The World Wide Molecular Matrix

Nick Day

Unilever Centre for Molecular Informatics,
University of Cambridge

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

The Internet Information Explosion

Symbolised by e.g. Google™, eBay™ and Wikipedia.

With the WWMM we are hoping to provide a chemical equivalent.

Skills for performing Web searches and locating information are common knowledge.

Bioinformatics – the forerunners

Authors are encouraged to make *factual* information from publications available in databases.

- Protein sequences deposited with NCBI,
- structures with PDB,
- disease alleles with (O)MIM etc…

Thus, this information is available to *anyone* connected to the Web.

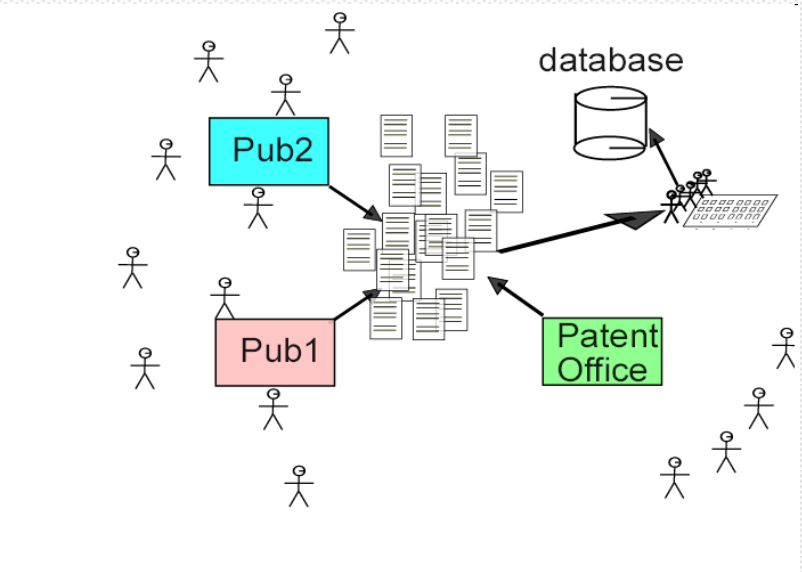
Cheminformatics – lagging…

Chemists can also ‘Google’ for facts and explanations:

- some high-quality curated info is available
    - webElements,
    - molBase,
    - PubChem.
- often data is not well curated or openly visible,
- thus, hard to make informed judgements.

Chemical Publication

Chemistry micropublished by humans then re-aggregated by humans.



The resulting chemical data is closed and generally in formats that are not reusable.

Example of data loss during publication

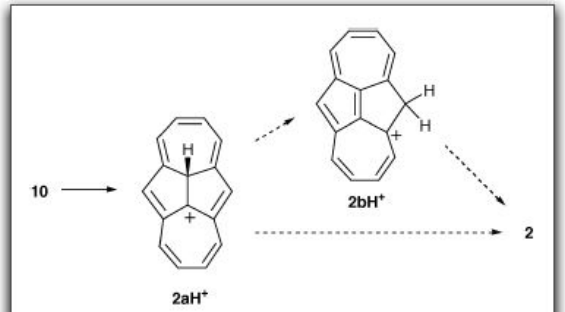
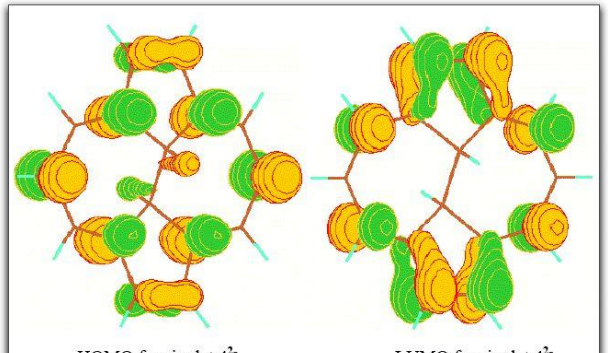


Fig. 10 The mechanism of formation of 2 from 10.

- Reaction is highly symbolic.



HOMO for singlet 4^{2-} LUMO for singlet 4^{2-}

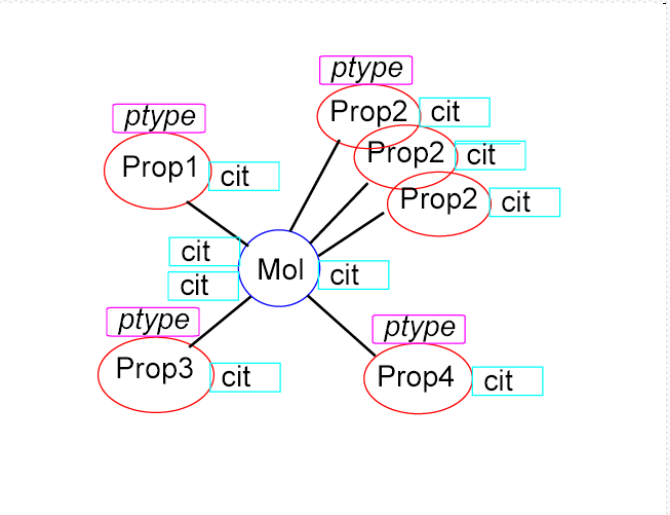
- Wavefunction is a GIF. All previously calculated data is not present.

Why create the WWMM?

- To provide a method for chemists to archive and share their data Openly by:
  - using community agreed markup and metadata, and providing tools to convert to them from ‘legacy’ files (e.g. mol, pdb, sdf etc).
  - storing the data in permanent, maintainable, easily searchable repositories.

What is the WWMM?

- The overall design is of *autonomous* sites that expose data and <u>metadata</u> openly.
- Statement of openness through Creative Commons licensing.
- The key concepts we will encode will represent Beilstein's vision of chemistry:
  - Molecules - ?
  - Properties
  - Provenance



Encoding molecules

We need a way of representing a chemical structure that:

- is unique - *a primary key,*
- today's search methods require the identifier be a text string,
- allows high-performance in database retrieval
  - high recall,
  - low false positives,
  - low false negatives.

Semantically free identifiers

Registry numbers e.g. CAS, RTECs or PubChem identifiers:

- are unique (e.g 58-08-2 is caffeine) but,
- contain no information on the molecule they represent – require a lookup
- lots of false positives when Web searched.

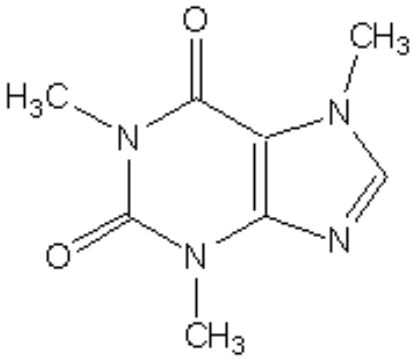
Canonical identifiers

- SMILES notation.
- Converts structure to unique string by algorithm.
- Can hold structural info on connections, stereochemistry, isotopic enrichment.
- …but is proprietary and there is more than one implementation in use.
- Different unique SMILES strings on the Web!

