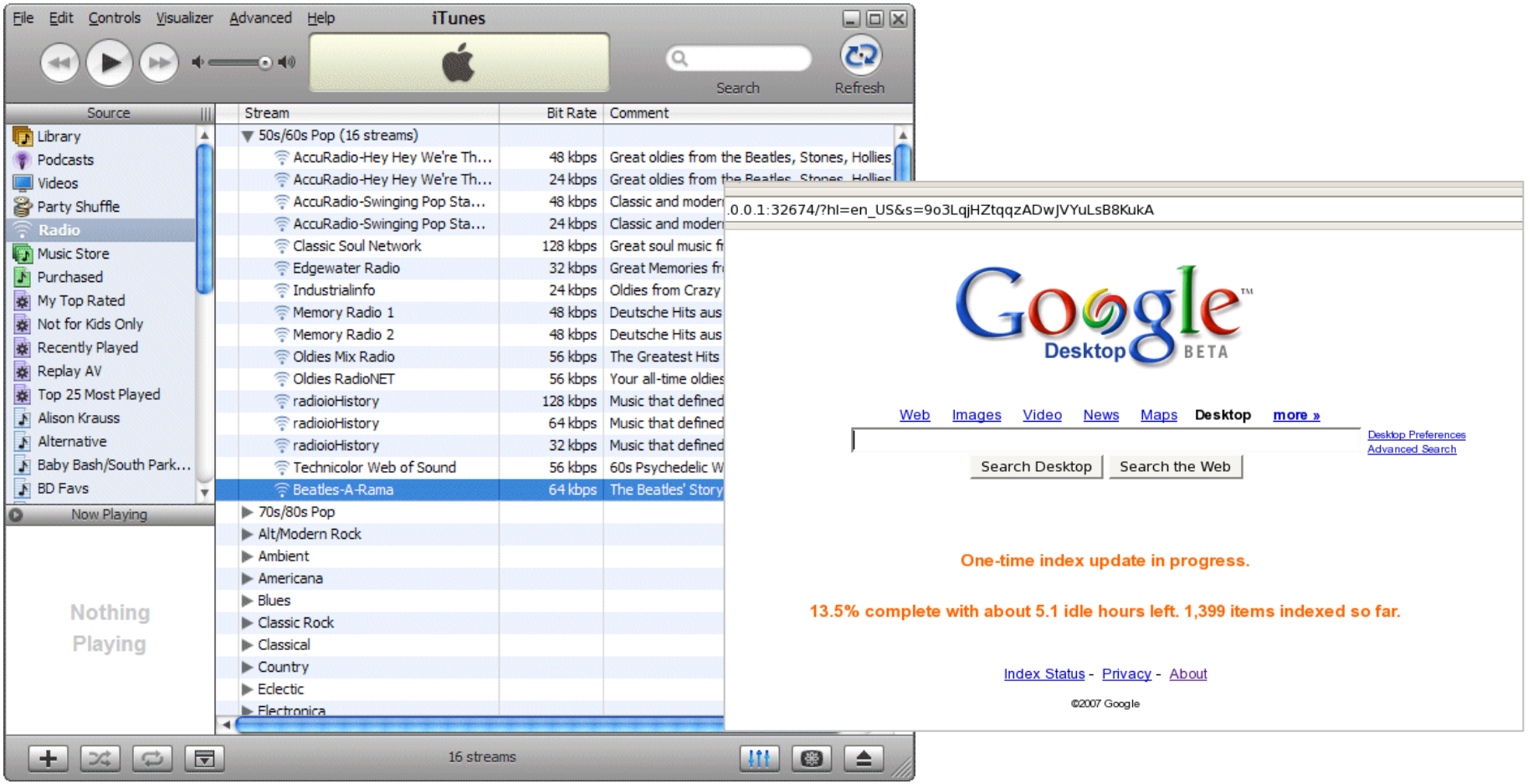
[1] Oliveri-Mandala, E. (1926), *Gazzetta Chimica Italiana 56*, 896-901

[2] Ochsner, A. B., Belloto, R. J., and Sokoloski, T. D. (1985), *Journal of Pharmaceutical Sciences 74*, 132-135

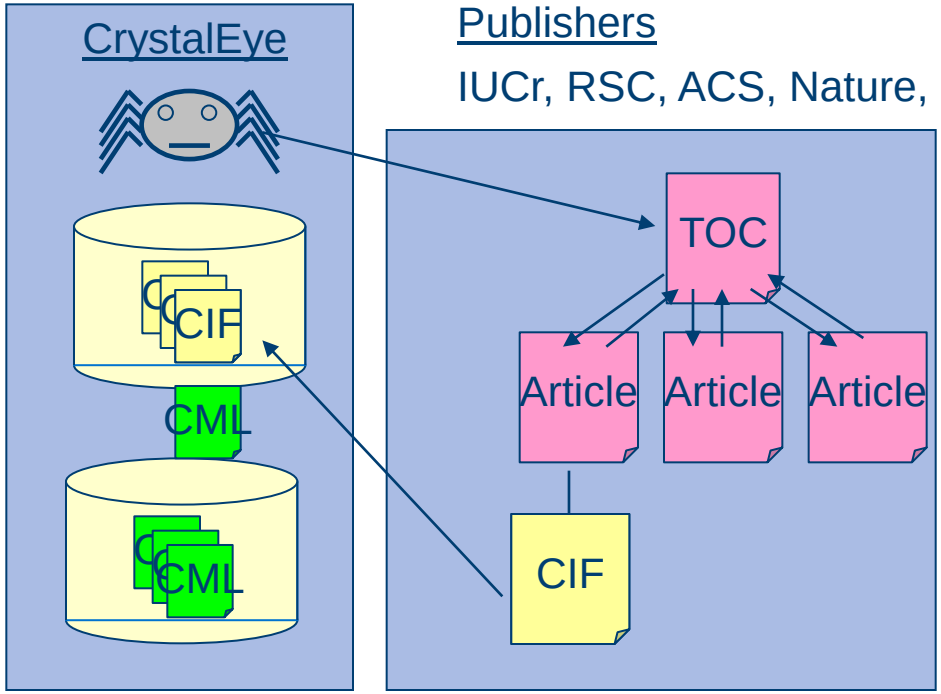
[3] Al-Maaieh, A., Flanagan, D. R. (2002), *Journal of Pharmaceutical Sciences 91*, 1000-1008

[4] Rytting, Erik, Lentz, Kimberley A., Chen, Xue-Qing, Qian, Feng, Venkatesh, Srini. AAPS Journal (2005), 7(1), E78-E105.

