Semantic Chemical Publishing

Nick Day*, Peter Corbett, Peter Murray-Rust
Unilever Centre for Molecular Informatics, University of Cambridge, UK.

March 27th, 2007

- All software Open Source

* ned24@cam.ac.uk

Overview

- What is ‘semantic chemistry’ and markup?

- **OSCAR3** – robotic analysis of chemistry in free text
  - recognition of chemical names
  - name-2-structure
  - chemical verbs, adjectives and reaction names
  - terminologies (e.g. techniques)
  - RSC **Project Prospect** …

- **CrystalEye** – creating semantic chemistry from crystallography:
  - High-throughput robotic harvesting
  - Re-use using CIF2CML
  - Dissemination through CMLRSS

The Semantic Web

“People keep asking what Web 3.0 is. I think maybe when you've got an overlay of scalable vector graphics […] on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource.”

- Tim Berners-Lee, A 'more revolutionary' Web (2006)

…Let’s change the vision to chemistry…

The Chemical Semantic Web

“…when you've got an overlay of

…..

- on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource.”

[our adaptations]

… what are chemical semantics and CML?…

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

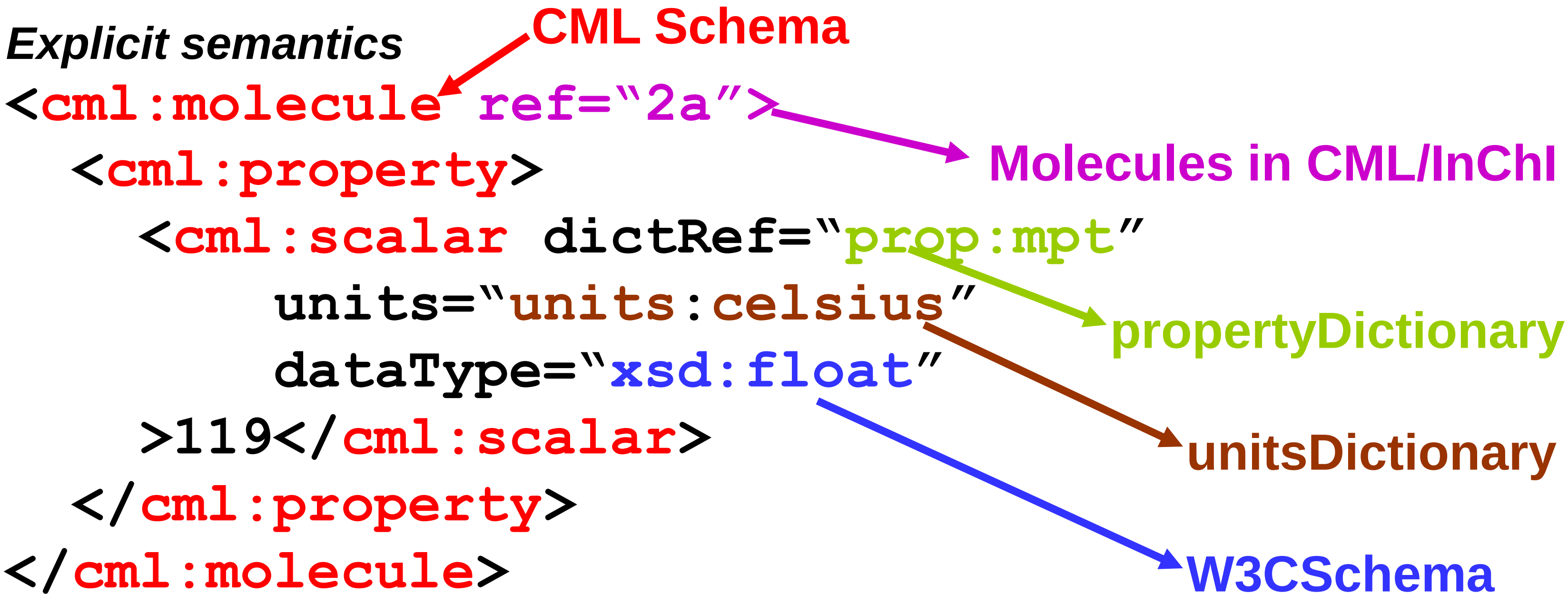
Implicit and explicit semantics

- *Implicit semantics*

  “Compound 2a melted at 119ºC”

  humans are good at interpreting this; machines see just a string.

- *Explicit semantics*



```
<cml:molecule ref=“2a”>
  <cml:property>
    <cml:scalar dictRef=“prop:mpt”
        units=“units:celsius”
        dataType=“xsd:float”
    >119</cml:scalar>
  </cml:property>
</cml:molecule>
```

4 namespaces, 3 dictionaries

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

UCC’s approach to creating Semantic Chemistry

- Authoring tools for theses and collaboration with publishers
- XML-ization (through FoX) of Comp. Chem. codes (MOPAC, CASTEP, SIESTA, GULP, ABINIT, DL_POLY GAMESS…)
- Capturing/conversion of CML data at source (SPECTRa)
- Rich clients (Bioclipse)
- Legacy Conversion (OpenBabel, CDK, JUMBO…)
- Intelligent Ontologies (Golem)

(today we will cover the following…)

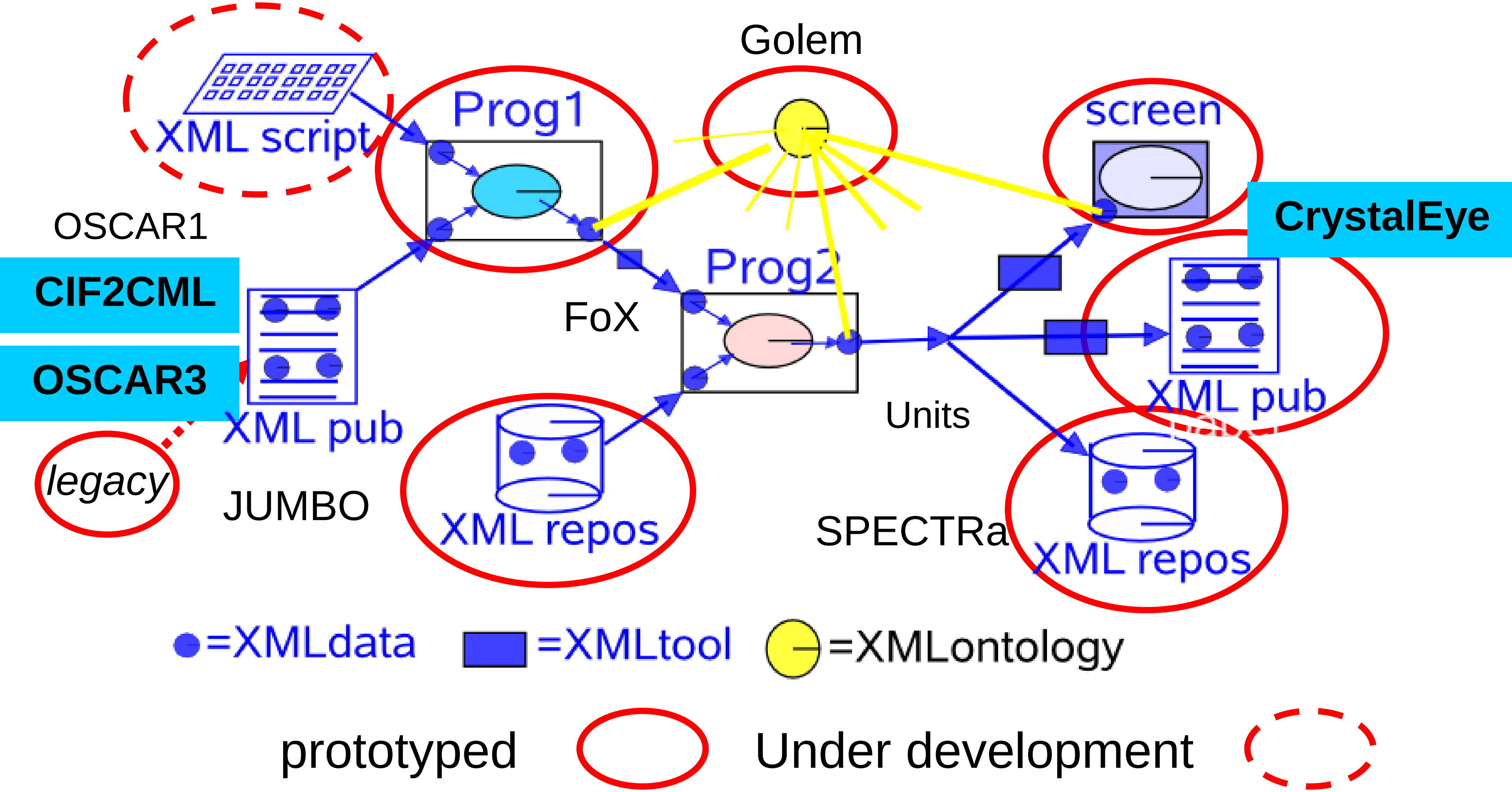
- **Chemical Linguistics and text-mining (OSCAR3)**
- **Legacy Conversion – CIF**

Many of these semantic chemical components are now deployed or prototyped…

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

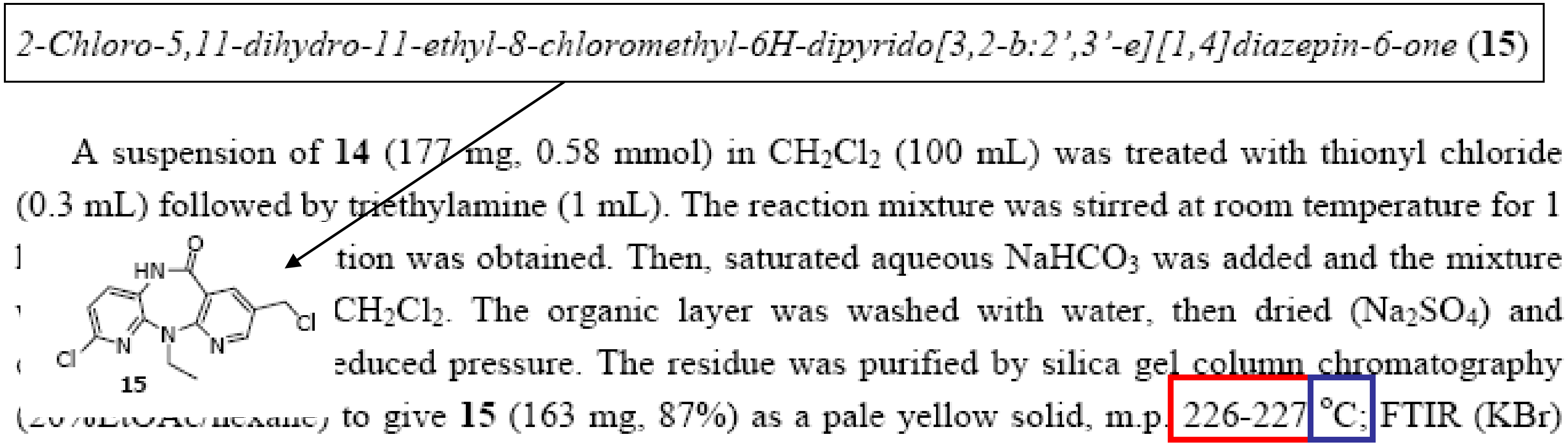
Chemical semantic framework at UCC



UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

“OSCAR”