SMILES for caffeine

1. [c]1([n+]([CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]
2. CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12
3. Cn1cnc2n(C)c(=O)n(C)c(=O)c12
4. Cn1cnc2c1c(=O)n(C)c(=O)n2C
5. N1(C)C(=O)N(C)C2=C(C1=O)N(C)C=N2
6. O=C1C2=C(N=CN2C)N(C(=O)N1C)C
7. CN1C=NC2=C1C(=O)N(C)C(=O)N2C

Caffeine

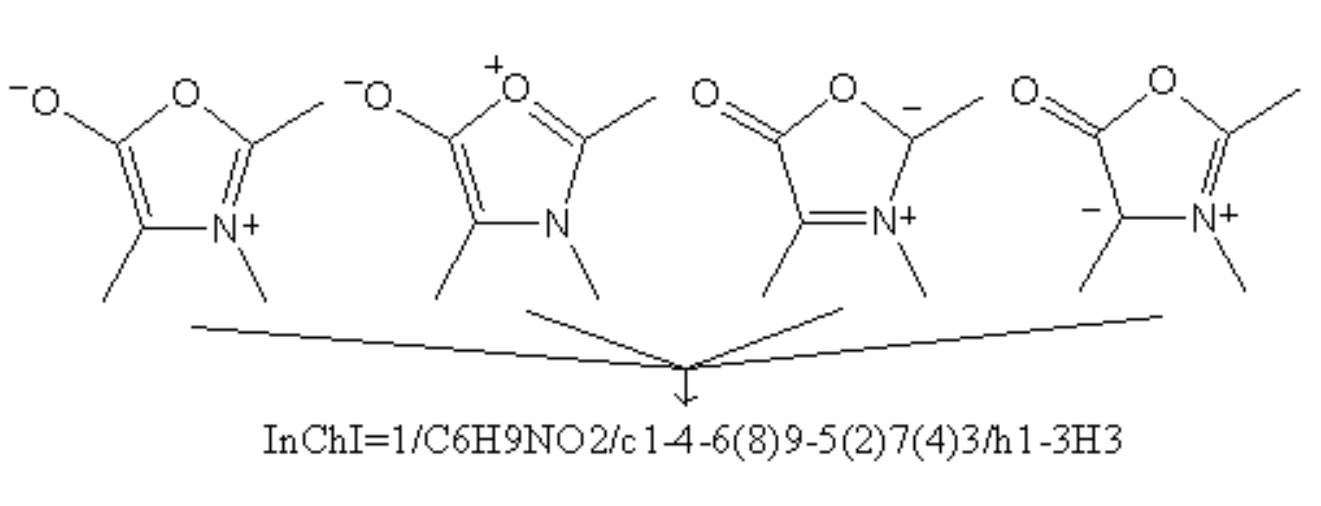


InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

InChI: IUPAC International Chemical Identifier

A non-proprietary unique identifier for the representation of chemical structures.

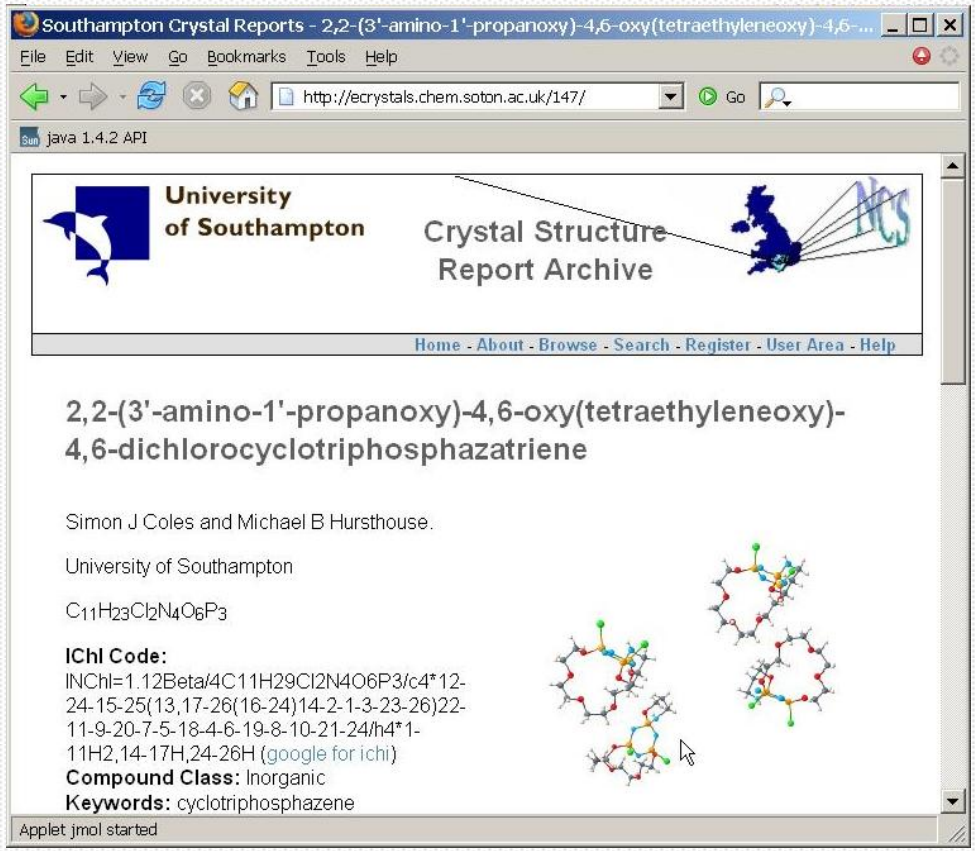
A normalised, canonicalised and serialised form of a chemical connection table.



InChI FAQ: http://wwmm.ch.cam.ac.uk/inchifaq/

Googling for InChIs

Searched for the entire Southampton Crystal Structure Report Archive – 104 structures (18-11-2004).



InChI Search Results

Table 3 InChI retrieval for a set of 104 crystal structures on 18 November, 2004

	Google™	Altavista™	Yahoo™	MSN™
Total XHTML files containing InChIs = 104				
XHTML recall[a]	104 (100%)	39 (38%)	33 (32%)	43 (42%)
Non-InChI false-positives[b]	0	0	0	0
Inter-InChI precision[c]	103	38	32	42
Total CML files containing InChIs = 93				
CML recall[d]	92	0	0	0

[a] Number (and percentage) of XHTML documents containing InChIs retrieved. [b] Number (and percentage) of non-InChIs (*e.g.*, football scores) retrieved. [c] Number of XHTML documents containing correct InChIs retrieved. [d] Number of CML documents containing correct InChIs retrieved.