Publishing data is difficult

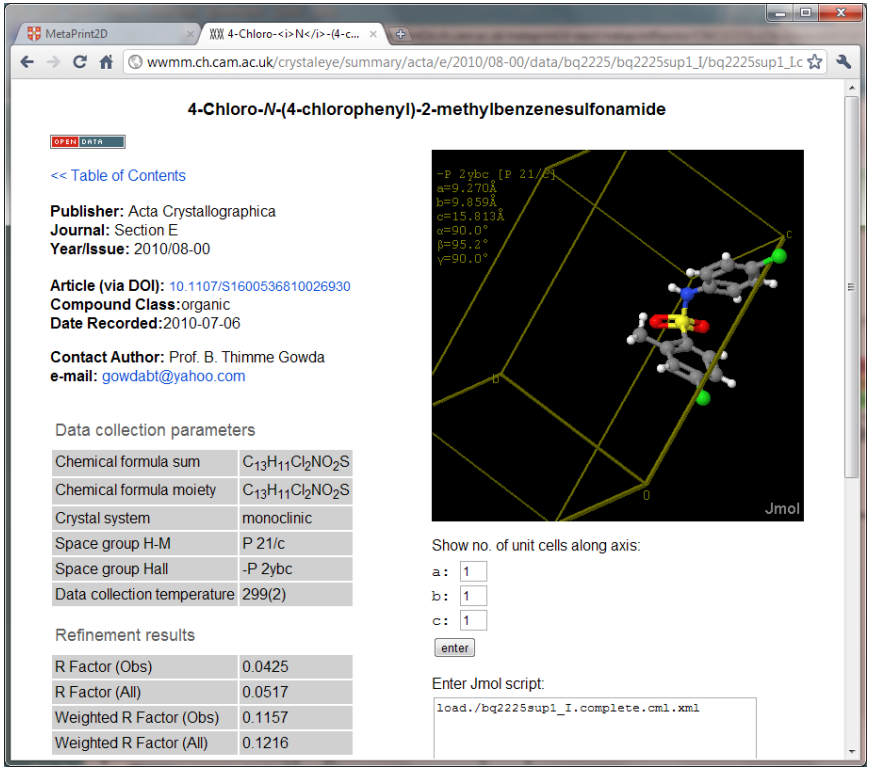


Data is really useful

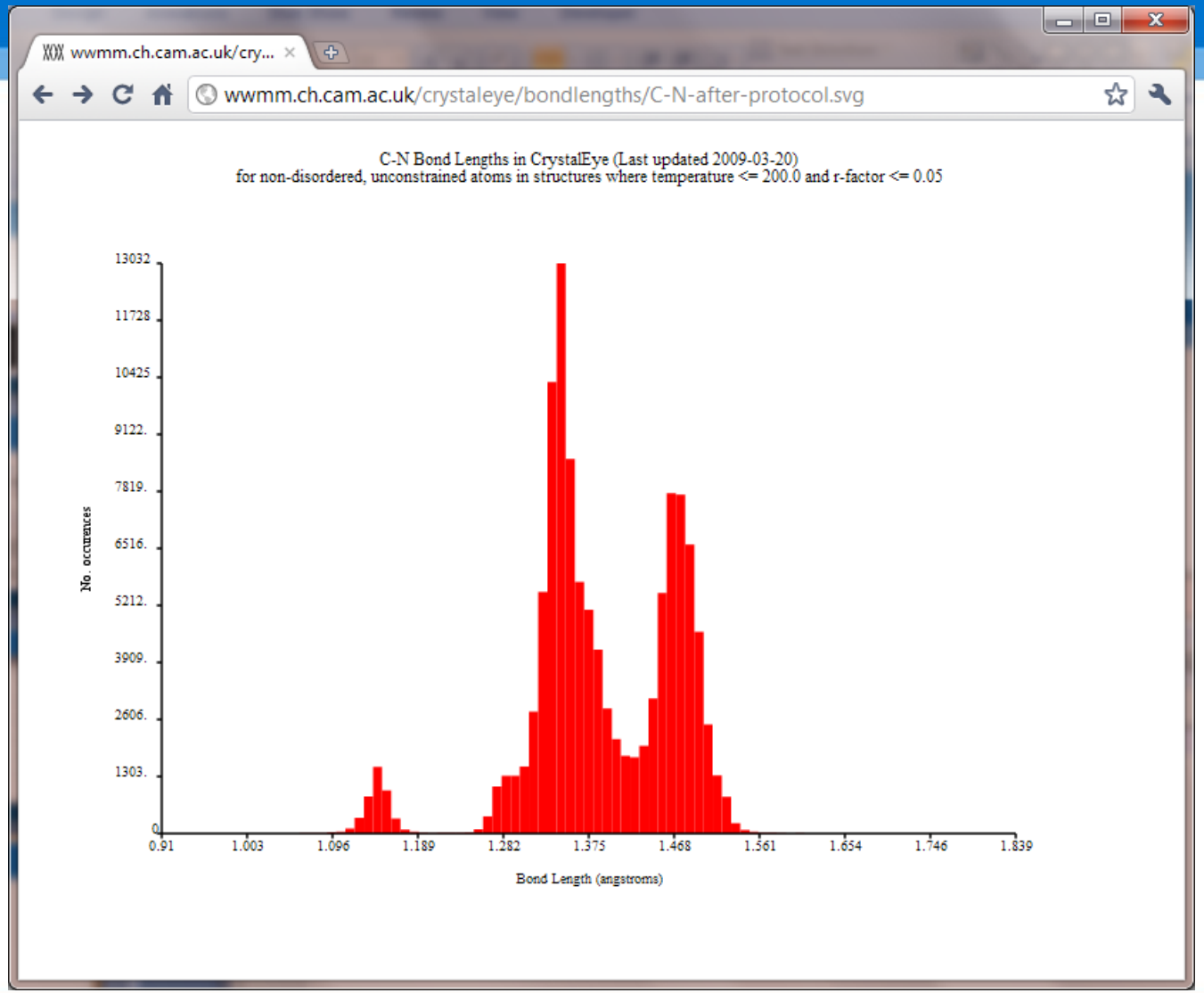


Publishers

IUCr, RSC, ACS, Nature, Chem. Soc. Japan



Especially in aggregate



Semantic Web of Data

Semantic Web

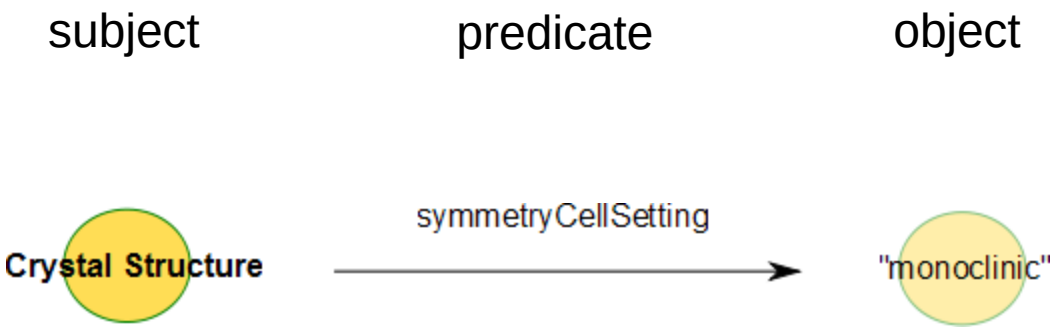
The *Semantic Web is a Web of data*. The vision of the Semantic Web is to extend principles of the Web from documents to data.

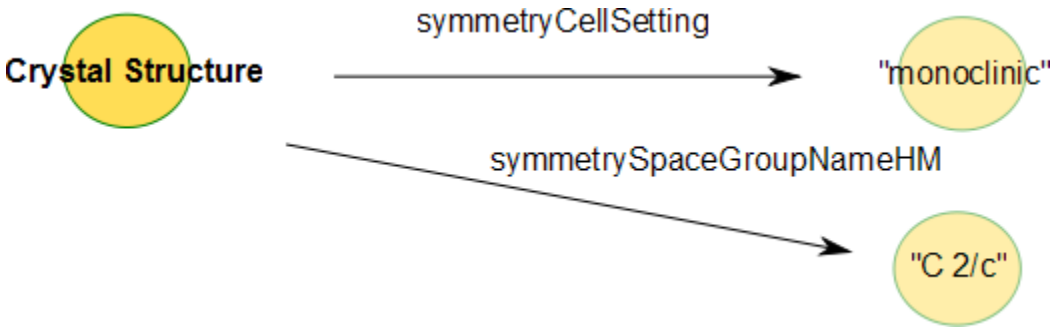
Right now data is controlled by applications, and each application keeps it to itself – the Semantic Web addresses this.

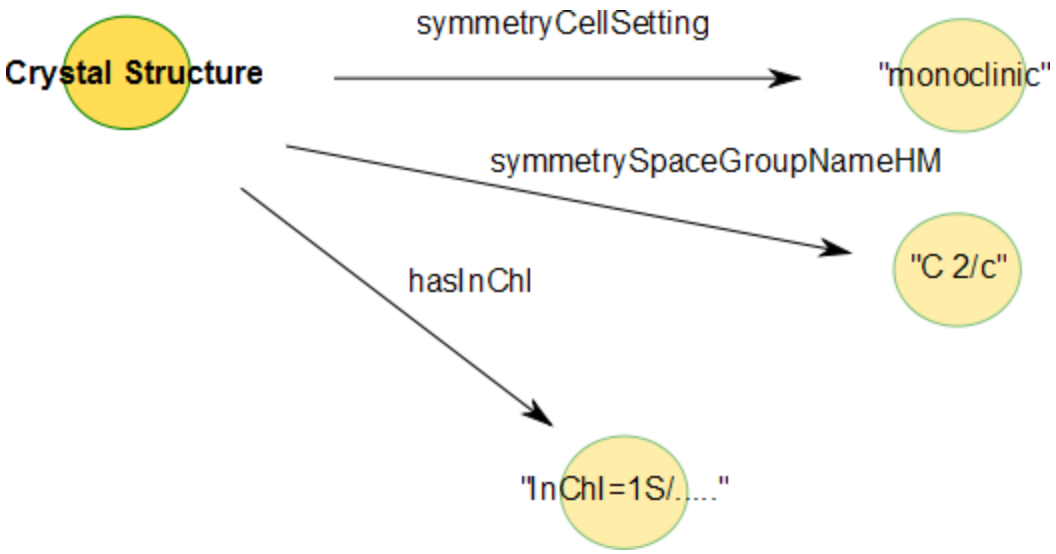
Data should be related to one another just as documents (or portions of documents) are already, and accessed using the general Web architecture.