O
HN
Cl
N
N
N
Cl

15

2-Chloro-5,11-dihydro-11-ethyl-8-chloromethyl-6H-dipyrido[3,2-b:2',3'-e][1,4]diazepin-6-one (**15**)

A suspension of **14** (177 mg, 0.58 mmol) in CH_2Cl_2 (100 mL) was treated with thionyl chloride (0.3 mL) followed by triethylamine (1 mL). The reaction mixture was stirred at room temperature for 1 l tion was obtained. Then, saturated aqueous $NaHCO_3$ was added and the mixture CH_2Cl_2. The organic layer was washed with water, then dried (Na_2SO_4) and educed pressure. The residue was purified by silica gel column chromatography (20%EtOAc/hexane) to give **15** (163 mg, 87%) as a pale yellow solid, m.p. 226-227 °C; FTIR (KBr)

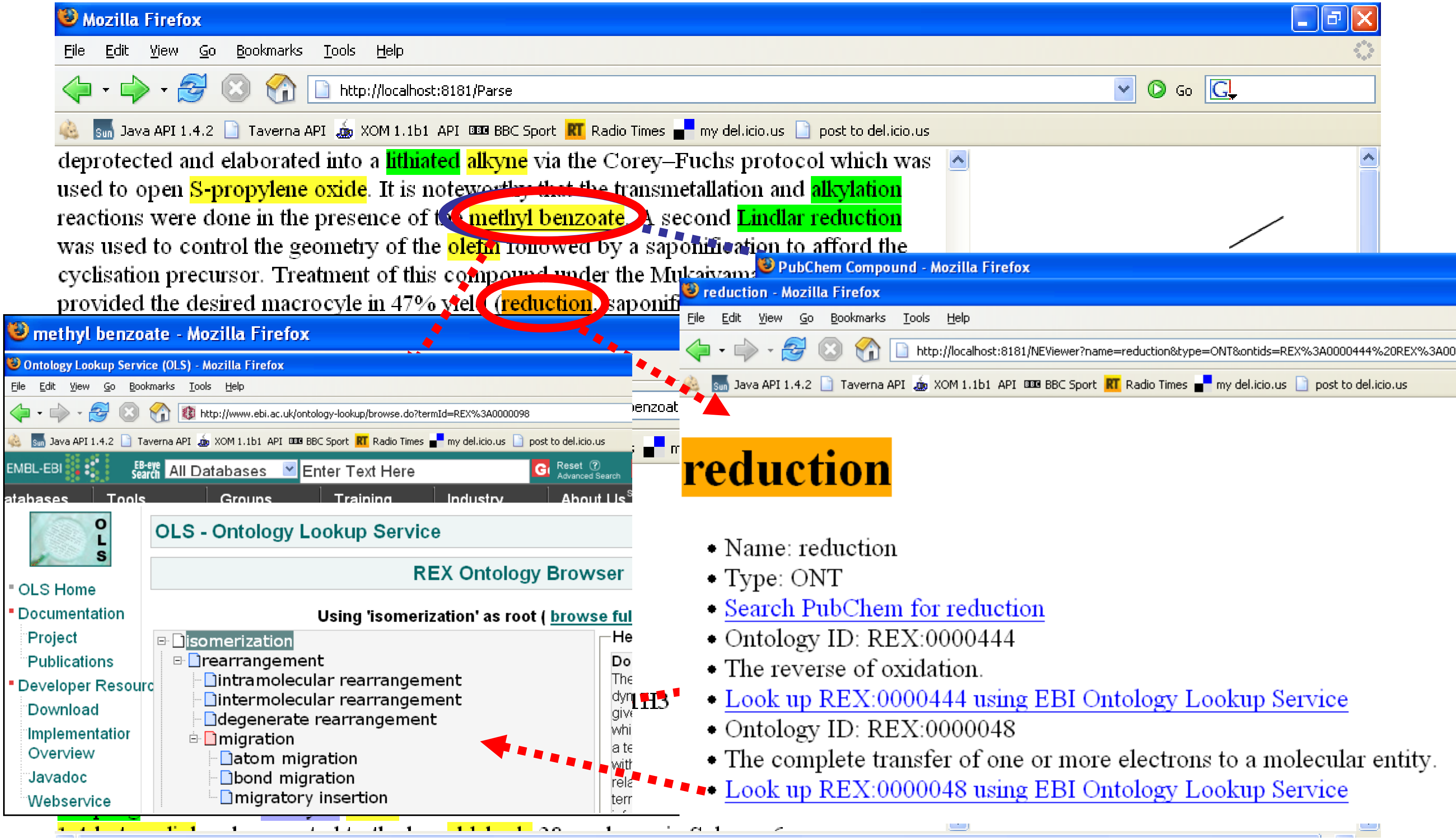
- Recognition of chemical entities.
- Name2structure, chemical diagrams, canonical identifiers
- Chemical heuristics to parse article full-text
- Links to ontologies and molecular databases.
- Open source
- High-throughput – 500, 000 PubMed abstracts parsed
- Substructure and similarity search on corpora

OSCAR1 + CheckCML (2003, 2004, 2005, 2006) Student projects supported by RSC
SciBorg (2005-2009)   EPSRC project (Computer Lab, Chemistry, Cambridge)
OSCAR3 (Peter Corbett)

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

OSCAR3 Concepts - Example


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8181/Parse
Go
Java API 1.4.2 Taverna API XOM 1.1b1 API BBC Sport Radio Times my del.icio.us post to del.icio.us
methyl benzoate - Mozilla Firefox
Ontology Lookup Service (OLS) - Mozilla Firefox
http://www.ebi.ac.uk/ontology-lookup/browse.do?termId=REX%3A0000098
EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Reset
Advanced Search
atabases Tools Groups Training Industry About Us
PubChem Compound - Mozilla Firefox
reduction - Mozilla Firefox
http://localhost:8181/NEViewer?name=reduction&type=ONT&ontids=REX%3A0000444%20REX%3A00

deprotected and elaborated into a lithiated alkyne via the Corey–Fuchs protocol which was used to open S-propylene oxide. It is noteworthy that the transmetallation and alkylation reactions were done in the presence of the methyl benzoate. A second Lindlar reduction was used to control the geometry of the olefin followed by a saponification to afford the cyclisation precursor. Treatment of this compound under the Mukaiyama provided the desired macrocyle in 47% yield (reduction, saponifi

OLS - Ontology Lookup Service

REX Ontology Browser

Using 'isomerization' as root (browse ful

- OLS Home
- Documentation
  - Project
  - Publications
- Developer Resourc
  - Download
  - Implementatior Overview
  - Javadoc
  - Webservice

- isomerization
  - rearrangement
    - intramolecular rearrangement
    - intermolecular rearrangement
    - degenerate rearrangement
    - migration
      - atom migration
      - bond migration
      - migratory insertion

reduction

- Name: reduction
- Type: ONT
- Search PubChem for reduction
- Ontology ID: REX:0000444
- The reverse of oxidation.
- Look up REX:0000444 using EBI Ontology Lookup Service
- Ontology ID: REX:0000048
- The complete transfer of one or more electrons to a molecular entity.
- Look up REX:0000048 using EBI Ontology Lookup Service

UNIVERSITY OF CAMBRIDGE

All markup is **automatic**

Unilever Cambridge
Centre for Molecular Informatics

…how can this be used for publishing?...

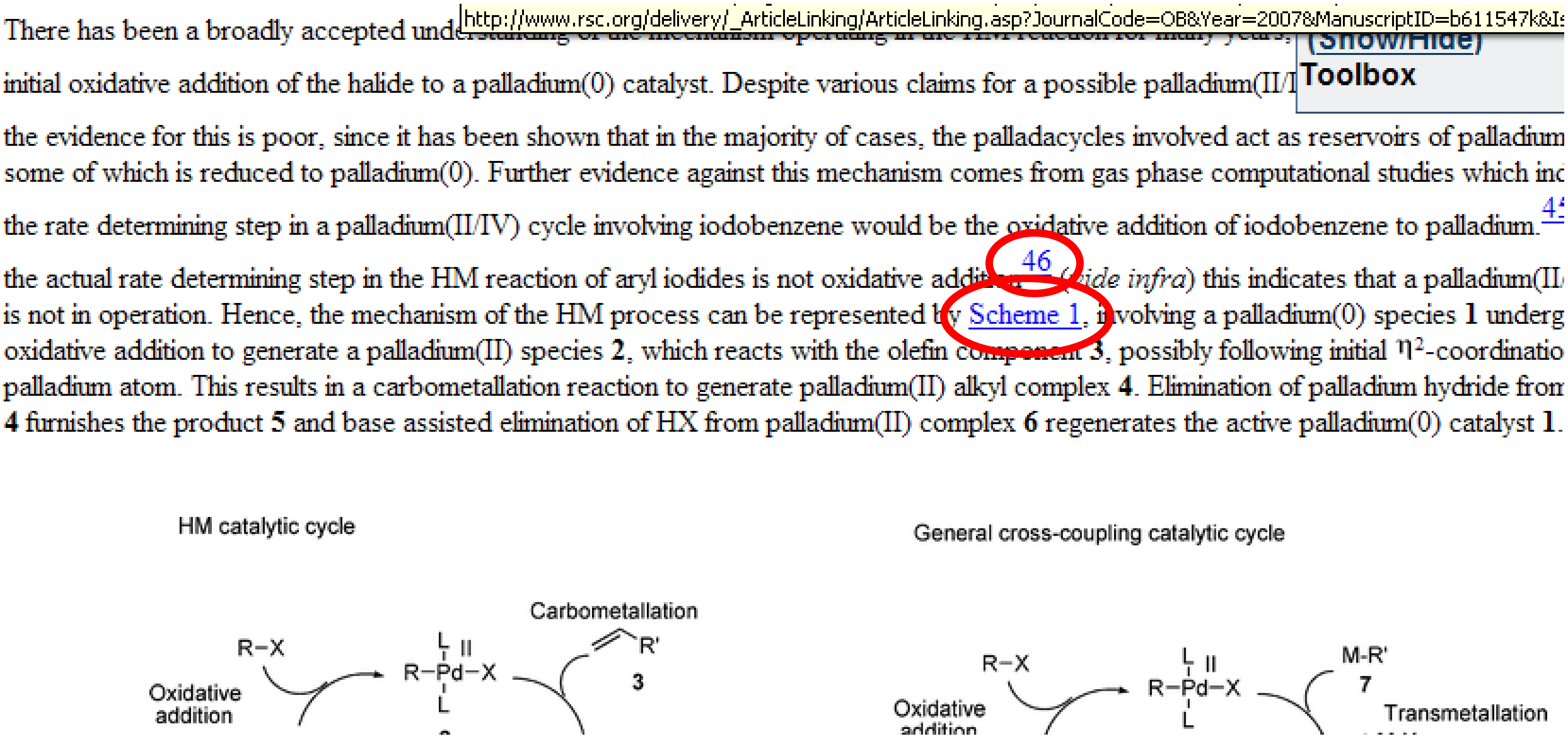
UCC and RSC have been collaborating on transferring this technology to journal articles…