832 searches performed in total on 8 different search engines with no false positives returned.

Org. Biomol. Chem., 2005, **3**, 1832-1834

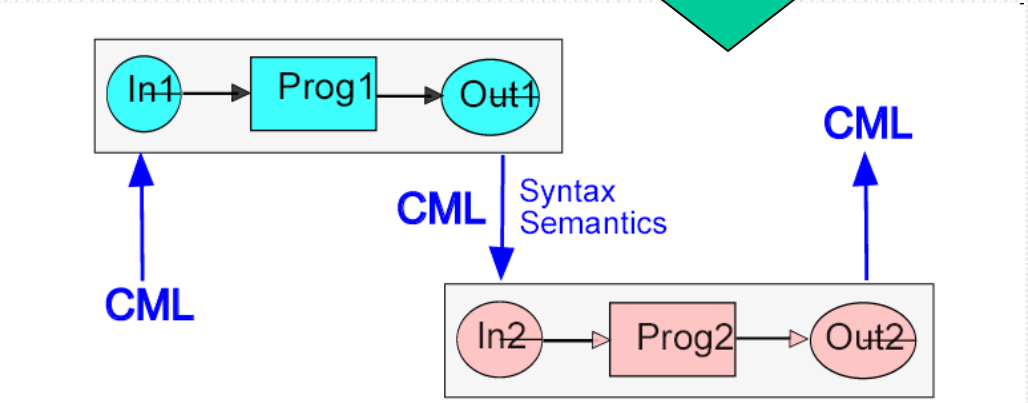
How do we encode properties?

The key concepts we will encode will represent Beilstein's vision of chemistry:

- Molecules – encoded as InChI
- Properties - ?
- Source (provenance)

Chemical Markup Language

- An XML-based language that provides a surface syntax and document structure.
- Can hold all information from legacy files.
- Easily reusable - strict structure means easy to write tools for further conversion or calculation → a good 'glue-ware'.
- Provides a container for InChIs.



Quick CML

```
<molecule id="no2">
  <atomArray>
    <atom id="n1" elementType="N" hydrogenCount="0"/>
    <atom id="o1" elementType="O" hydrogenCount="0"/>
    <atom id="o2" elementType="O" hydrogenCount="0"/>
  </atomArray>
  <bondArray>
    <bond id="bo1" atomRefs2="n1 o1" order="2"/>
    <bond id="bo2" atomRefs2="n1 o2" order="1"/>
  </bondArray>
</molecule>
```

How do we encode provenance?

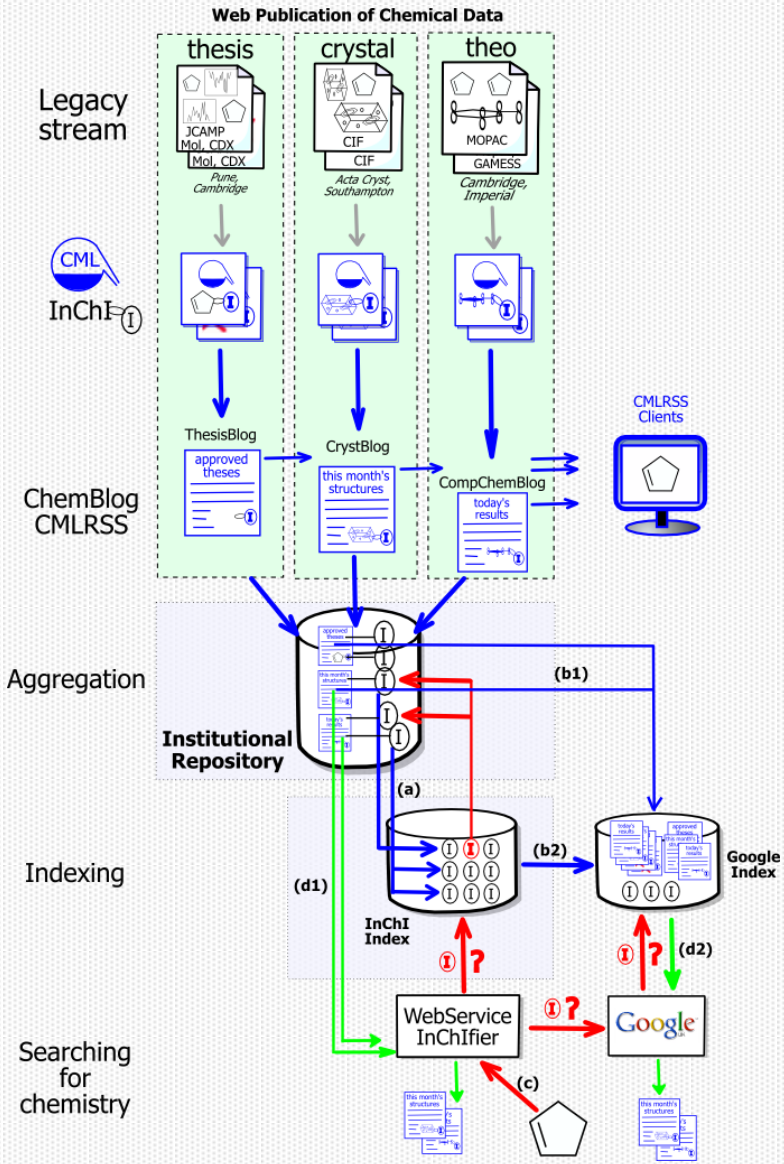
The key concepts we will encode will represent Beilstein's vision of chemistry:

- Molecules – encoded as InChI
- Properties – encoded as CML
- Source (provenance) - ?