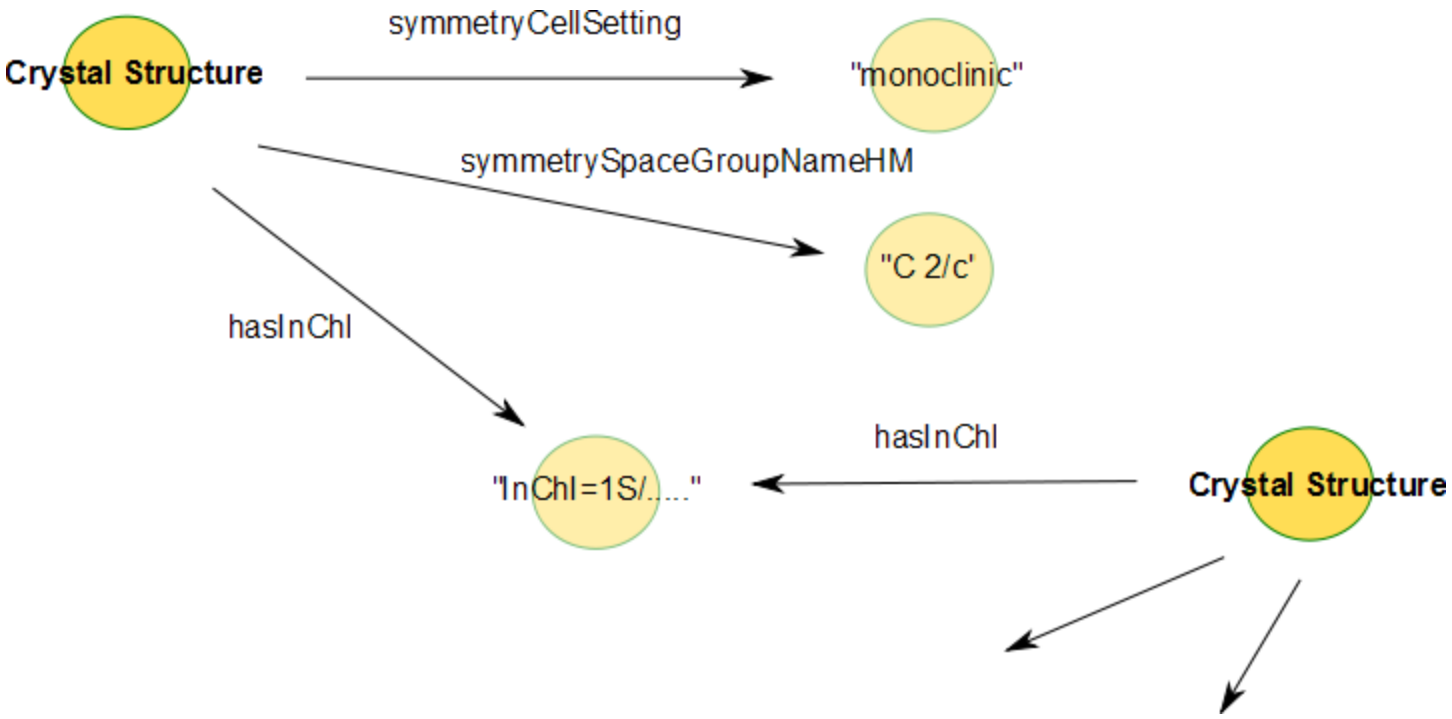
This also means creation of a common framework that allows data to be shared and reused across application, enterprise, and community boundaries, to be processed automatically by tools as well as manually, including revealing possible new relationships among pieces of data.

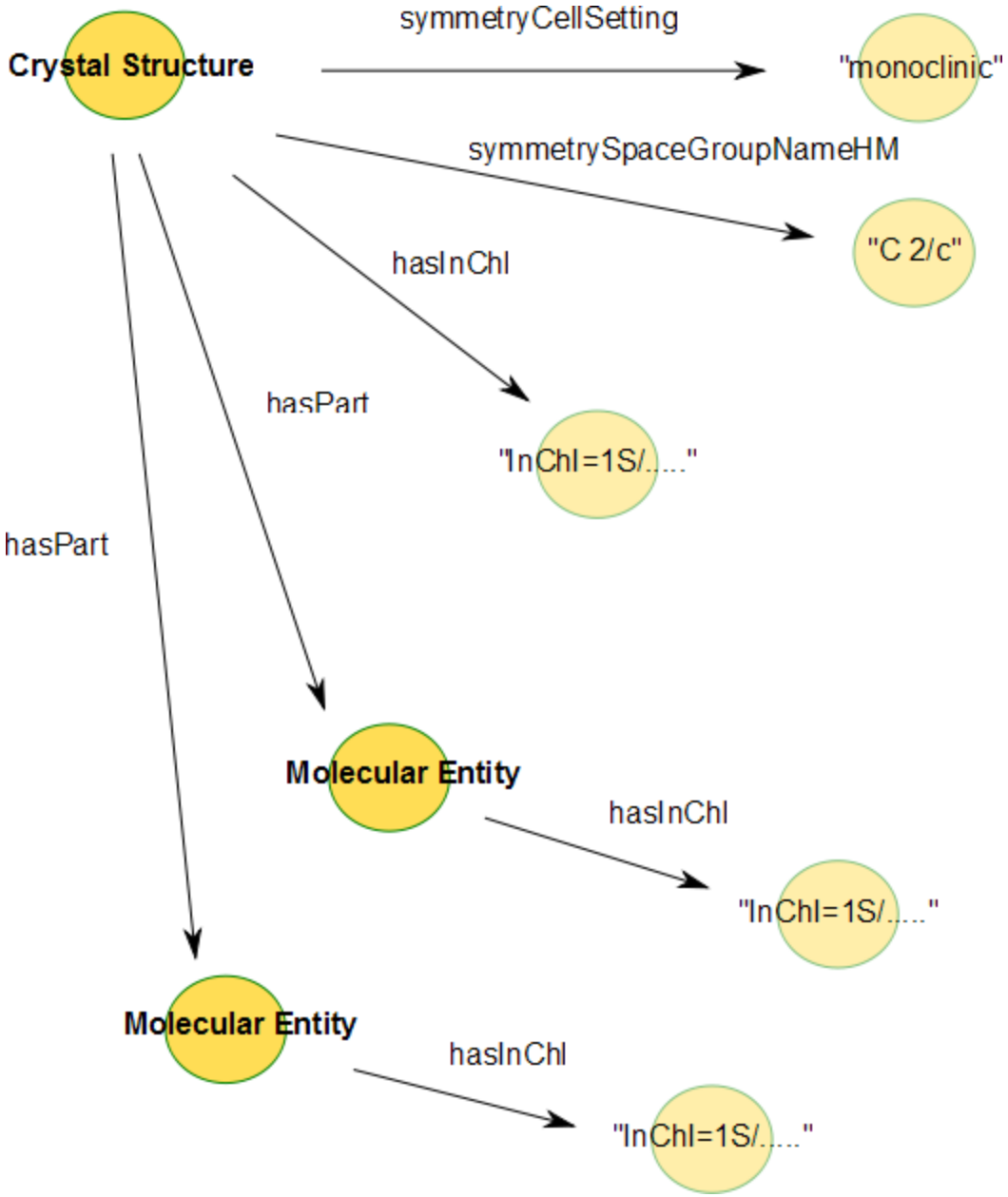
-- http://www.w3.org/2001/sw/SW-FAQ

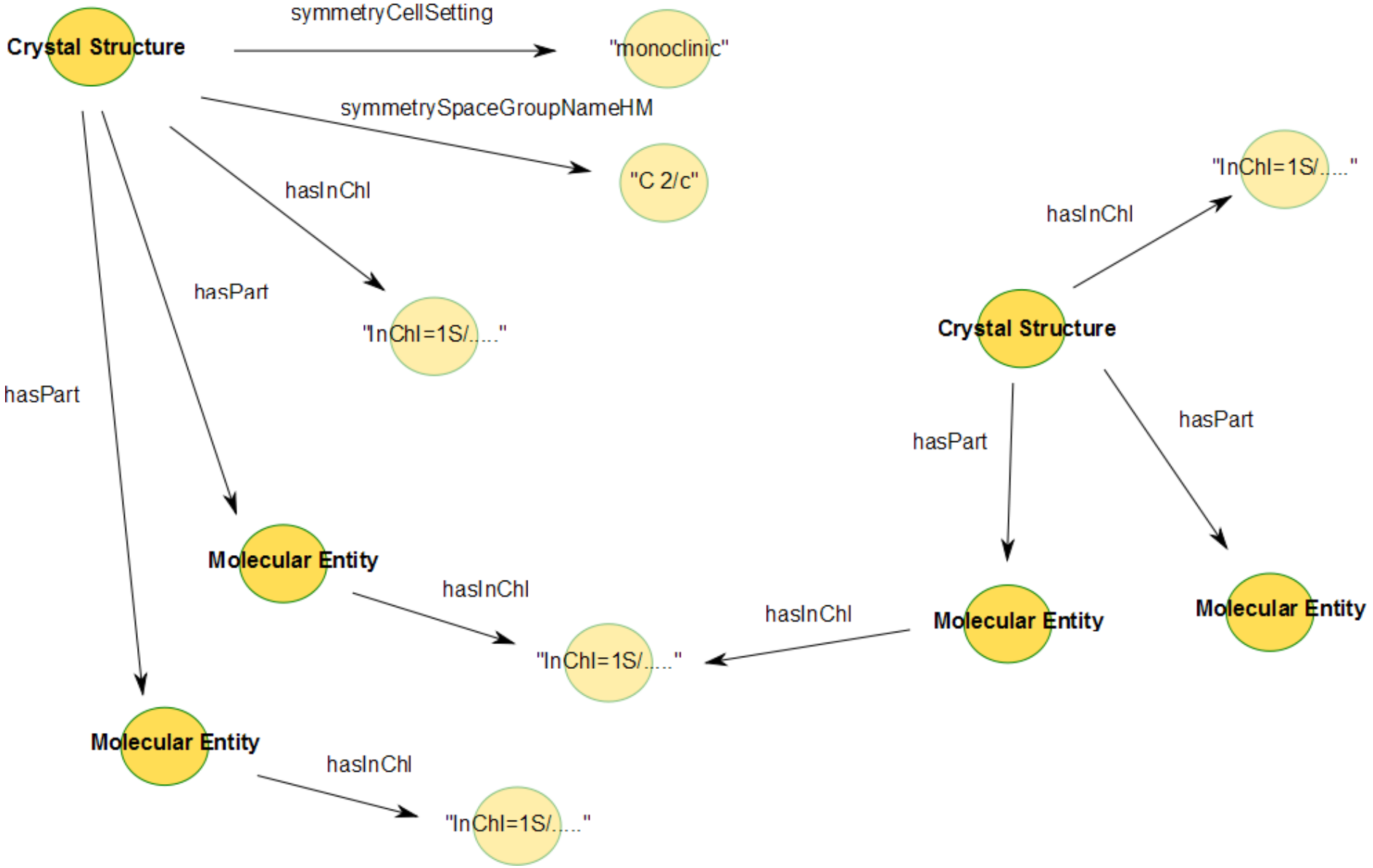

subject
predicate
object
Crystal Structure
symmetryCellSetting
"monoclinic"