Project Prospect (2007) adds semantics…

Project Prospect (RSC)

Typical HTML paper

Semantics confined to hyperlinks



There has been a broadly accepted und

http://www.rsc.org/delivery/_ArticleLinking/ArticleLinking.asp?JournalCode=OB&Year=2007&ManuscriptID=b611547k&Is

(Show/Hide)

Toolbox

initial oxidative addition of the halide to a palladium(0) catalyst. Despite various claims for a possible palladium(II/I
the evidence for this is poor, since it has been shown that in the majority of cases, the palladacycles involved act as reservoirs of palladium
some of which is reduced to palladium(0). Further evidence against this mechanism comes from gas phase computational studies which inc
the rate determining step in a palladium(II/IV) cycle involving iodobenzene would be the oxidative addition of iodobenzene to palladium.[4
the actual rate determining step in the HM reaction of aryl iodides is not oxidative addition[46] (vide infra) this indicates that a palladium(II/
is not in operation. Hence, the mechanism of the HM process can be represented by Scheme 1, involving a palladium(0) species **1** underg
oxidative addition to generate a palladium(II) species **2**, which reacts with the olefin component **3**, possibly following initial η^2-coordinatio
palladium atom. This results in a carbometallation reaction to generate palladium(II) alkyl complex **4**. Elimination of palladium hydride fron
4 furnishes the product **5** and base assisted elimination of HX from palladium(II) complex **6** regenerates the active palladium(0) catalyst **1**.

Prospect adds more semantics…

UNIVERSITY OF CAMBRIDGE

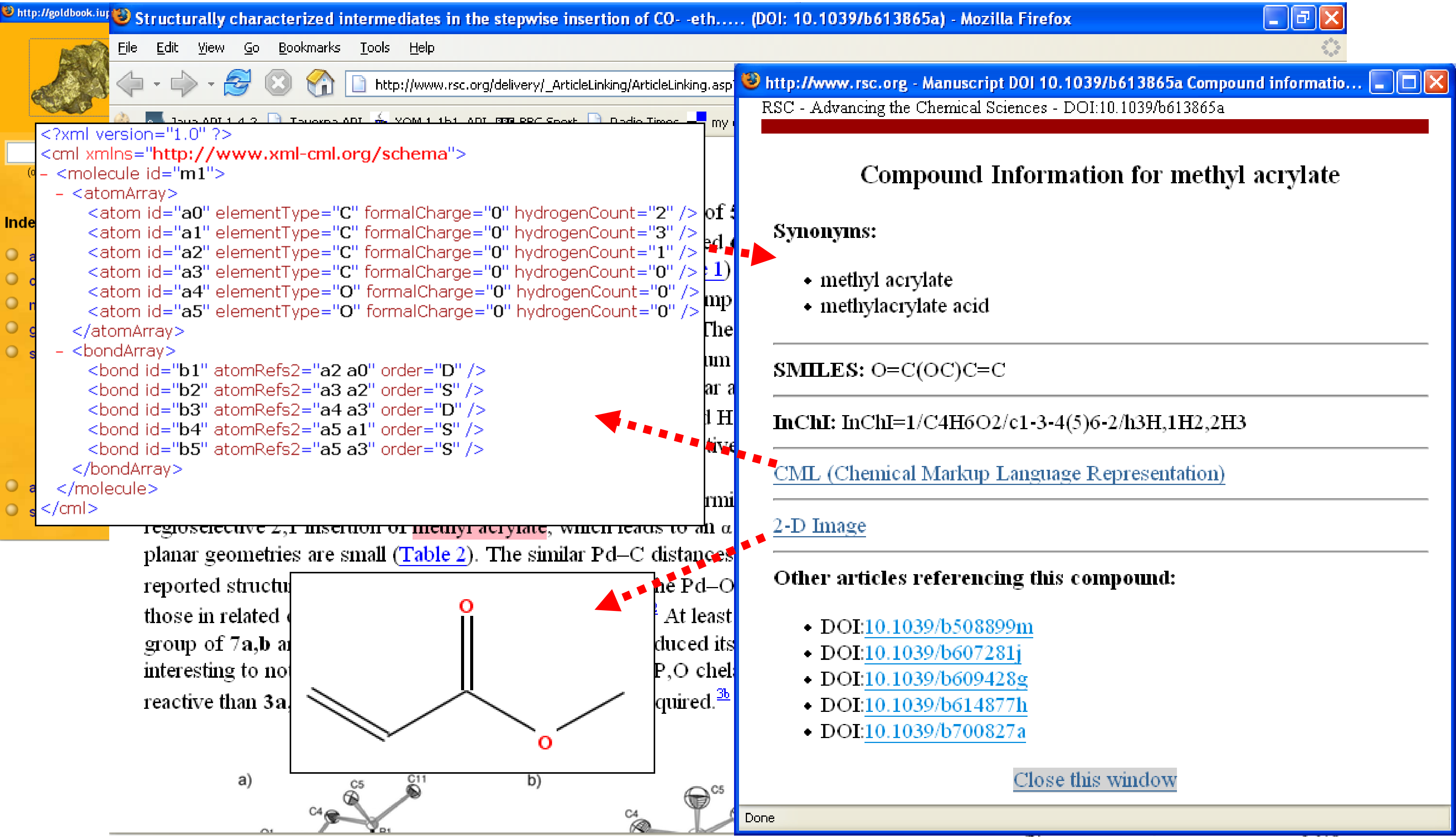
Unilever Cambridge
Centre for Molecular Informatics

… Prospect markup includes:

- CML (Chemical Markup Language)
- InChI
- IUPAC Gold Book
- Gene Ontology
- …and more coming …

…the marked-up semantic paper…

Project Prospect RSC



Structurally characterized intermediates in the stepwise insertion of CO- -eth..... (DOI: 10.1039/b613865a) - Mozilla Firefox

http://www.rsc.org/delivery/_ArticleLinking/ArticleLinking.asp

```
<?xml version="1.0" ?>
<cml xmlns="http://www.xml-cml.org/schema">
- <molecule id="m1">
  - <atomArray>
      <atom id="a0" elementType="C" formalCharge="0" hydrogenCount="2" />
      <atom id="a1" elementType="C" formalCharge="0" hydrogenCount="3" />
      <atom id="a2" elementType="C" formalCharge="0" hydrogenCount="1" />
      <atom id="a3" elementType="C" formalCharge="0" hydrogenCount="0" />
      <atom id="a4" elementType="O" formalCharge="0" hydrogenCount="0" />
      <atom id="a5" elementType="O" formalCharge="0" hydrogenCount="0" />
    </atomArray>
  - <bondArray>
      <bond id="b1" atomRefs2="a2 a0" order="D" />
      <bond id="b2" atomRefs2="a3 a2" order="S" />
      <bond id="b3" atomRefs2="a4 a3" order="D" />
      <bond id="b4" atomRefs2="a5 a1" order="S" />
      <bond id="b5" atomRefs2="a5 a3" order="S" />
    </bondArray>
  </molecule>
</cml>
```

regioselective 2,1 insertion of methyl acrylate, which leads to an α planar geometries are small (Table 2). The similar Pd–C distances reported structu… he Pd–O those in related … At least group of 7a,b a… duced its interesting to no… P,O chel reactive than 3a… quired.[3b]

http://www.rsc.org - Manuscript DOI 10.1039/b613865a Compound informatio...

RSC - Advancing the Chemical Sciences - DOI:10.1039/b613865a

Compound Information for methyl acrylate

Synonyms:

- methyl acrylate
- methylacrylate acid

SMILES: O=C(OC)C=C

InChI: InChI=1/C4H6O2/c1-3-4(5)6-2/h3H,1H2,2H3

CML (Chemical Markup Language Representation)

2-D Image

Other articles referencing this compound:

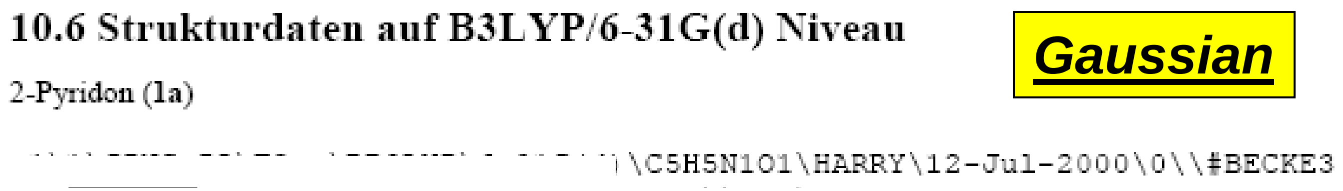
- DOI:10.1039/b508899m
- DOI:10.1039/b607281j
- DOI:10.1039/b609428g
- DOI:10.1039/b614877h
- DOI:10.1039/b700827a

Close this window

Done

… but not all chemistry is in free text …

Chemistry is also “Data”

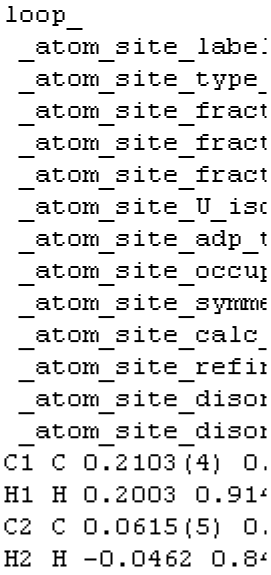


Gaussian

10.6 Strukturdaten auf B3LYP/6-31G(d) Niveau

2-Pyridon (1a)

```
\C5H5N1O1\HARRY\12-Jul-2000\0\\#BECKE3
```



```
loop_
  _atom_site_label
  _atom_site_type_
  _atom_site_fract
  _atom_site_fract
  _atom_site_fract
  _atom_site_U_isc
  _atom_site_adp_t
  _atom_site_occup
  _atom_site_symme
  _atom_site_calc_
  _atom_site_refir
  _atom_site_disor
  _atom_site_disor
C1 C 0.2103(4) 0.
H1 H 0.2003 0.91
C2 C 0.0615(5) 0.
H2 H -0.0462 0.8
```