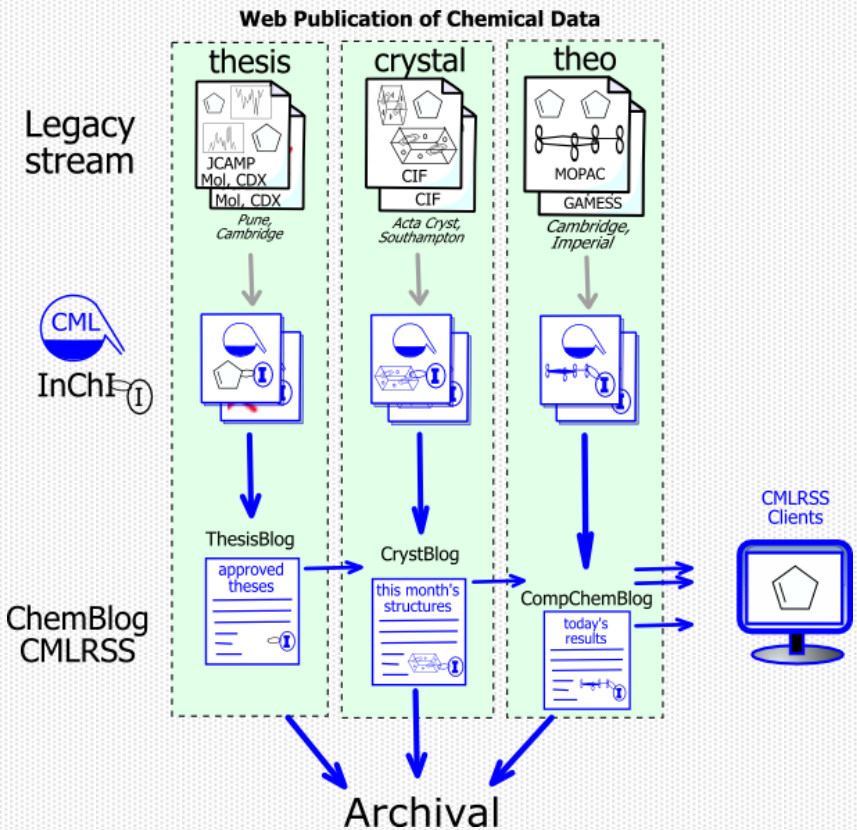
Provenance of data

- Provided by RDF (Resource Description Framework) metadata:
  - Dublin Core – document level metadata
  - FOAF (Friend-of-a-friend) – personal detail metadata
  - DOAP (Description-of-a-project) – used to describe Open Source projects.

WWMM Architecture



Aggregation to archival



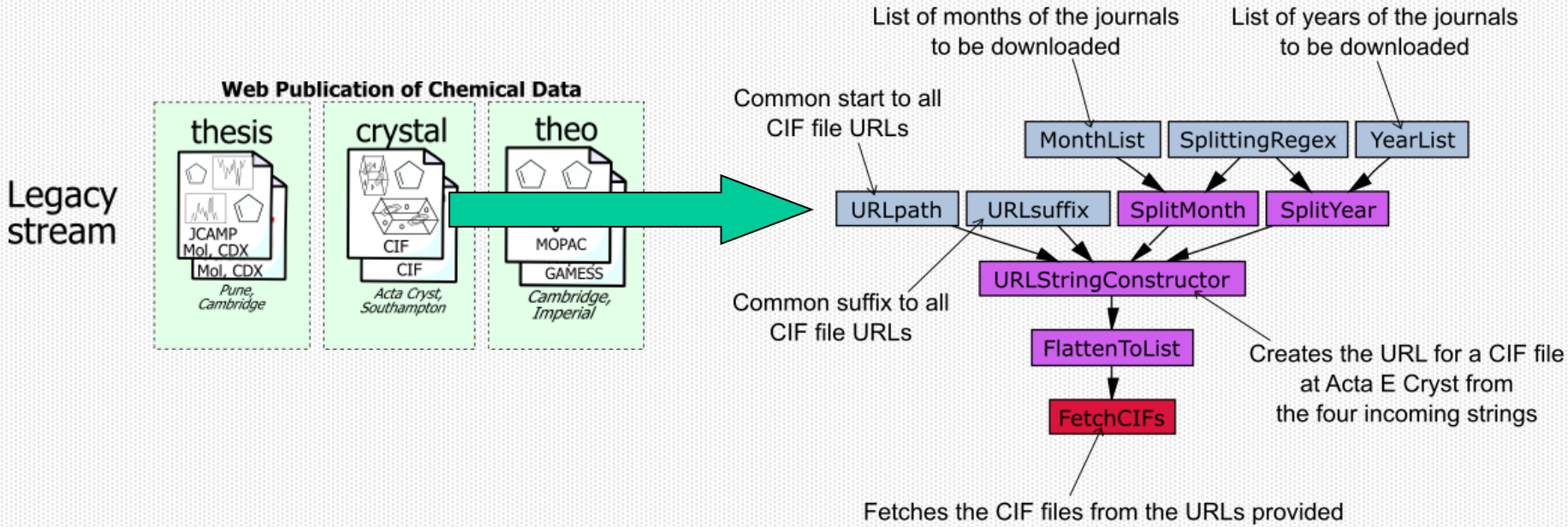
- Creation of our data and metadata for archival.
- Stream based on small modular components.
- Use a low cost, high-throughput workflow system to link the components and manage data flow between.
- Aim to be fully automated.

Taverna

- An Open Source, Java-based workflow management system from the myGrid project.
- Workflow processors can be created from libraries through the use of the ‘API Consumer’.
- We have incorporated JUMBO, the Open modular toolkit into the system.
- Once created, processors can be ‘clicked’ together to create complex technologies from simple building blocks...

Aggregating Legacy Documents

Before any processing is done, we need to collect the legacy formats. Done with a workflow!



