Indexing the Chemical Semantic Web

Nick Day, Peter Murray-Rust.

Unilever Centre for Molecular Science Informatics, Chemistry Department, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK.

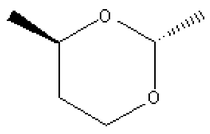
Unilever Centre for Molecular Informatics

The Chemical Semantic Web (CSW) is a global collection of structured, indexed Open chemistry

The CSW consists of XML documents (in CML, Chemical Markup Language) which are understandable by modern informatics tools such as Google, MSN, RSS (Blogs) and RDF. Science is now being archived in ***Institutional Repositories (IRs)*** with metadata standards such as OAI-PMH (Open Archives Initiative). Molecular metadata is provided by the new IUPAC identifier, InChI. We have shown that for structure-searching, this approach can replace traditional chemical aggregation and search methodologies.

The IUPAC International Chemical Identifier[1]

InChI is a non-proprietary unique identifier for chemical structures, including tautomerism, mobile hydrogens, stereochemistry (different isomers, Figure 1) and isotopes.



Absolute chirality known: InChI=1/C6H12O2/c1-5-3-4-7-6(2)8-5/h5-6H,3-4H2,1-2H3/t5-,6-/m1/s1
Absolute chirality unknown: InChI=1/C6H12O2/c1-5-3-4-7-6(2)8-5/h5-6H,3-4H2,1-2H3/t5-,6-/s2
Racemic mixture: InChI=1/C6H12O2/c1-5-3-4-7-6(2)8-5/h5-6H,3-4H2,1-2H3/t5-,6-/s3

We have shown that Google and MSN can index InChIfied molecules in Institutional Repositories (eCrystals at Southampton, DSpace at Cambridge) with almost 100% recall and almost 100% precision[2]