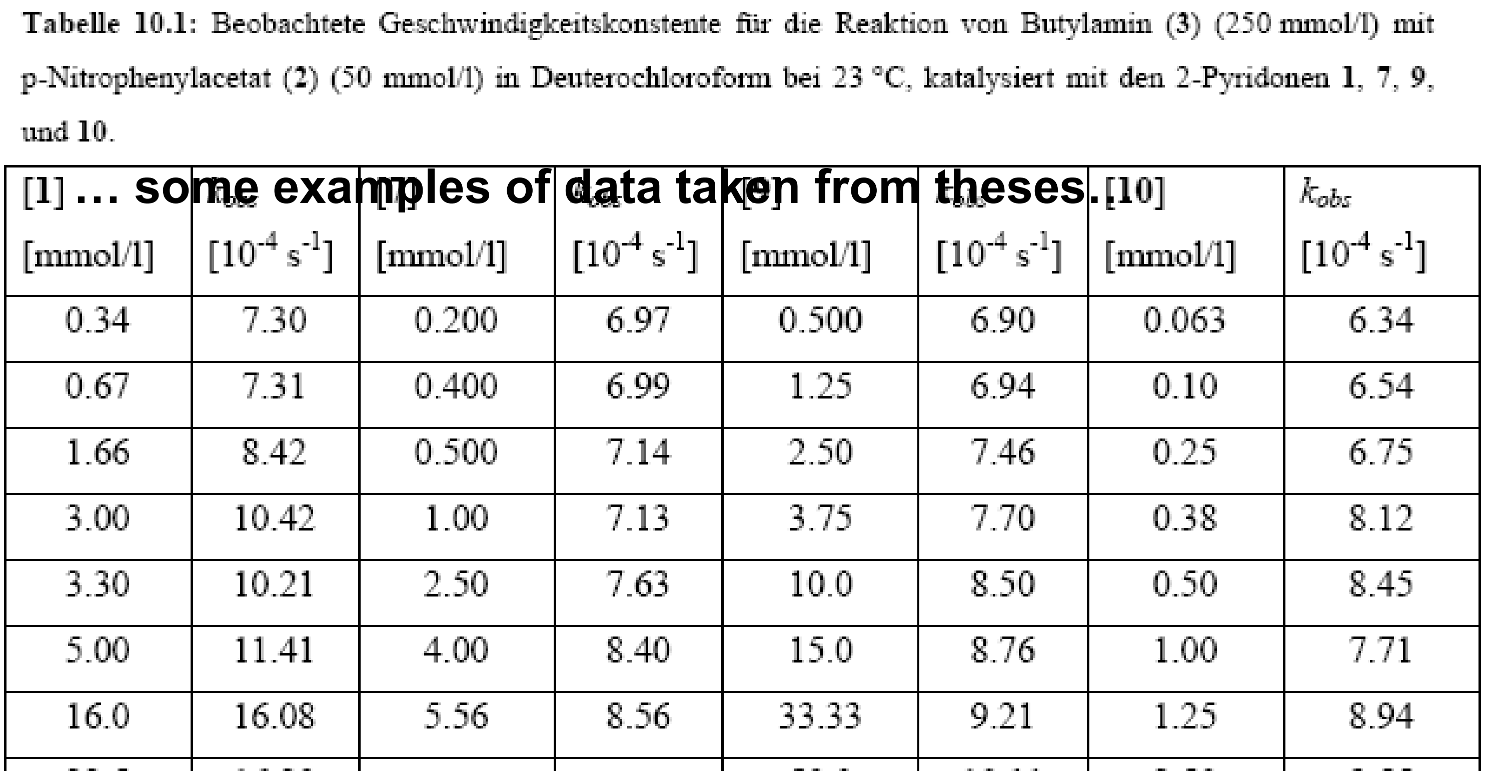


Tabelle 10.1: Beobachtete Geschwindigkeitskonstente für die Reaktion von Butylamin (3) (250 mmol/l) mit p-Nitrophenylacetat (2) (50 mmol/l) in Deuterochloroform bei 23 °C, katalysiert mit den 2-Pyridonen 1, 7, 9, und 10.

… some examples of data taken from theses…

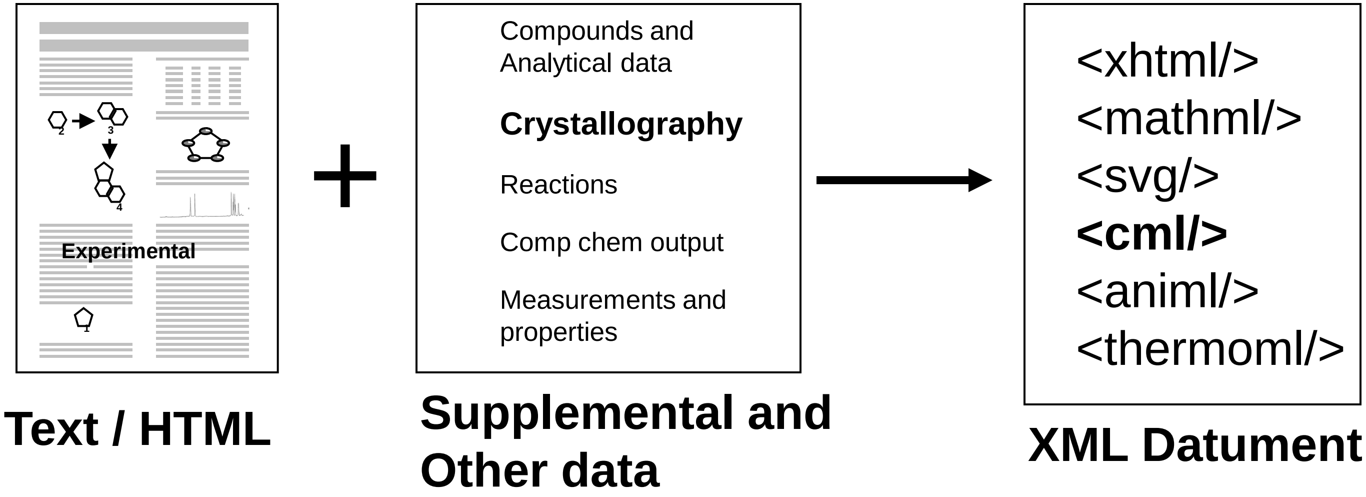
[1]	k_{obs}	[illegible]	k_{obs}	[illegible]	k_{obs}	[10]	k_{obs}
[mmol/l]	$[10^{-4} s^{-1}]$	[mmol/l]	$[10^{-4} s^{-1}]$	[mmol/l]	$[10^{-4} s^{-1}]$	[mmol/l]	$[10^{-4} s^{-1}]$
0.34	7.30	0.200	6.97	0.500	6.90	0.063	6.34
0.67	7.31	0.400	6.99	1.25	6.94	0.10	6.54
1.66	8.42	0.500	7.14	2.50	7.46	0.25	6.75
3.00	10.42	1.00	7.13	3.75	7.70	0.38	8.12
3.30	10.21	2.50	7.63	10.0	8.50	0.50	8.45
5.00	11.41	4.00	8.40	15.0	8.76	1.00	7.71
16.0	16.08	5.56	8.56	33.33	9.21	1.25	8.94

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

Semantic Chemistry and the Datument

The **document** is only part of the scientific record
We can transform the experimental **data** to CML.
The integrated result is a **datument**…



Compounds and Analytical data

Crystallography

Reactions

Comp chem output

Measurements and properties

<xhtml/>
<mathml/>
<svg/>
<cml/>
<animl/>
<thermoml/>

Text / HTML

Supplemental and Other data

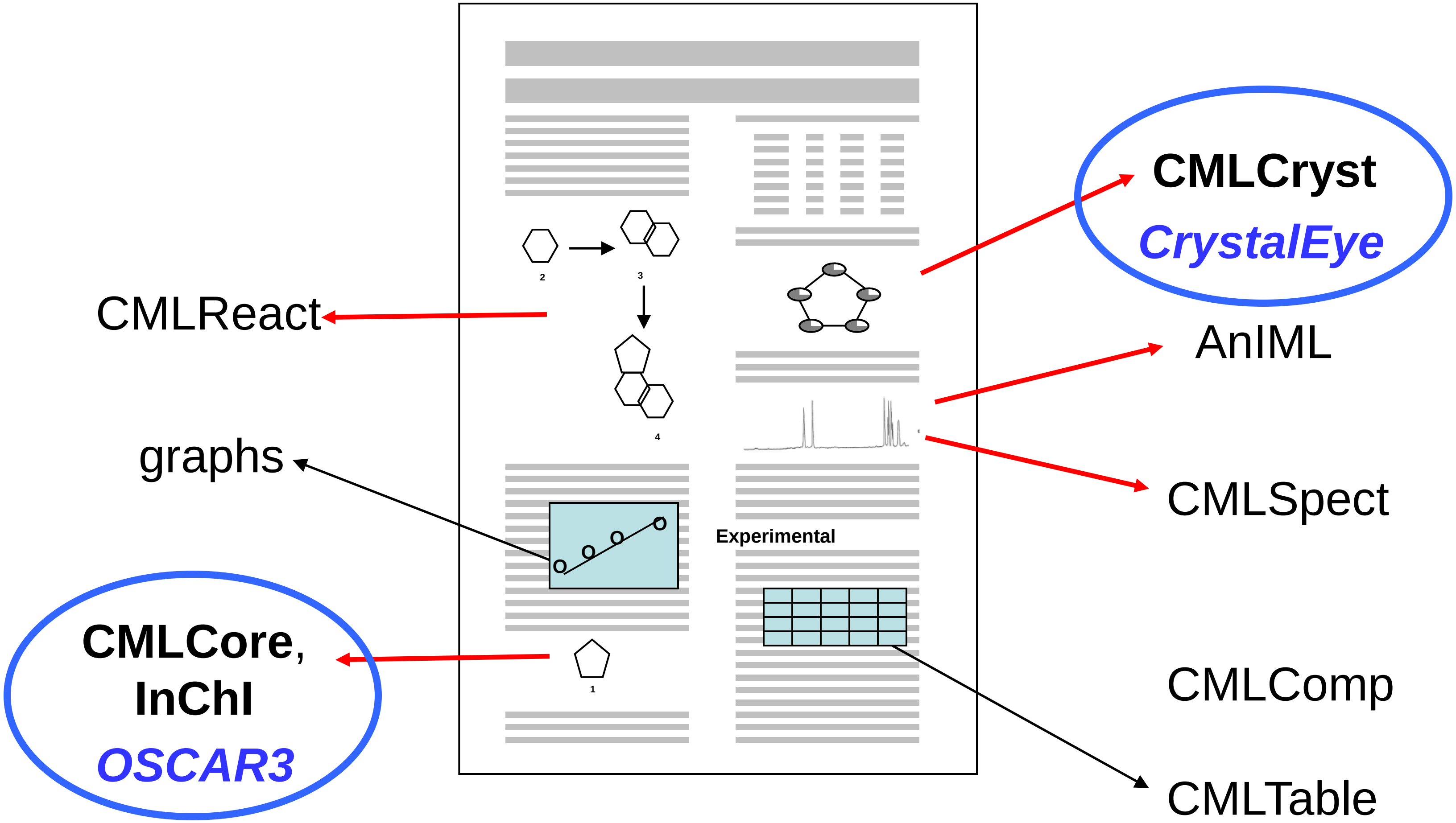
XML Datument

- P. Murray-Rust, *The complete chemical E-publication*, 216th ACS National Meeting, Boston, August 23-27 (**1998**), CINF-033.
- Peter Murray-Rust, Henry S. Rzepa and Michael Wright, Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content, *New J. Chem.*, **2001**, 618-634.
- P. Murray-Rust and H. S. Rzepa, "The Next Big Thing: From Hypermedia to Datuments", *J. Digital Inf.*, **2004**, **5**, article 248, 2004-03-18.