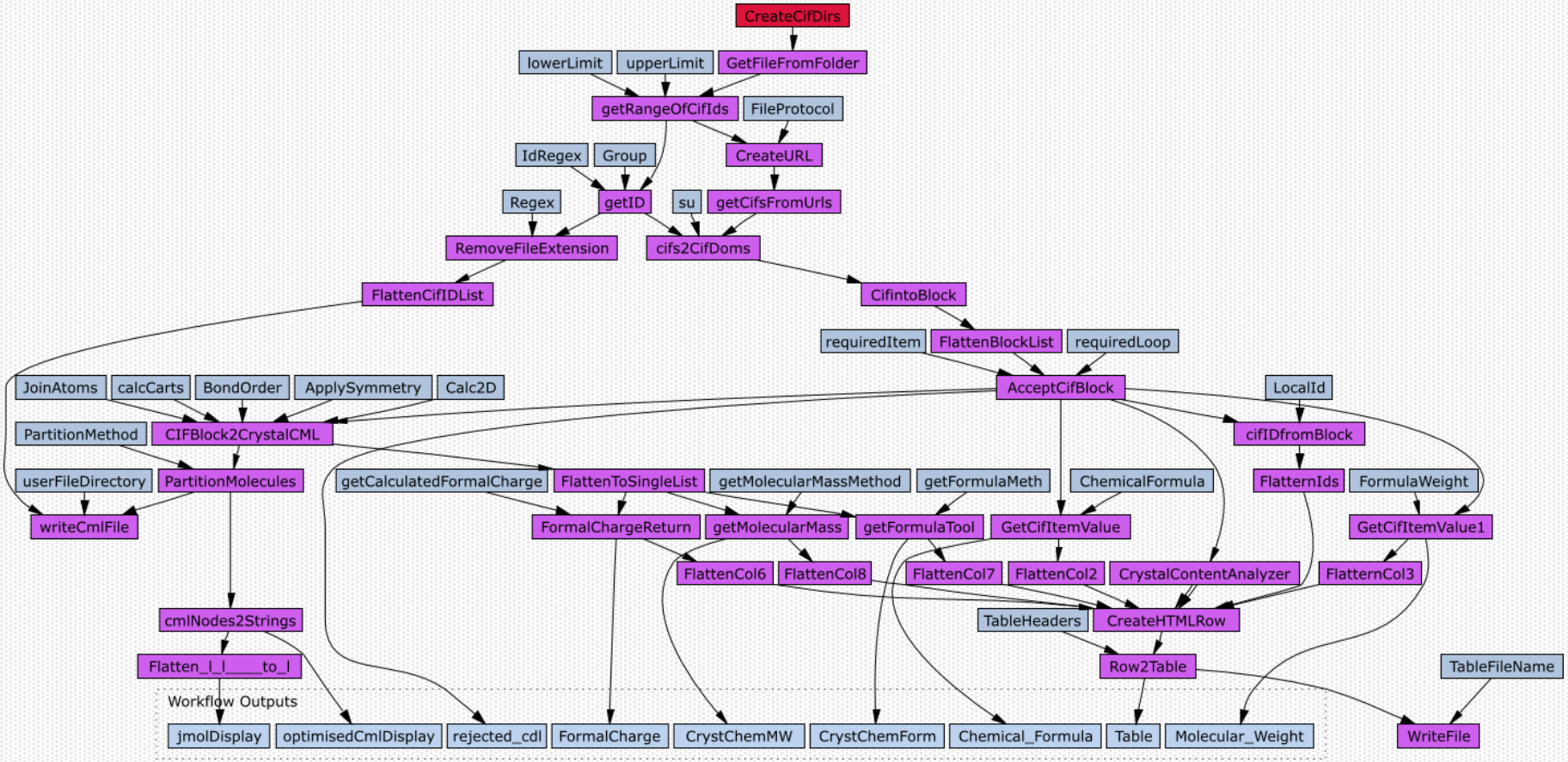
Downloaded 12,000+ CIFs from Acta E. Cryst in ~40mins.

Legacy→CML

- Many legacy formats can be converted to CML using OpenBabel.
- We also have tools for converting
  - CIFs (Crystallographic Interchange Format)
  - MOPAC/GAMESS input and output

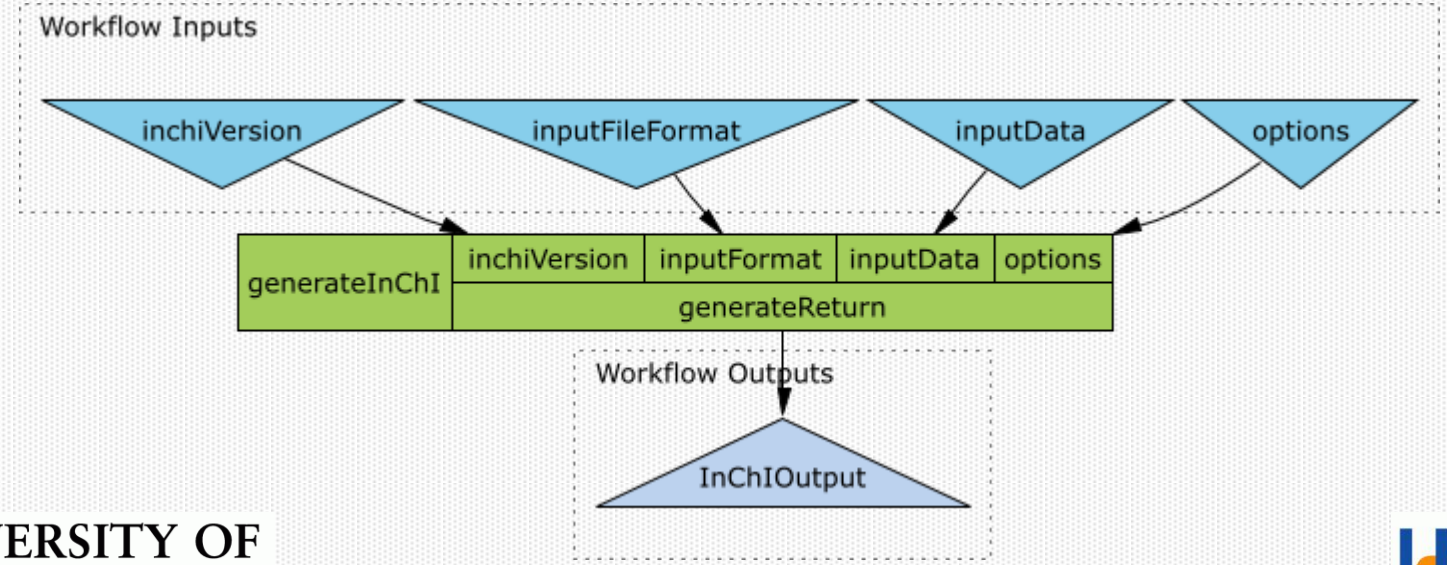
to CML.

CIF2CML Example



Adding InChI

- InChIs are created by sending the CML representation of a molecule to our InChI Web Service, which implements the IUPAC InChI generation app.
- Processing done on our Web server then returned.
- We have implemented this WS in a Taverna workfow.



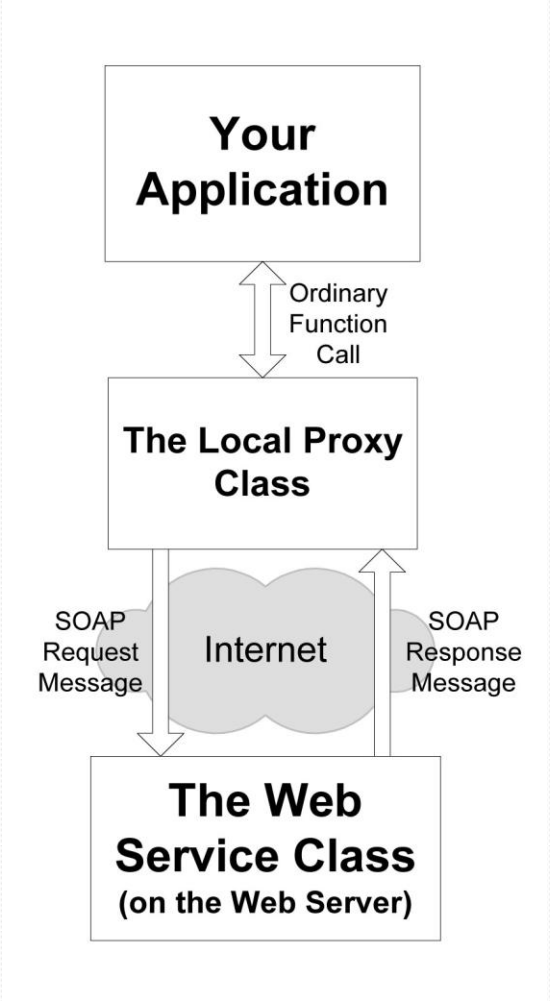
Web Services

A set of protocols that allows applications on remote terminals to communicate through a standard XML-based langauge.