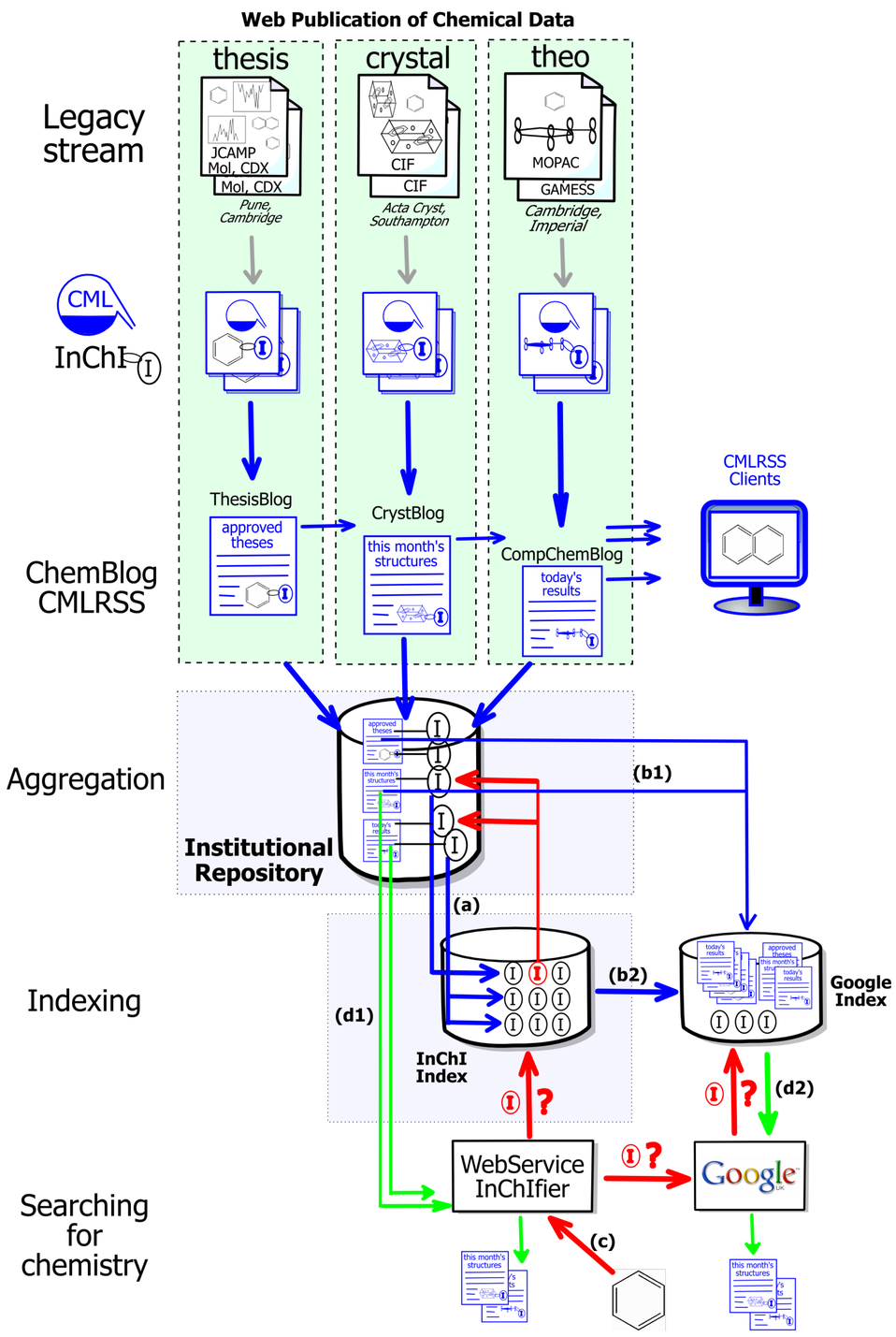


Figure 2.
Legacy chemical data streams (black) converted to CML (blue), InChIfied (I) and blogged (CMLRSS) read by news clients or ingested into IRs. (a)IR-blogs indexed by InChI. (b1) IRs and (b2) InChI Index abstracted by Google. (c) Molecular query InChIfied (d1) retrieves blogs (XQuery, green) (d2) used by Google to retrieve blog. All CML-ised information displayable as hyperactive molecules (e.g. Jmol).

Publishing to the CSW

We provide Open tools and Web Services for converting conventional chemistry (molecules and data) to InChI-indexed XML documents. Figure 2 shows the flow with documents further converted to "Chemical Blogs" which can either be "news-fed" to subscribers or ingested by OAI-PMH-compliant repositories These store the blogged data, including a complete CML representation of spectra, crystallographic experiments, computational chemistry, etc. A smaller separate index of InChIs points to each blog in which that InChI is mentioned.

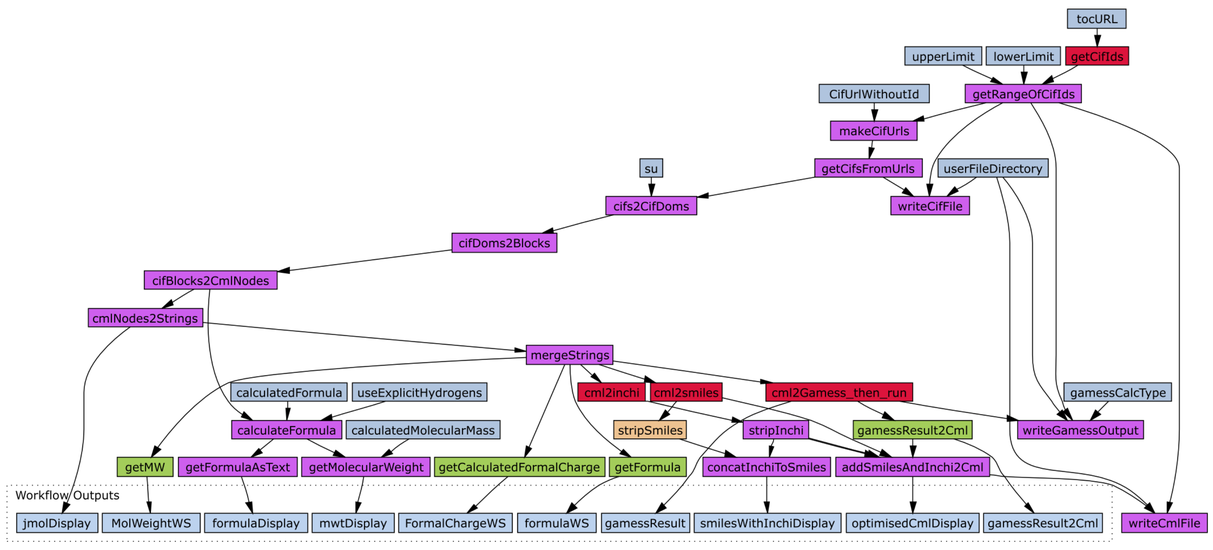


Figure 3. *Workflow which automatically extracts CIF files from the Acta Crystallographica website, converts to CML and InChI, computes GAMESS structures and deposits in an IR.*

Querying the CSW

Documents and metadata in XML are supported by XQuery engines (eXist) and modern search engines (Google, MSN) all of which use full-text indexing. Queries include strings, numeric data and structured XML.

XPath Example:

```
//molecule[atomArray/atom[@elementType='Fe']
and count(atom) < 10]
```

will find all CML molecules containing Fe and less than 10 atoms

Google Example:

`"1/C6H6/c1-2-4-6-5-3-1/h1-6H"`

finds all example of InChIfied benzene on the web.
We are extending this to spectra and reactions

Services and tools

All tools are Open and downloadable
Home page: http://wwmm.ch.cam.ac.uk
Web Services: http://wwmm.ch.cam.ac.uk/gridsphere/gridsphere
Downloads: http://wwmm.ch.cam.ac.uk/moin/SoftWare
DSpace: http://www.dspace.cam.ac.uk/handle/1810/724
Email: ned24@cam.ac.uk, pm286@cam.ac.uk

References

(1) Unofficial InChI FAQ: http://wwmm.ch.cam.ac.uk/inchifaq

(2) Simon J. Coles, Nick E. Day, Peter Murray-Rust, Henry S. Rzepa, Yong Zhang, "Enhancement of the chemical semantic web through the use of InChI identifiers", *Organic & Biomolecular Chemistry*, 2005, 3(10), 1832 - 1834, DOI:10.1039/b502828k

Acknowledgements

•Dr. Yong Zhang
•The InChI team (Steve Heller, Steve Stein and Dmitrii Tchekovskoi)