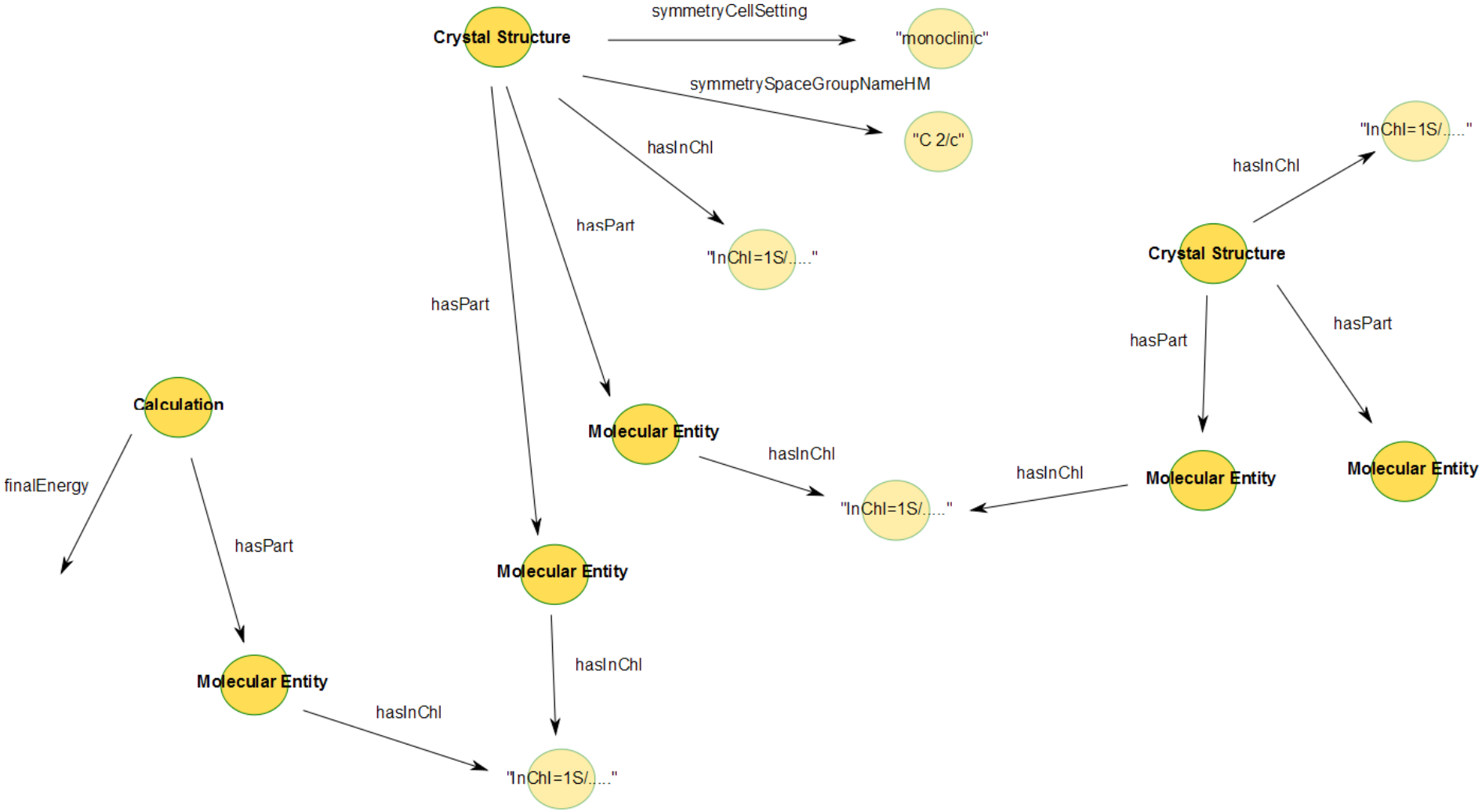

Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"

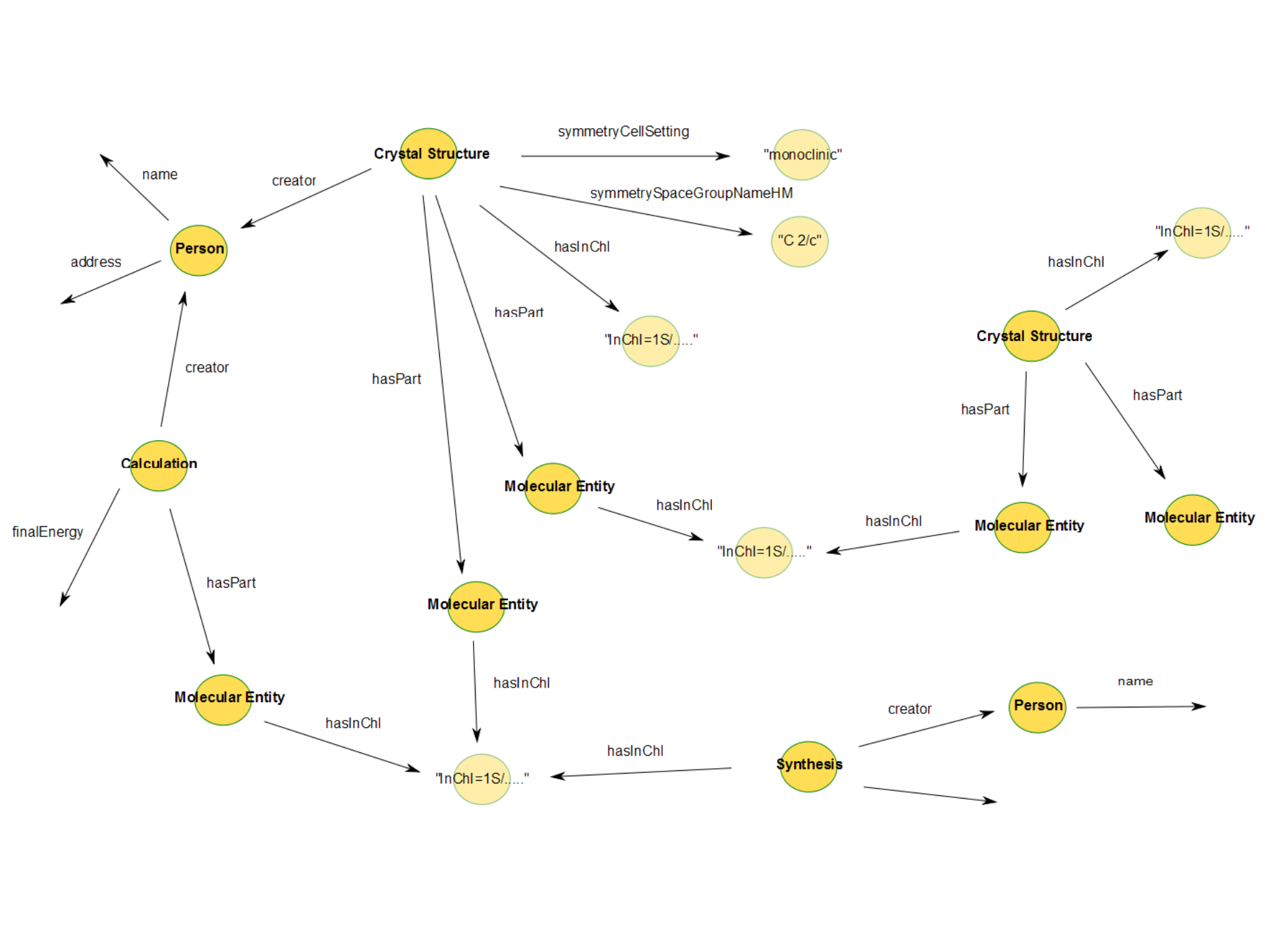

Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"
hasInChI
"InChI=1S/....."


Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c'
hasInChI
"InChI=1S/....."
hasInChI
Crystal Structure


Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Molecular Entity
hasInChI
"InChI=1S/....."


Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Molecular Entity
hasInChI
"InChI=1S/....."
Crystal Structure
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Molecular Entity


Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Calculation
finalEnergy
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Crystal Structure
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Molecular Entity


Crystal Structure
symmetryCellSetting
"monoclinic"
symmetrySpaceGroupNameHM
"C 2/c"
hasInChI
"InChI=1S/....."
hasPart
name
creator
Person
address
creator
Calculation
finalEnergy
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
hasPart
Molecular Entity
hasInChI
"InChI=1S/....."
Crystal Structure
hasInChI
"InChI=1S/....."
hasPart
hasPart
Molecular Entity
Molecular Entity
hasInChI
"InChI=1S/....."
Synthesis
creator
Person
name

Chemical Markup Language

eXtensible Markup Language (XML) vocabulary for Chemistry

Extensible:

not limited by the designer's vision

Preserve semantics: machine understandable

But the flexibility can be daunting at first...

Dictionaries and Conventions

- **Conventions:**
  - molecular
  - compcomp
  - crystallographic
  - spectra

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

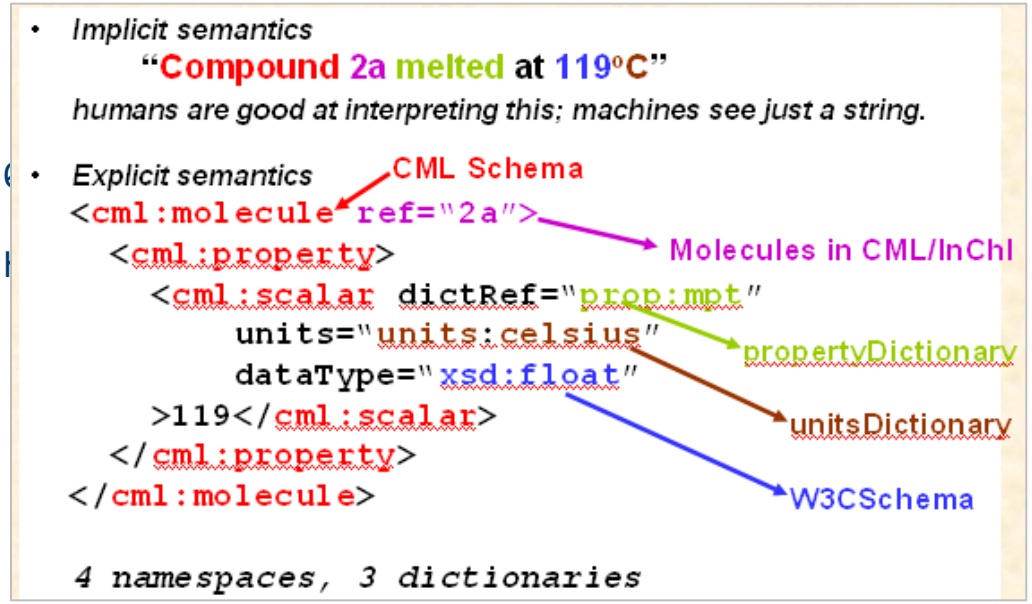
Validator Service:
http://validator.xml-cml.org/