Provides:

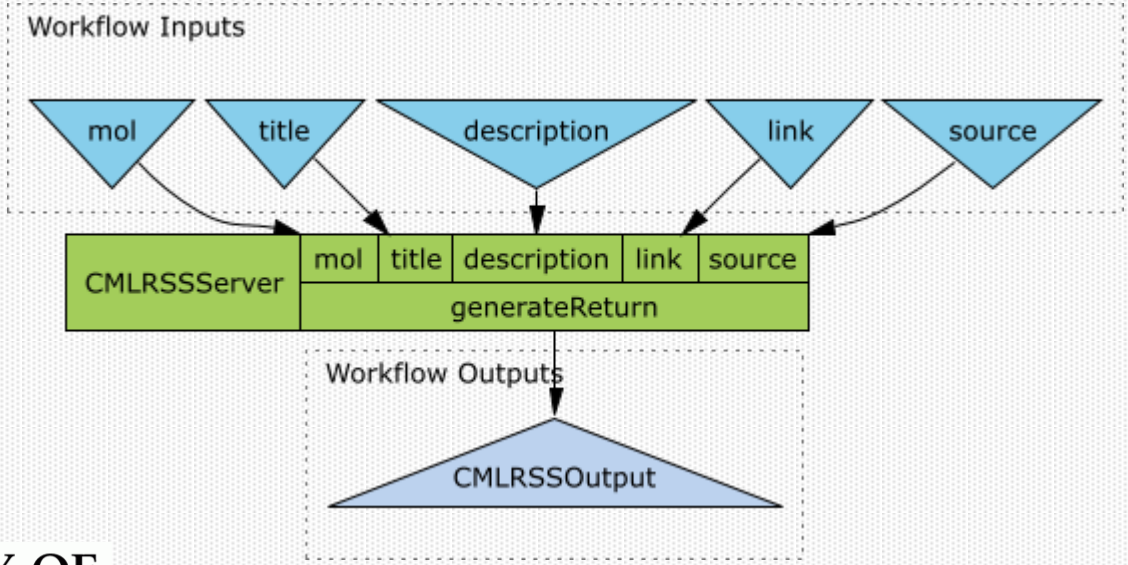
• interoperability – apps in different languages on different platforms can interact.

• ease of reuse – no need for any software downloading or installation.



CML/InChI 2 CMLRSS

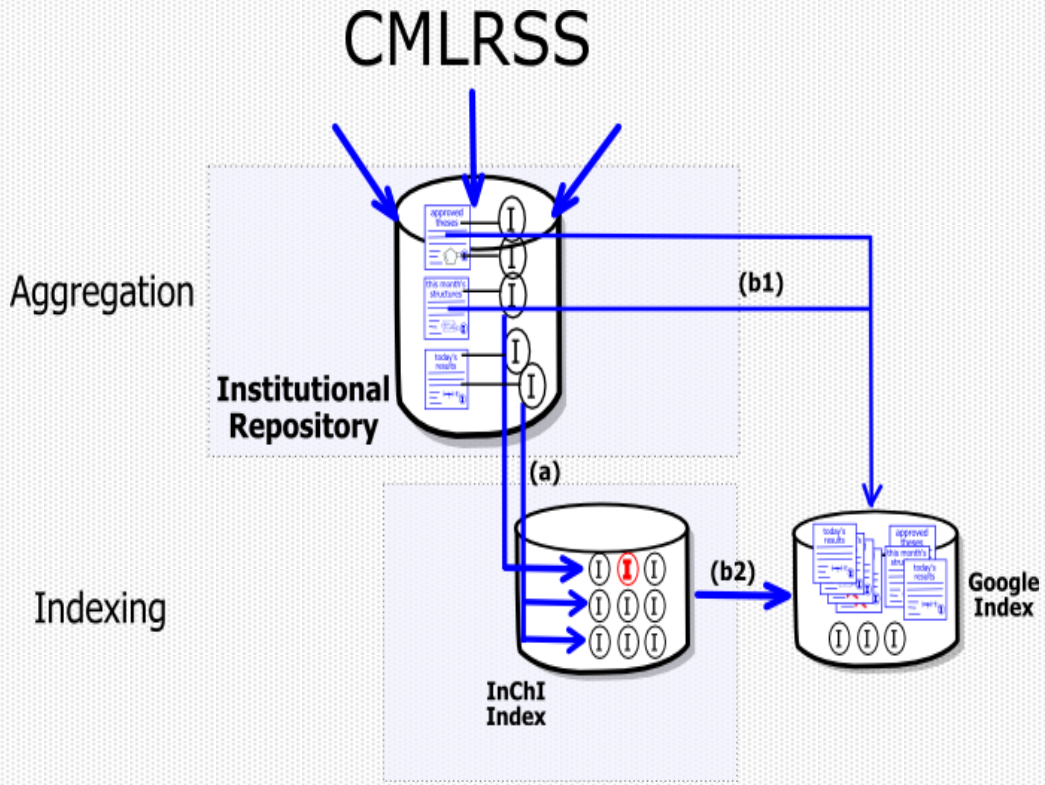
- CMLRSS is an extension of RSS 1.0 which holds CML data.
- CMLRSS creation implemented as a Web Service in Taverna.