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

The Datument



CMLCryst

CrystalEye

CMLReact

AnIML

graphs

CMLSpect

CMLCore,
InChI

OSCAR3

CMLComp

CMLTable

Chemical Crystallography

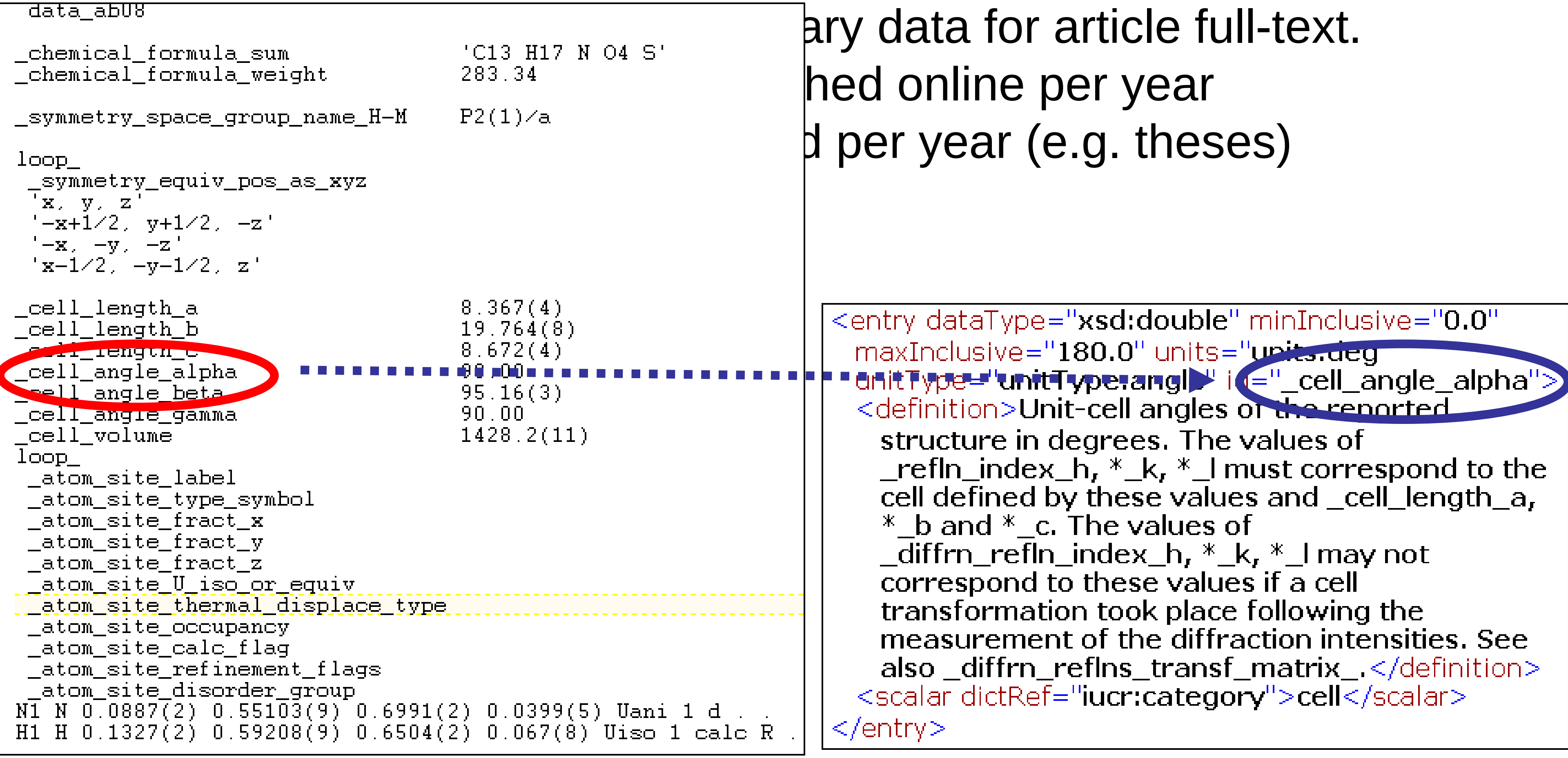
Universally published as **CIF**.

- Complete output of structure experiment,

ary data for article full-text.

hed online per year

d per year (e.g. theses)



```
 data_ab08

_chemical_formula_sum              'C13 H17 N O4 S'
_chemical_formula_weight           283.34

_symmetry_space_group_name_H-M     P2(1)/a

loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, y+1/2, -z'
 '-x, -y, -z'
 'x-1/2, -y-1/2, z'

_cell_length_a                     8.367(4)
_cell_length_b                     19.764(8)
_cell_length_c                     8.672(4)
_cell_angle_alpha                  90.00
_cell_angle_beta                   95.16(3)
_cell_angle_gamma                  90.00
_cell_volume                       1428.2(11)
loop_
 _atom_site_label
 _atom_site_type_symbol
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_U_iso_or_equiv
 _atom_site_thermal_displace_type
 _atom_site_occupancy
 _atom_site_calc_flag
 _atom_site_refinement_flags
 _atom_site_disorder_group
N1 N 0.0887(2) 0.55103(9) 0.6991(2) 0.0399(5) Uani 1 d . .
H1 H 0.1327(2) 0.59208(9) 0.6504(2) 0.067(8) Uiso 1 calc R .
```

```
<entry dataType="xsd:double" minInclusive="0.0"
  maxInclusive="180.0" units="units:deg
  unitType="unitType:angle" id="_cell_angle_alpha">
  <definition>Unit-cell angles of the reported
    structure in degrees. The values of
    _refln_index_h, *_k, *_l must correspond to the
    cell defined by these values and _cell_length_a,
    *_b and *_c. The values of
    _diffrn_refln_index_h, *_k, *_l may not
    correspond to these values if a cell
    transformation took place following the
    measurement of the diffraction intensities. See
    also _diffrn_reflns_transf_matrix_.</definition>
  <scalar dictRef="iucr:category">cell</scalar>
</entry>
```

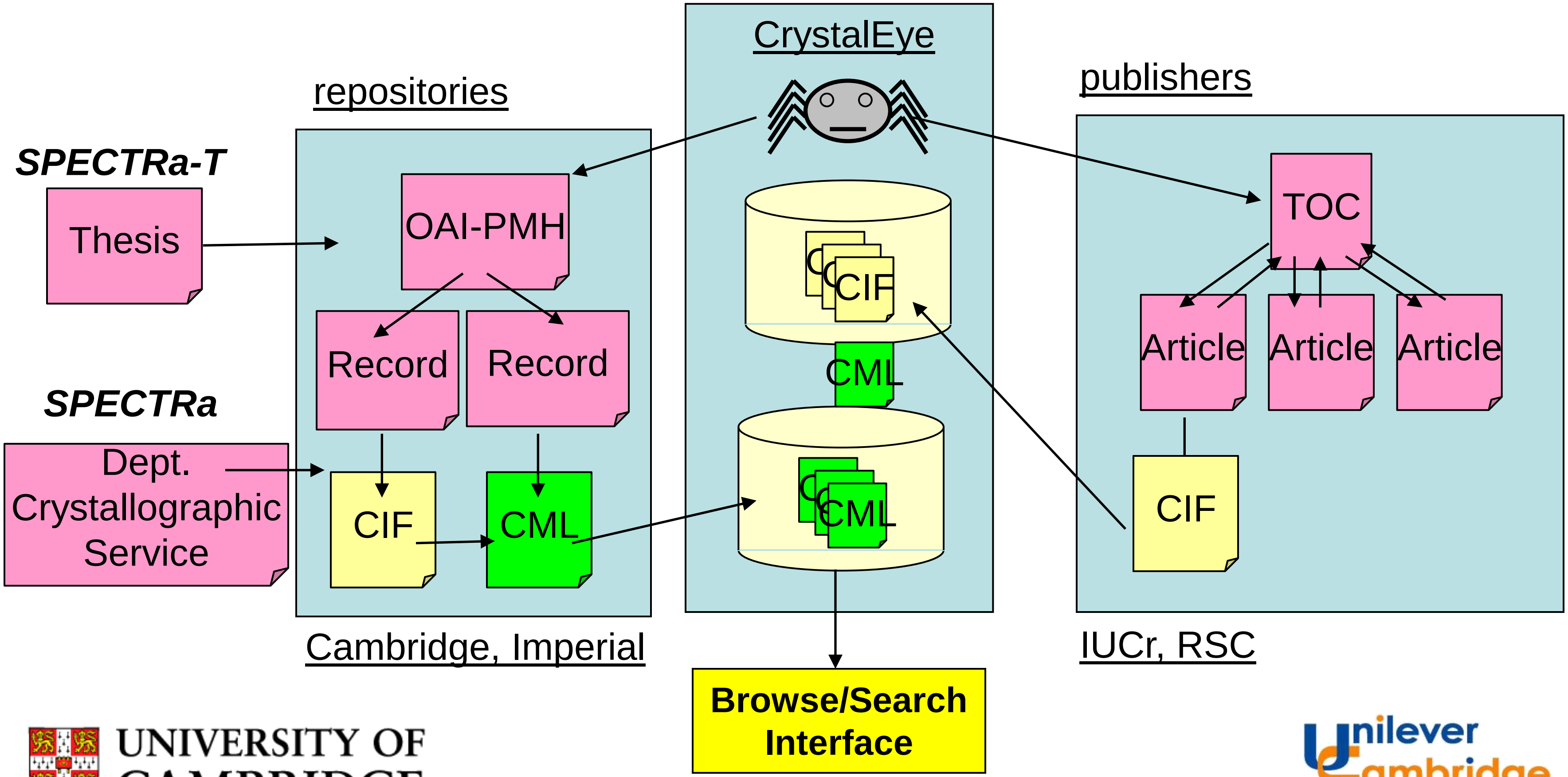
CrystalEye

The aim:

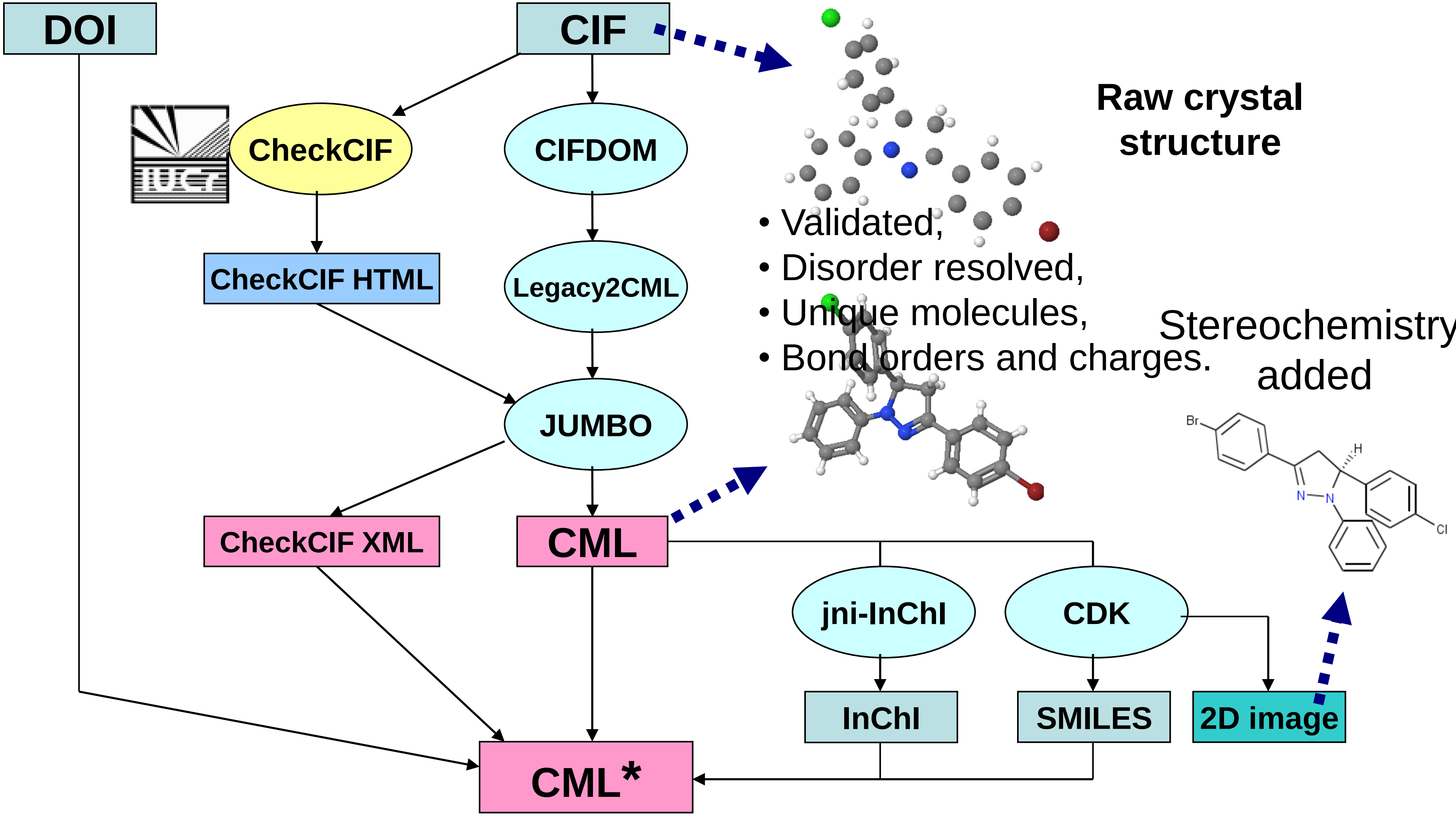
To automatically create semantic chemistry from crystallography (CIFs) published on the Web.

Aggregation

- Web spider checks publishers and repositories every day.
- Currently over **60,000 validated** CIF files.



Marking up and Validation



DOI

CIF

CheckCIF

CIFDOM

CheckCIF HTML

Legacy2CML

JUMBO

CheckCIF XML

CML

jni-InChI

CDK

InChI

SMILES

2D image

CML*

Raw crystal structure

- Validated,
- Disorder resolved,
- Unique molecules,
- Bond orders and charges.

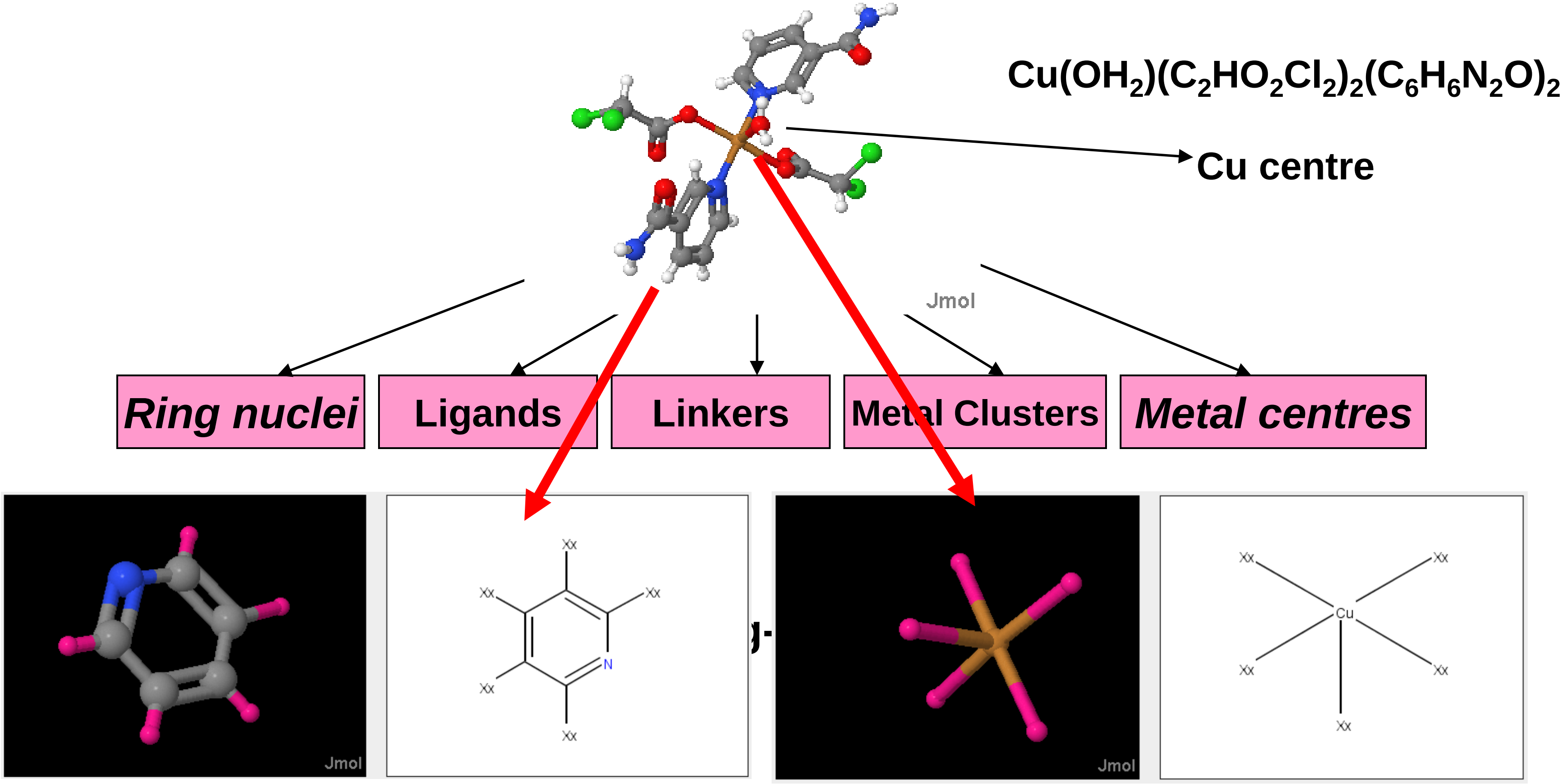
Stereochemistry added

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

CrystalEye: Re-Use through XML/CML

- Automatic generation of fragments



$Cu(OH_2)(C_2HO_2Cl_2)_2(C_6H_6N_2O)_2$

Cu centre

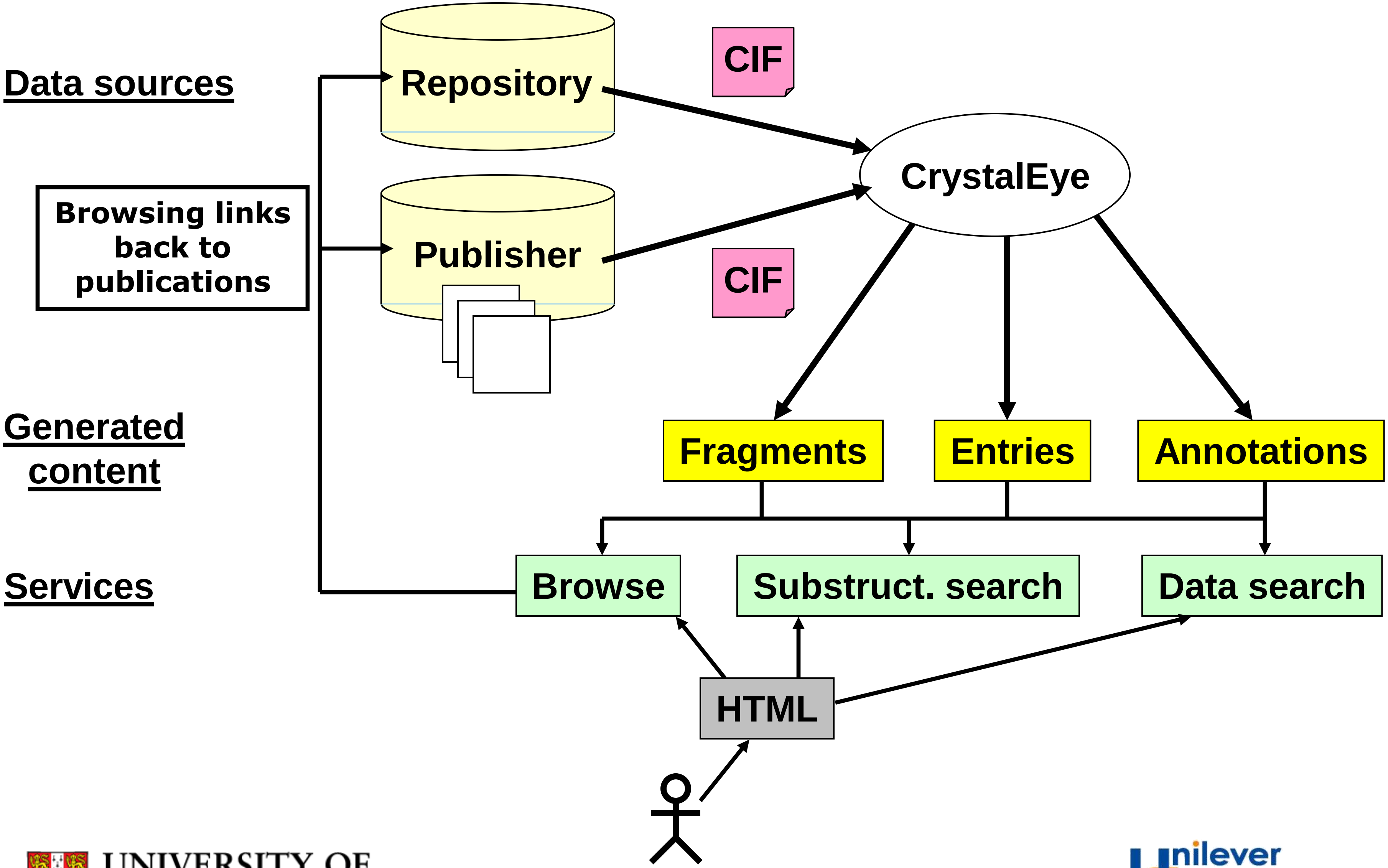
Ring nuclei | Ligands | Linkers | Metal Clusters | *Metal centres*