- **Dictionaries:**
  - properties
  - units
  - unitTypes
  - data types

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>
```

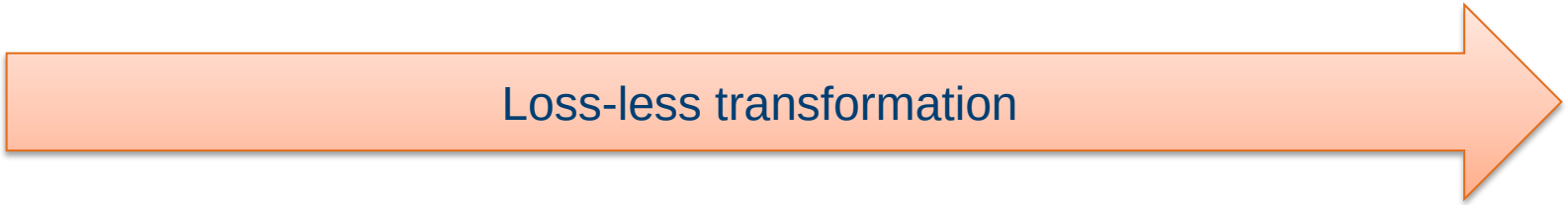


JUMBO-Converters

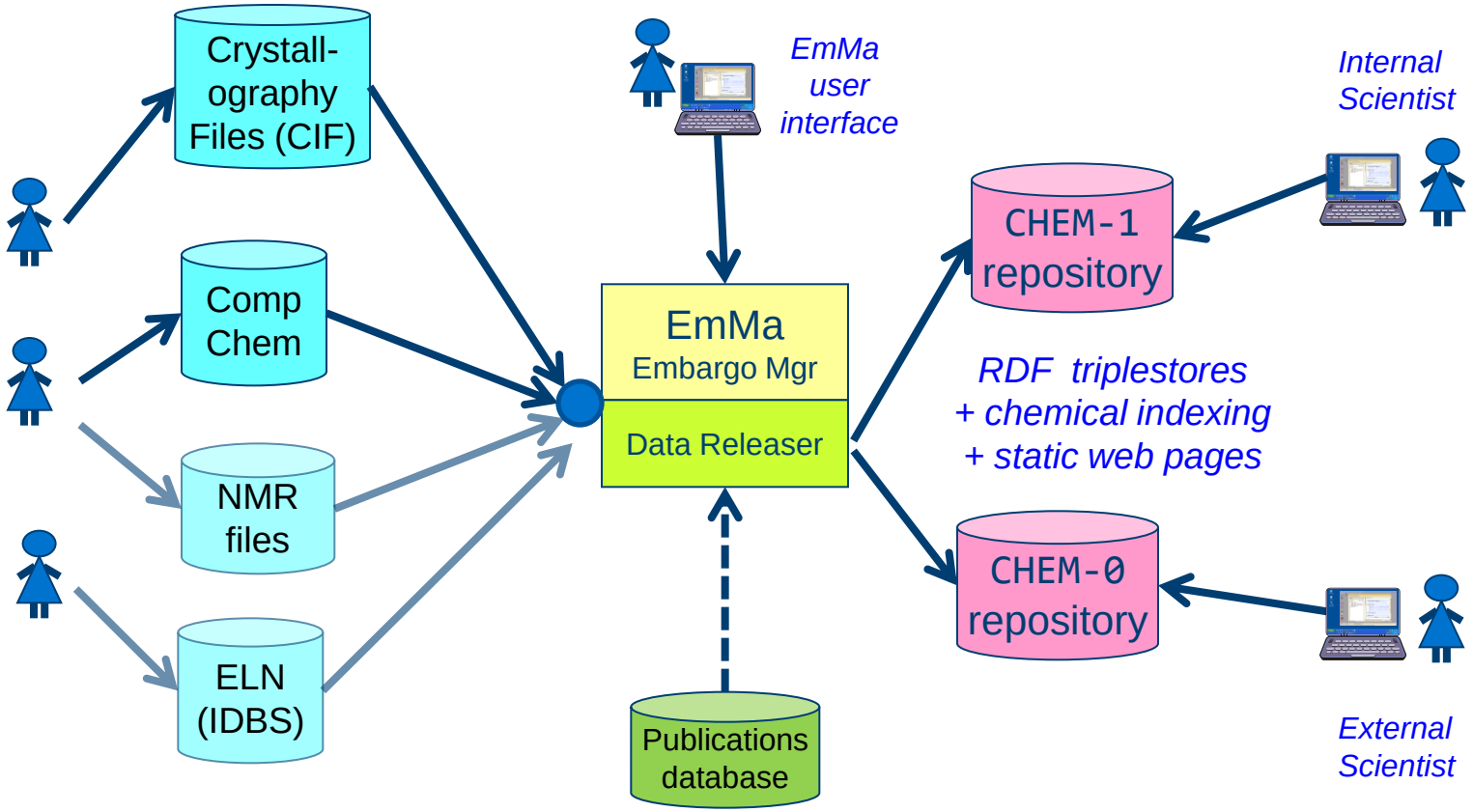
CIF
GAMESS-UK
GAMESS-US
GAUSSIAN
NWCHEM

Reflects structure of legacy input format
Properties and parameters are referenced against application specific dictionaries

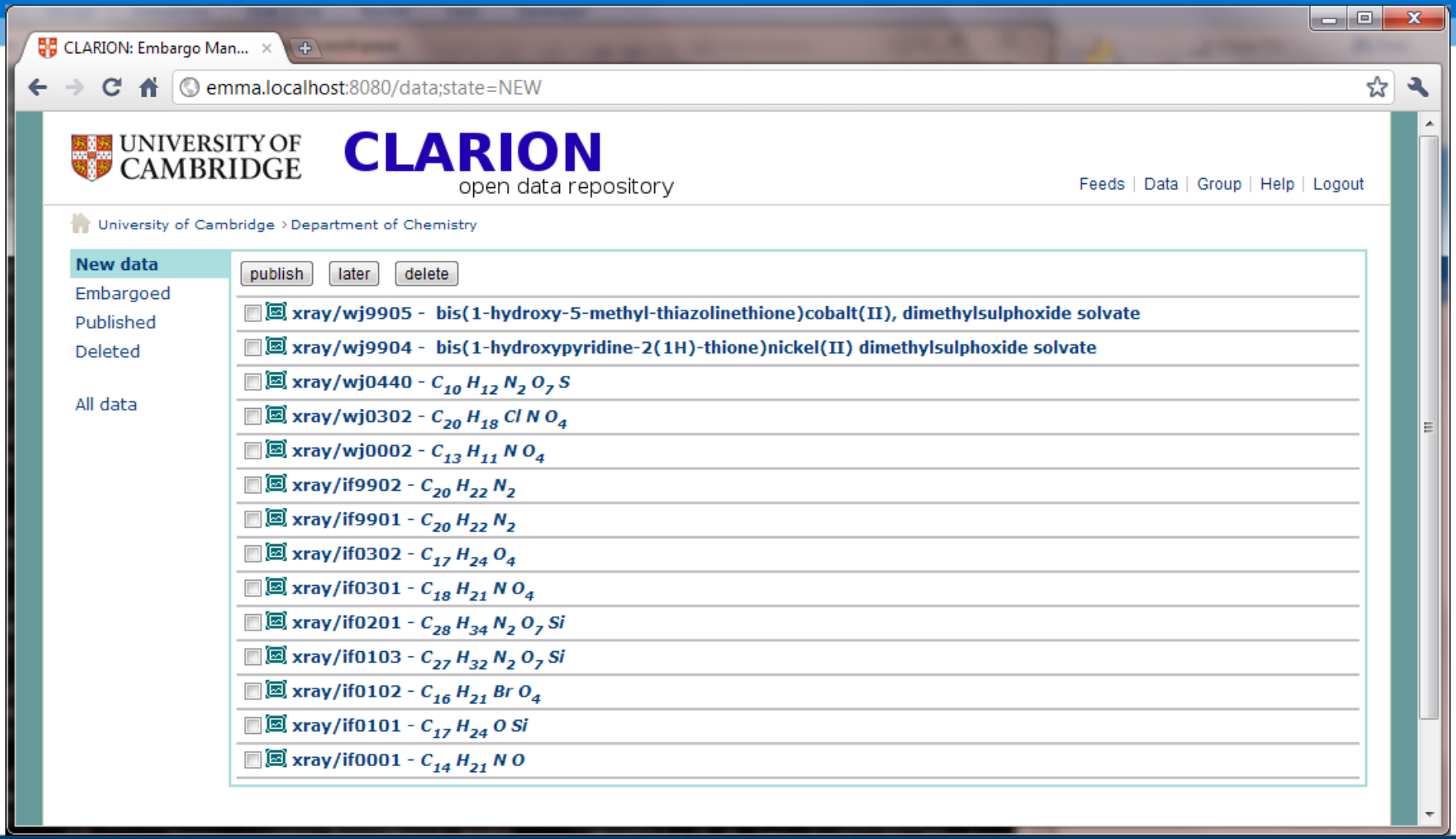
Data re-factored to fit standard data model
Where possible properties and parameters referenced against standard dictionary



CLARION


Crystall-ography Files (CIF)
Comp Chem
NMR files
ELN (IDBS)
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

CLARION: Embargo Manager



CLARION: Embargo Manager

CLARION: Embargo Man...

emma.localhost:8080/data/xray/wj9904

Department
Public
Deleted

All data

Title:
bis(1-hydroxypyridine-2(1H)-thione)nickel(II) dimethylsulphoxide solvate

Crystal system:
orthorhombic

Chemical formula sum:
C_{12} H_{20} N_2 Ni O_4 S_6

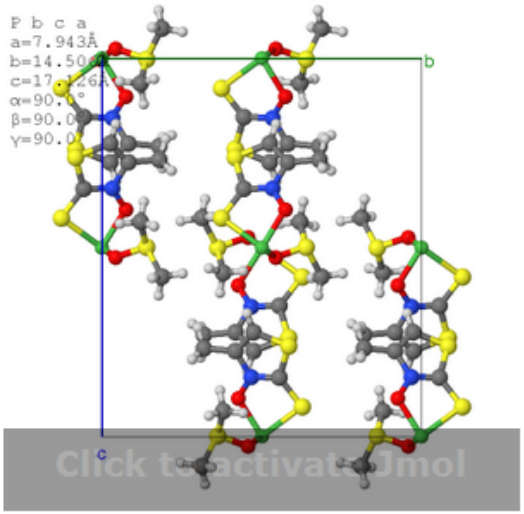
Compound class:
organometallic

Cell parameters:
a=7.943 b=14.506 c=17.126
α=90.0 β=90.0 γ=90.0

Chemists:
A.D.Bond; W.Jones

Solvent:

Notes:

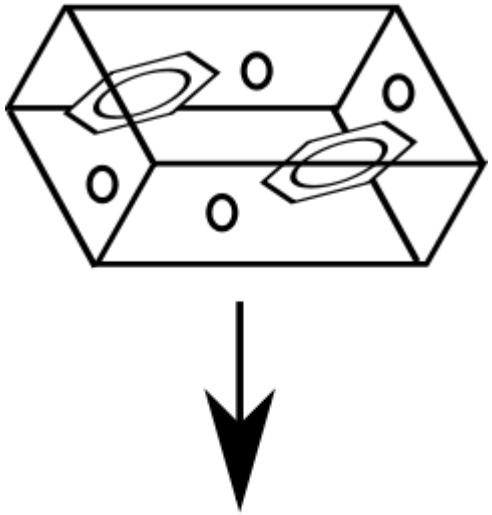


Data publication

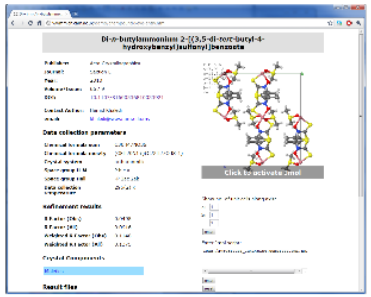
Publish to Department:
- Not at this time
- ASAP
- on dd-mm-yyyy

Publish to Public:
- Not at this time
- ASAP

Chempound



linked open data:
the chemical semantic web



Chem#

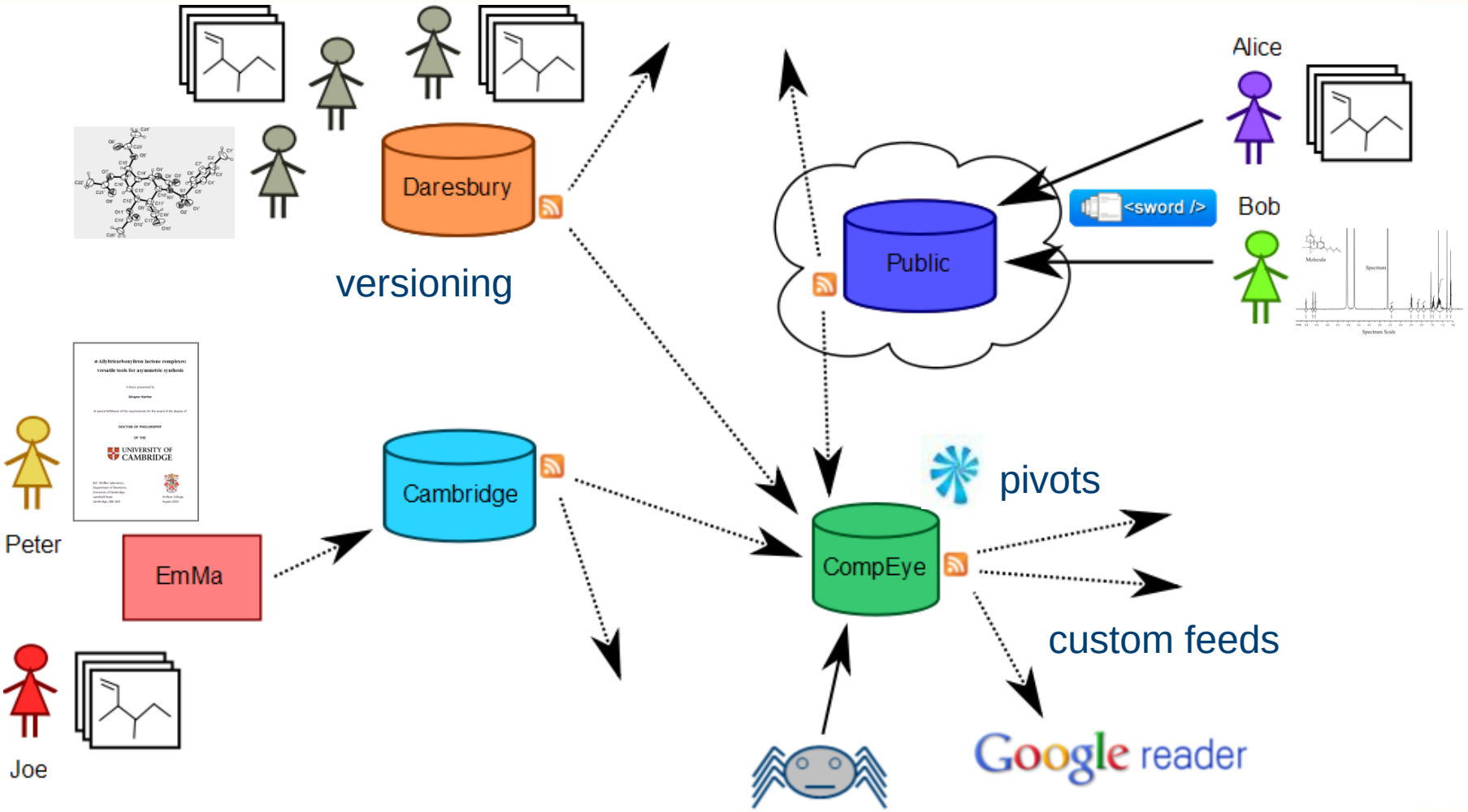
W3C SPARQL

CML RDF

Atom

Chempound stores