Automatic Dissemination

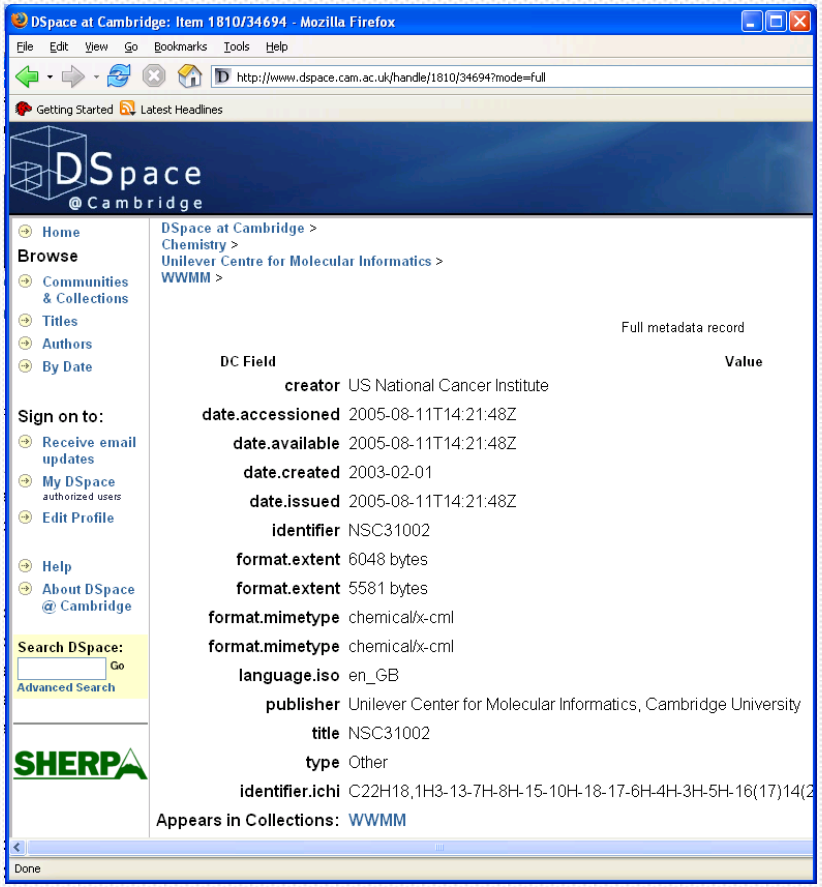
- The CMLRSS for each stream is deposited in separate RSS ‘newsfeeds’ on our server.
- Users can subscribe to these to get the latest chemistry from different sources.

Archiving the data



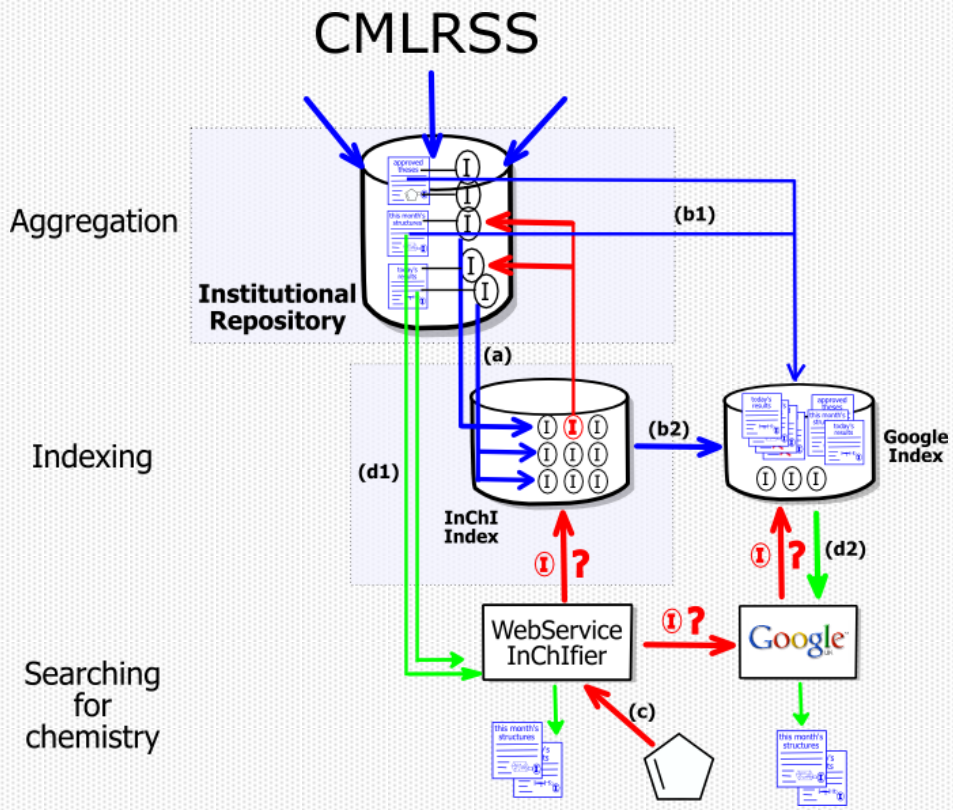
- The CMLRSS is to be directly ingested in an Institutional Repository.
- The data will then be indexed by InChI in a separate repository.
- Provides search engines with a simpler indexing method.

Institutional Repositories



- Provides permanence and maintenance of data.
- Cambridge has a ‘DSpace’ repository.
- Already deposited 250,000 molecules and calculated properties from NCI database.

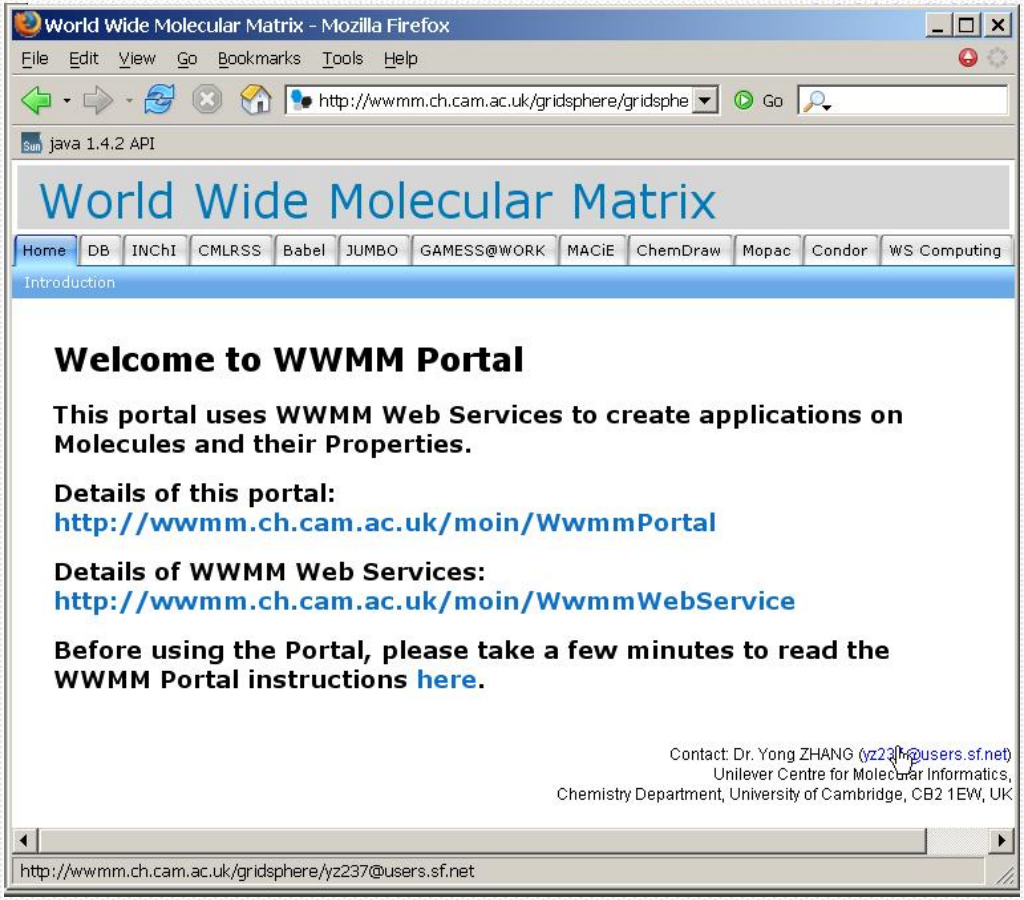
Searching the WWMM



- Search engine queries our only method of searching…for now.
- In the future we may rely on OAI-PMH for searching.
-