- ca. 1 million fragments with 50,000 different chemical types
- Open Access via automatically generated HTML

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge Centre for Molecular Informatics

CrystalEye for humans



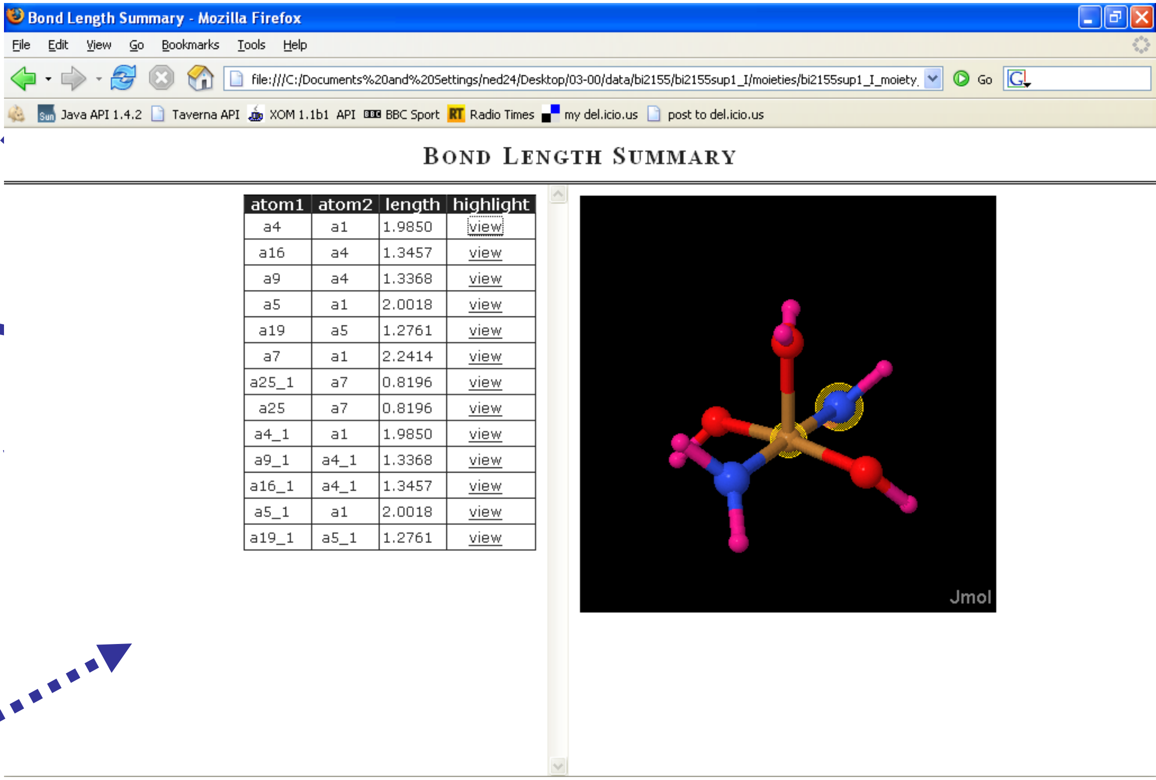
UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

CrystalEye Webpage Demo

Let’s assume we’re interested in Cu-N bonds:

- Browse to title page
- View structure data
- Explore fragments
- Inspect bond lengths


Bond Length Summary - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Java API 1.4.2 Taverna API XOM 1.1b1 API BBC Sport Radio Times my del.icio.us post to del.icio.us
Jmol

file:///C:/Documents%20and%20Settings/ned24/Desktop/03-00/data/bi2155/bi2155sup1_I/moieties/bi2155sup1_I_moiety

BOND LENGTH SUMMARY

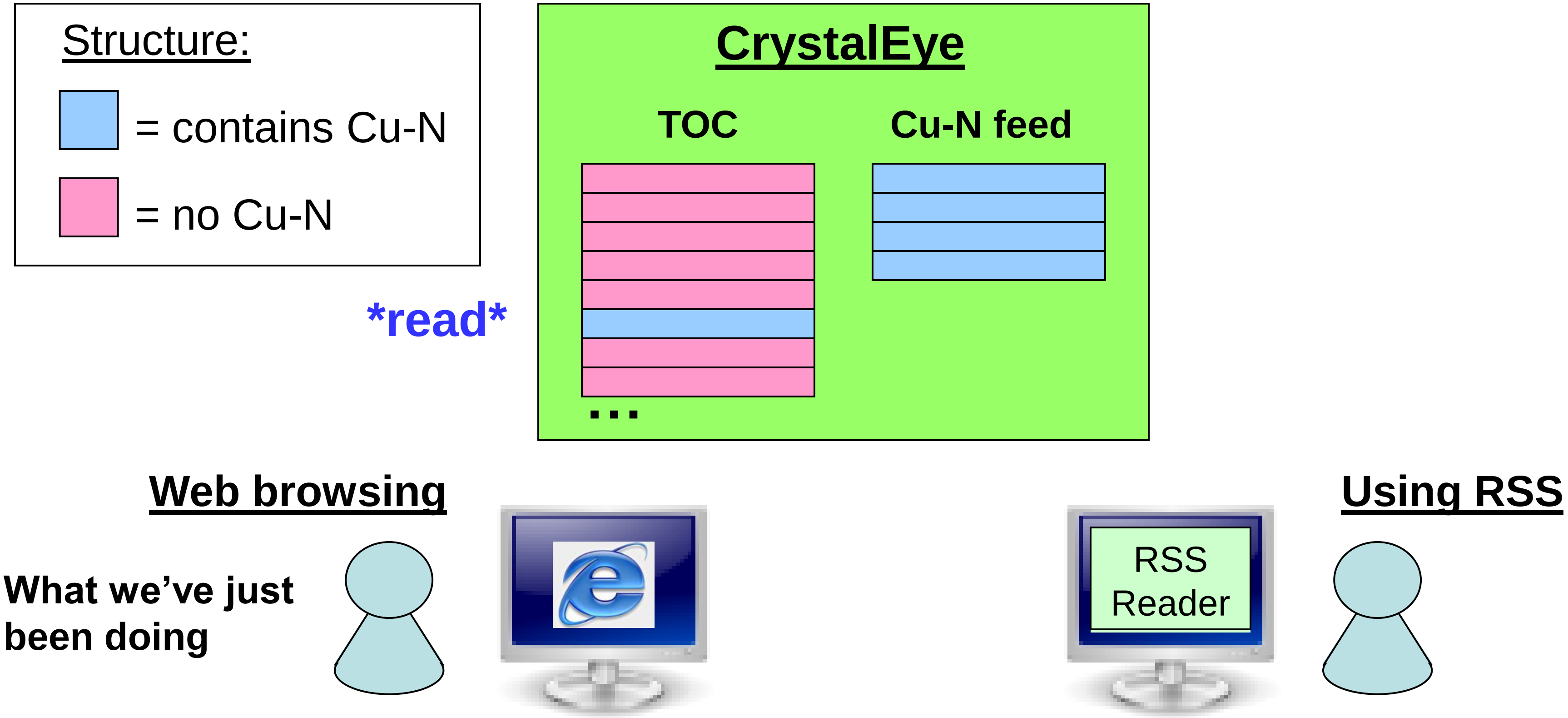
atom1	atom2	length	highlight
a4	a1	1.9850	view
a16	a4	1.3457	view
a9	a4	1.3368	view
a5	a1	2.0018	view
a19	a5	1.2761	view
a7	a1	2.2414	view
a25_1	a7	0.8196	view
a25	a7	0.8196	view
a4_1	a1	1.9850	view
a9_1	a4_1	1.3368	view
a16_1	a4_1	1.3457	view
a5_1	a1	2.0018	view
a19_1	a5_1	1.2761	view

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

CMLRSS newsfeeds

- How can the chemist find every Cu-N bond immediately?