legacy and semantic files
indexed using RDF

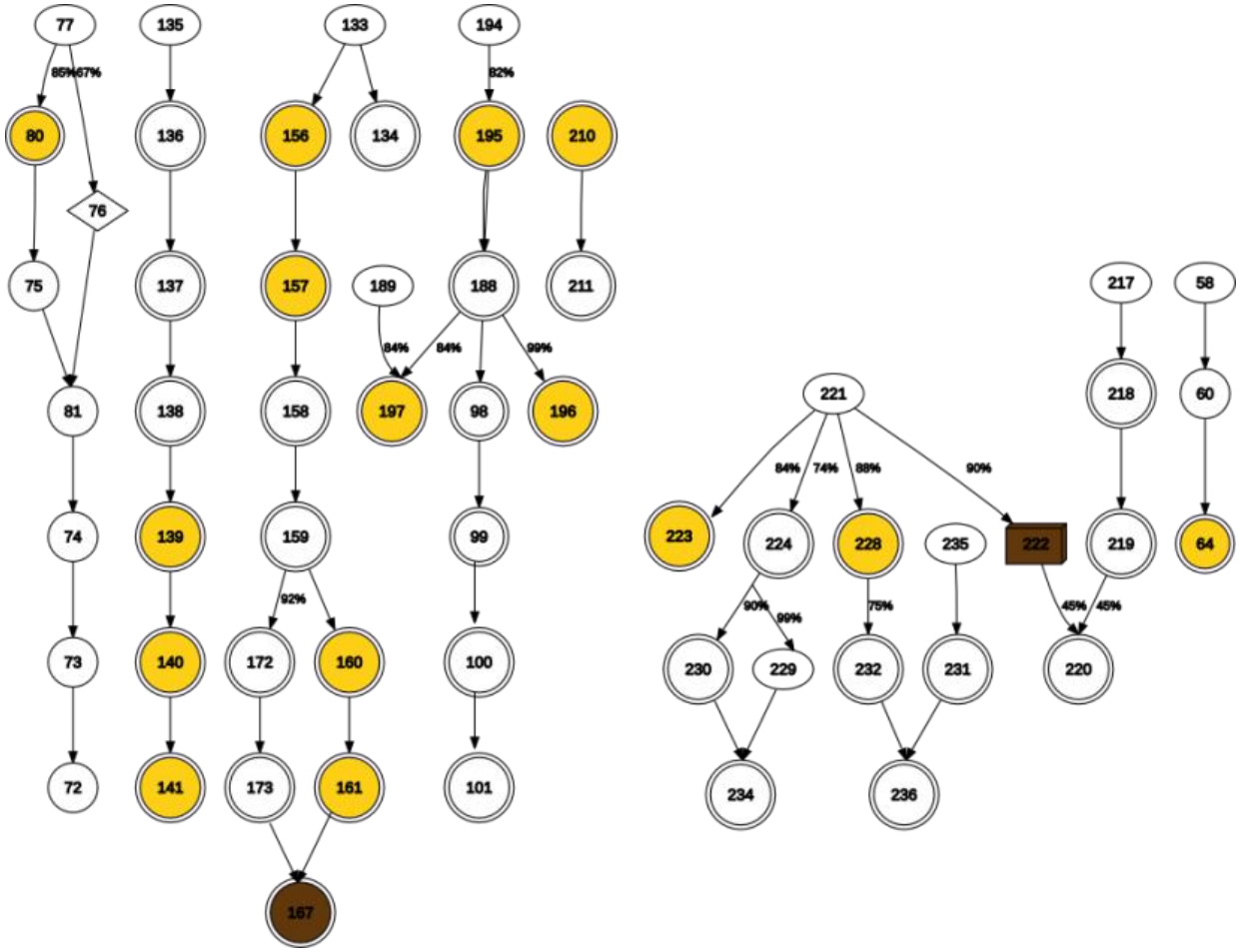
The future?

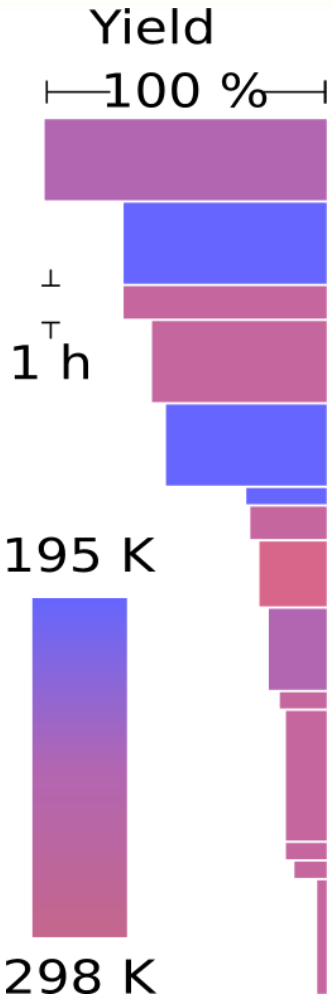

Daresbury
versioning
Alice
Bob
<sword />
Public
Peter
EmMa
Cambridge
CompEye
pivots
custom feeds
Google reader
Joe

Data is exciting!

and we can’t predict what people will do with it

New ways of visualising chemistry





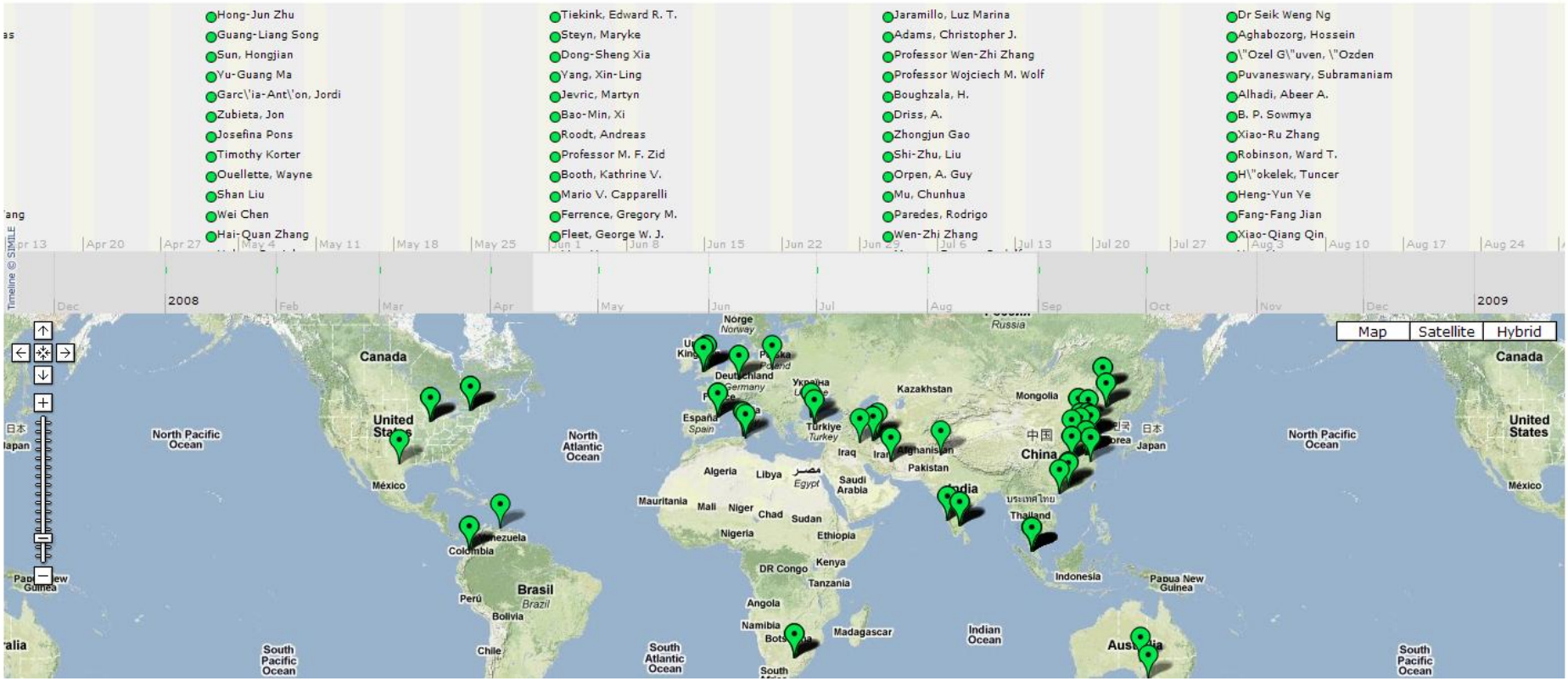
Map Timeline Mashup

benosteen.com/timemap/index

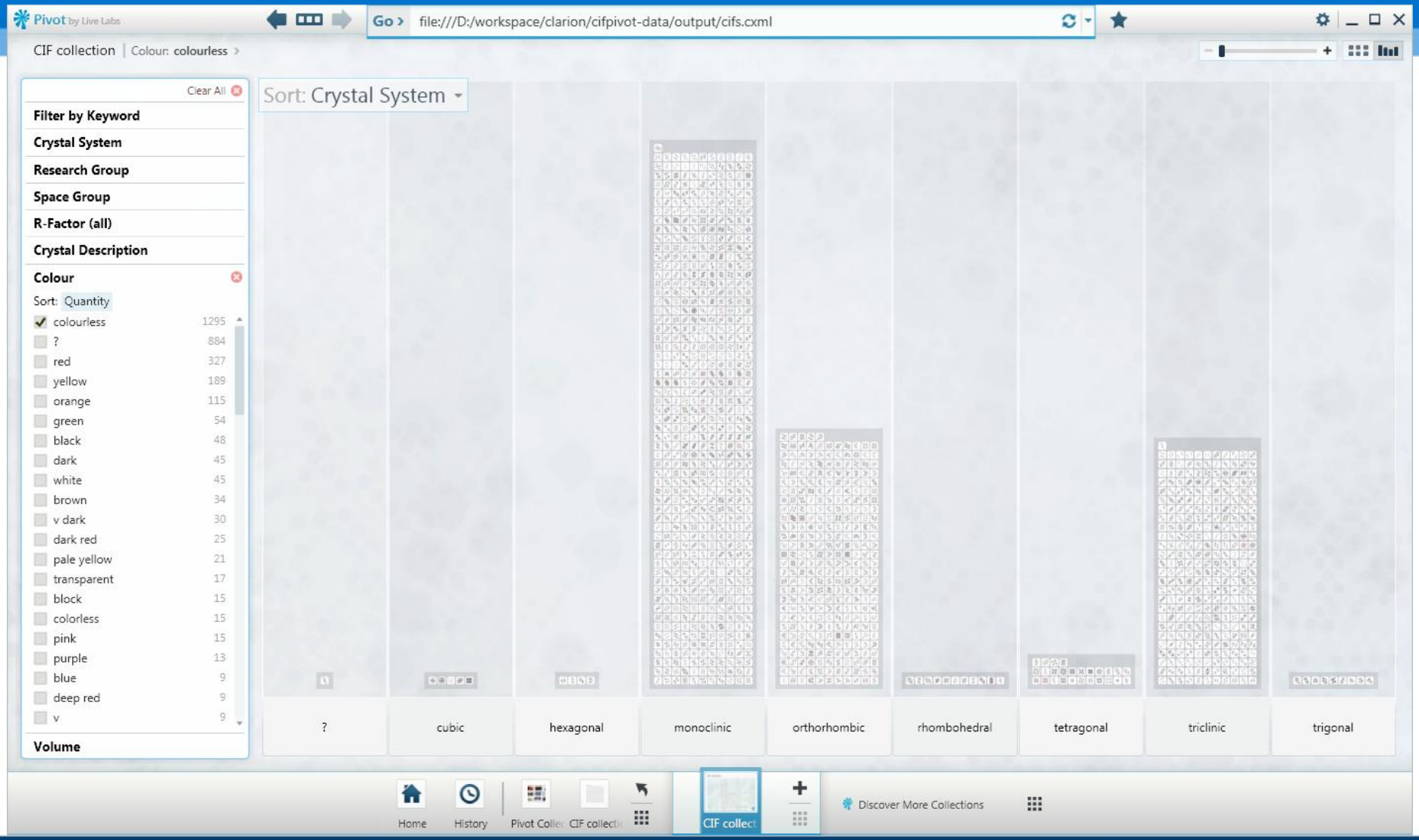
IUCr Crystal publication data

Authorship - as an interactive display based on the date of publication and the location of the author.

The IUCr articles have rich and open metadata, allowing for repurposing and reuse. Here, is a visualisation of where each author of a paper is affiliated, running against a timeline for when the articles were published. Scroll the timeline from the right-hand side to the left to go forward in time. The map below is a google map and can be zoomed and panned as desired.



Visualisation



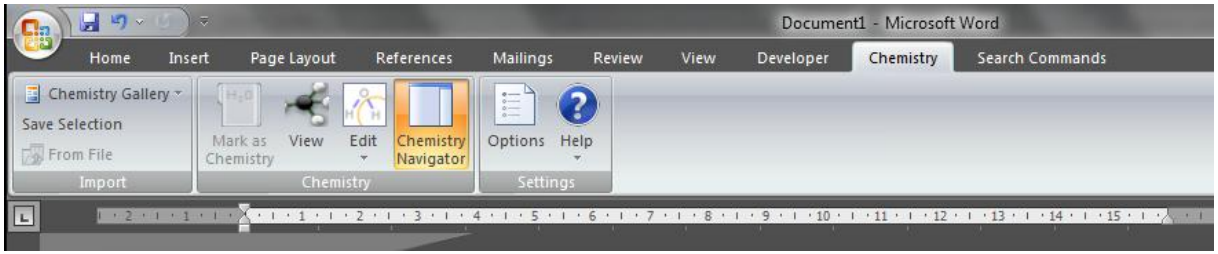
Summary

- Data drives science, but masses of scientific data is currently lost
- Publication needs to be easy – fit into scientist's existing workflows
- Archiving is not enough – must plan for reuse
- Semantic, linked open data is the solution
- Existing standards where possible, but need domain knowledge
- **Data publication will grow.**

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

Thank You

Chemistry Add-in for Word



directory

file