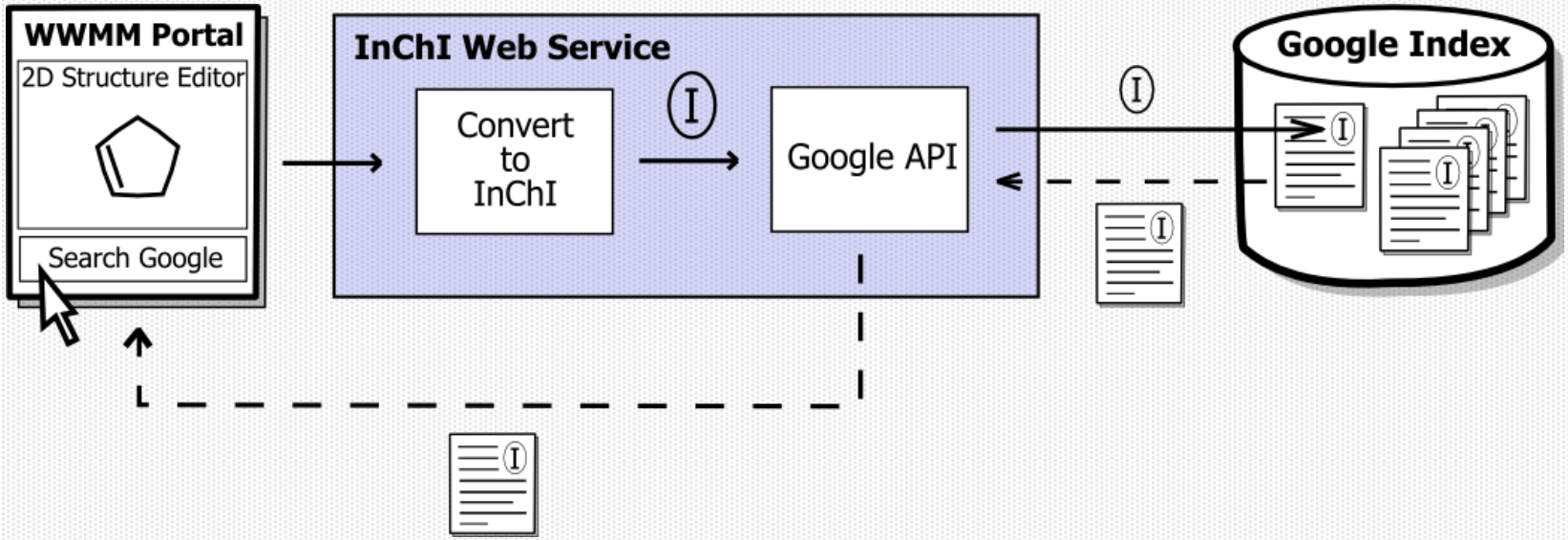
The WWMM Portal

- Provides a GUI interface to our Web Services.
- A method to trivially run Web Services with point-and-click.
- Based on Gridsphere technology.



The Google/InChI Web Service

A Web Service based at our Portal which allows users to search the Web by drawing a 2D structure.



Searching…

World Wide Molecular Matrix - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://wwmm.ch.cam.ac.uk/gridsphere/gridsphere?cid=inchigoogle&Java

Getting Started Latest Headlines

World Wide Molecular Matrix

Home | DB | INChI | CMLRSS | Babel | JUMBO | GAMESS@WORK | MACiE | ChemDraw | Mopac | Condor | WS Computing | Test

Introduction | Generate INChI | Structure Search

Search the Web Using InChI(TM) and Google(TM) (Used: **818** times)

This Service searches the public Web for molecules whose InChIs have been indexed by Google. See the InChI FAQ

INChI version: 1

File Edit View Insert Tools Help

H C N O F React Select Erase Paste Undo Redo Zoom

- + P S Cl

More Br I

- Create a molecule in the drawing space of the applet. (If you want a test, select naphthalene (the 2-ring molecule) and drag it to the drawing area)
- click "Search" ("Lucky search" gives the first hit only)
- You should get a list of links to pages containing that molecular structure
- If you get no hits, that is because no web page contaiining that InChI has been indexed yet - see FAQ. Check that naphthalene works (normally 3 or more hits).

Results

Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://wwmm.ch.cam.ac.uk/gridsphere/jsp/inchigoogle.jsp

Go

Getting Started Latest Headlines

Google Search Results:

The estimated total number of results returned for the query: 4

The query string that generated this result: 1/C10H8/c1-2-6-10-8-4-3-7-9(10)5-1/h1-8H

═══Result elements (4) ═══

1. The size in kilobytes of the cached version:

   URL: http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=931

2. The size in kilobytes of the cached version:

   URL: http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=2998

3. The size in kilobytes of the cached version: 98k

   URL: http://wwmm.ch.cam.ac.uk/inchifaq/

4. The size in kilobytes of the cached version: 234k

   URL: http://potency.berkeley.edu/chemical-structures.text

Done

Conclusion

We therefore provide an infrastructure of distributable components where robots can:

- read journals,
- extract molecules,
- compute their properties and,
- publish them to newsfeeds and Open repositories.

Thanks

- Peter MR, Yong Zhang and Joe Townsend.
- The InChI team - Steve Heller, Steve Stein, Dmitrii Tchekovskoi and Alan McNaught.
- The Taverna team – Tom Oinn et al.
- EPSRC is thanked for funding.

Links

- Group HomePage – http://wwmm.ch.cam.ac.uk
- WWMM Portal – http://wwmm.ch.cam.ac.uk/gridsphere/gridsphere
- DSpace – http://www.dspace.cam.ac.uk
- InChI FAQ – http://wwmm.ch.cam.ac.uk/inchifaq
- InChI application – http://www.iupac.org/inchi/license.html