Structure:

🟦 = contains Cu-N

🟥 = no Cu-N

CrystalEye

TOC

Cu-N feed

...

read

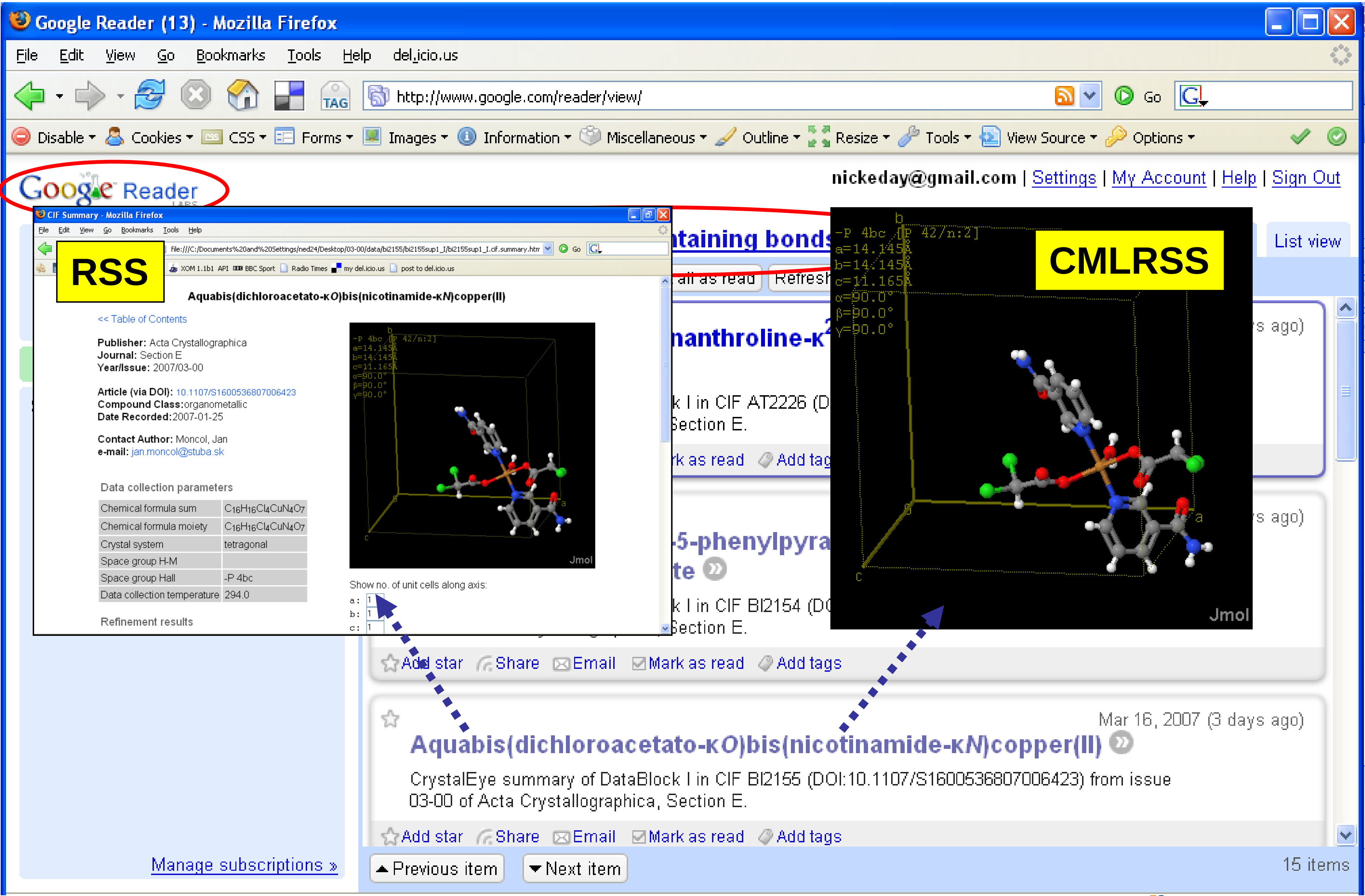
Web browsing

What we've just been doing

RSS Reader

Using RSS

- CrystalEye uses RSS 1, RSS 2 and Atom 1.0 to create both RSS and CMLRSS feeds.


Google Reader (13) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help del.icio.us
http://www.google.com/reader/view/
Go
Disable Cookies CSS Forms Images Information Miscellaneous Outline Resize Tools View Source Options
Google Reader
nickeday@gmail.com | Settings | My Account | Help | Sign Out
List view
CIF Summary - Mozilla Firefox
file:///C:/Documents%20and%20Settings/ned24/Desktop/03-00/data/bi2155/bi2155sup1_I/bi2155sup1_I.cif.summary.htm
XOM 1.1b1 API BBC Sport Radio Times my del.icio.us post to del.icio.us
RSS
CMLRSS
-P 4bc [P 42/n:2]
a=14.145Å
b=14.145Å
c=11.165Å
α=90.0°
β=90.0°
γ=90.0°
a
b
c
Jmol
Manage subscriptions »
▲ Previous item
▼ Next item
15 items

Aquabis(dichloroacetato-κ*O*)bis(nicotinamide-κ*N*)copper(II)

<< Table of Contents

Publisher: Acta Crystallographica
Journal: Section E
Year/Issue: 2007/03-00

Article (via DOI): 10.1107/S1600536807006423
Compound Class:organometallic
Date Recorded:2007-01-25

Contact Author: Moncol, Jan
e-mail: jan.moncol@stuba.sk

Data collection parameters

Chemical formula sum	$C_{16}H_{16}Cl_4CuN_4O_7$
Chemical formula moiety	$C_{16}H_{16}Cl_4CuN_4O_7$
Crystal system	tetragonal
Space group H-M	
Space group Hall	-P 4bc
Data collection temperature	294.0

Refinement results

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

...taining bonds
...all as read Refresh
...nanthroline-κ...
...k I in CIF AT2226 (D...
...Section E.
...rk as read Add tag...
...s ago)
...5-phenylpyra...
...te
...k I in CIF BI2154 (DO...
...Section E.
...s ago)

☆ Add star Share Email ☑ Mark as read Add tags

☆ Mar 16, 2007 (3 days ago)

Aquabis(dichloroacetato-κ*O*)bis(nicotinamide-κ*N*)copper(II)

CrystalEye summary of DataBlock I in CIF BI2155 (DOI:10.1107/S1600536807006423) from issue 03-00 of Acta Crystallographica, Section E.

☆ Add star Share Email ☑ Mark as read Add tags

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

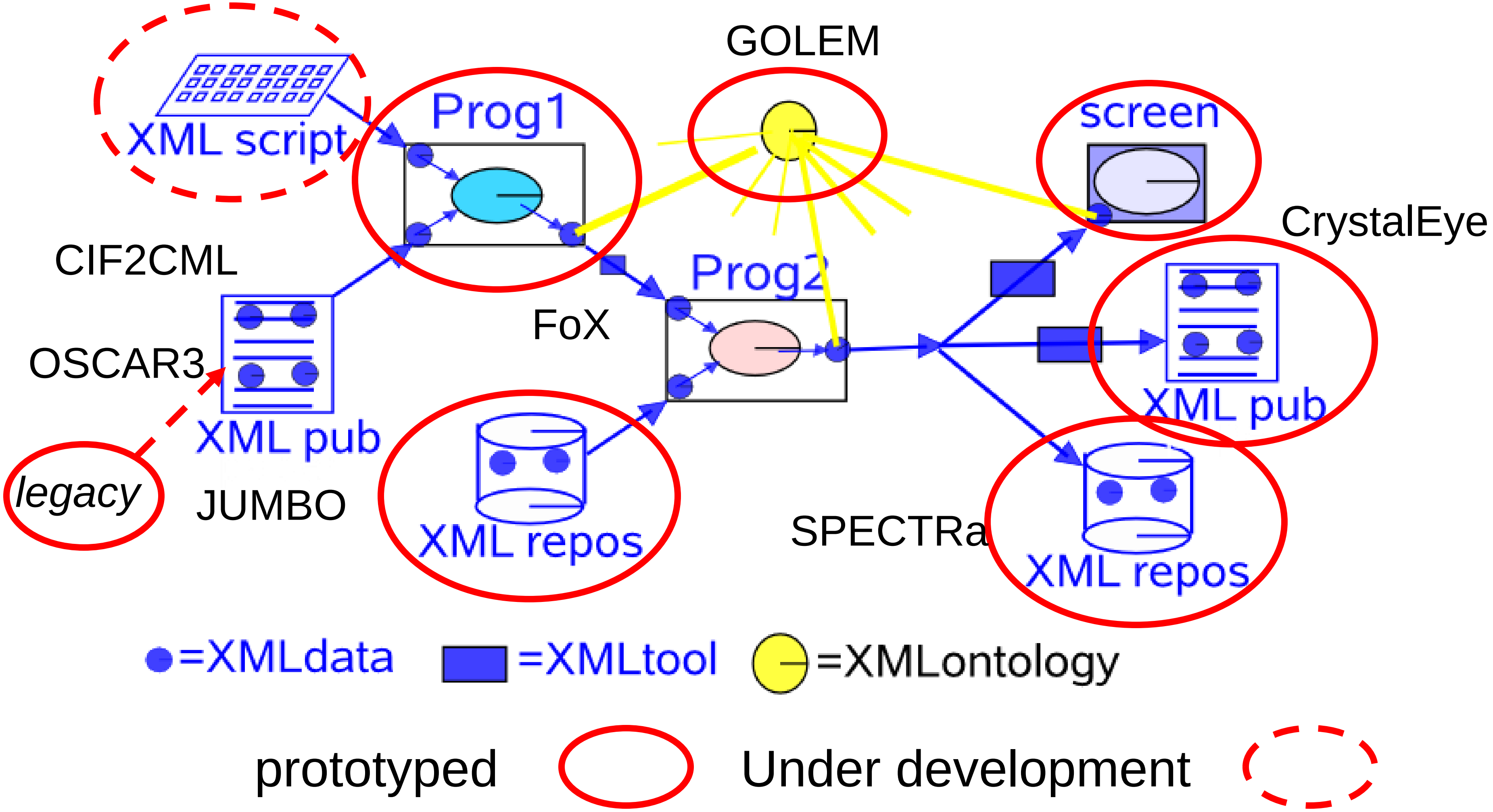
CrystalEye: KnowledgeBase not DataBase

- Aggregation by robots, not humans
- All types of chemistry (organic, inorganic, etc…)
- Social computing - aggregates **COD*,** theses
- Software validation by robots, not humans,
- Open and free;

- Goes live in April…

* **Crystallographic Open Database**

Chemical semantic framework at UCC



XML script

Prog1

GOLEM

screen

CrystalEye

CIF2CML

FoX

Prog2

OSCAR3

XML pub

legacy

JUMBO

XML repos

SPECTRa

XML pub

XML repos

●=XMLdata =XMLtool =XMLontology

prototyped Under development

Cambridge Semantic Chemistry

- Released:
  - CML – XML for chemistry
  - JUMBO – library for CML
  - OSCAR1/CheckCML – data validation
  - OSCAR3 – text mining
  - FoX – Fortran XML

- Soon:
  - CrystalEye – crystallographic knowledgebase
  - SPECTRa – chemical repositories

- Later…
  - Golem - ontologies
  - CMLUnits

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

Acknowledgements

- CML - Henry Rzepa
- OSCAR1 – Sam Adams, Joe Townsend, Fraser Norton, Justin Davies, Richard Marsh, Jonathan Goodman
- OSCAR 3 – Ann Copestake, SimoneTeufel (Computer Lab)
- SPECTRa – Jim Downing, Alan Tonge, Peter Morgan
- Software - CDK, Jmol, jni-InChI and many Blue Obelisk contributions
- Timo Hannay (NPG), Richard Kidd (RSC), Colin Batchelor (RSC), Brian McMahon (IUCr)

npg nature publishing group

RSC | Advancing the Chemical Sciences

IUCr

EPSRC

JISC

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

Thankyou.