A lookup for each of the items referenced in the document (e.g. images, sounds, footers, styles and embedded OLE objects)

package

- _rels
  - *document.xml.rels*
- customXml
  - *item1.xml*
  - *itemProps1.xml*
- docProps
- word
  - _rels
  - media
    - *image1.png*
  - theme
  - *document.xml*
- *[Content_Types].xml*

Data linked to a content control in the document

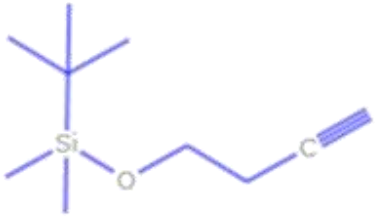
A static image - this is displayed in the document if the content control is a picture.

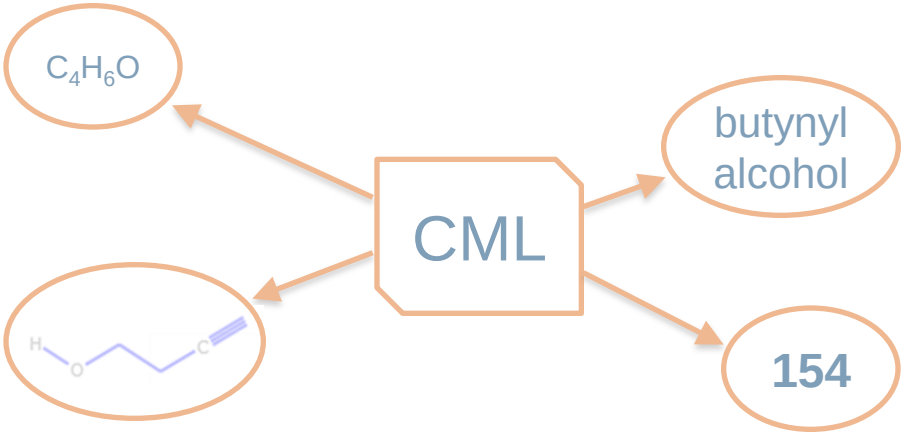
The text of the document. This also contans links to other objects retrieved via lookup.

Contains MIME type information for parts of the package.

Chemistry Add-in for Word

Preparation of 4 – (tert – butyl – dimethyl – silanyloxy)butyne **155**





To a solution of butynyl alcohol **154** in DCM (250 ml) at 0°C was added TBDMSCl (1.02 eq., 0.24 mol, 36.18 g) and DMAP(MW=122.17, 200 mg, 1.6 mmol). Et_3N (MW=101.19, d=0.726, bp=89°C, 0.25 mol, 25.29 g = 34.85 ml) was subsequently added via syringe and the reaction mixture stirred at 0°C for 2 h, after which it was warmed to rt and stirred for a further 6 h, until completion as judged by tlc. The mixture was poured onto saturated NH_4Cl solution (800 ml) and extracted with Et_2O (3 x 300 ml). The combined organic extracts were washed with brine (500 ml) and dried ($MgSO_4$), filtered, and concentrated *in vacuo*, yielding an oil which was purified by filtration through a pad of SiO_2 (eluent PE:Et_2O 10:1) giving compound **155** (36.10 g, 19.6 mmol, 83%) as colourless oil.