COMMUNICATION

OBC
www.rsc.org/obc

Enhancement of the chemical semantic web through the use of InChI identifiers†

Simon J. Coles,[a] Nick E. Day,[b] Peter Murray-Rust,[b] Henry S. Rzepa[c] and Yong Zhang[b]

[a] *School of Chemistry, University of Southampton, Southampton, UK SO17 1BJ*

[b] *Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW*

[c] *Department of Chemistry, Imperial College, London, UK SW7 2AY*

Received 24th February 2005, Accepted 6th April 2005
First published as an Advance Article on the web 18th April 2005

Molecules, as defined by connectivity specified *via* the International Chemical Identifier (InChI), are precisely indexed by major web search engines so that Internet tools can be transparently used for unique structure searches.

The Chemical Semantic Web[1] is a vision where all chemical information is immediately accessible and processible by semantically aware tools based on W3C protocols.[2] Full-text based search engines are universal on the Open Web and their indexes act as an effective knowledge base for much personal and corporate use. Specialist engines such as CiteSeer[3] and Google Scholar™[4] which index citations in those Scientific–Technical–Medical (STM) articles available to them, are recently introduced enhancements for traditional abstracting services. However the quality of search retrieval from such sources is still regarded as patchy, in part because metadata (data about data and its relationships) for semantic enhancement of the Web-based document content is still sparse. A good (but still rare) example is Swoogle,[5] which focuses on searching the Semantic Web specifically based on metadata. Only in some disciplines (*e.g.*, bioinformatics) is content near-comprehensive. In this article, we suggest one mechanism for enabling molecular information to become similarly universal and accessible on the public Web.

About 1.5 million new compounds are published each year and many more existing ones are mentioned, often with new measurements of properties. Traditionally these are abstracted from primary publications and aggregated in centralised databases. If all these molecules were semantically marked and published on the web, then the Internet could transparently become a global knowledge base for chemical information, with much less human effort in secondary data aggregation. Indeed, if molecular information could be reliably indexed by existing search engines, this would automatically create a base index for a Chemical Semantic Web. The new InChI identifier[6] provides exactly this for pure chemical compounds[7] for which a connection table exists, in the form of a unique text string (Table 1).

An InChI encodes structure, including stereochemistry, isotopes and tautomers in a lossless manner. This can then be translated back to the connection table, so a separate registry is not required for identifier resolution. The code to create InChIs is Open[6] and in addition we have created an Open GUI and Web services for InChI creation, based on Chemical Markup Language (CML).[9] Both authors and readers can use these InChI-enhanced tools in their normal environment. The InChI itself can be layered (only some of the layers are shown here) and high layers can be omitted if the information is imprecise or if one wants to allow isomers to give exact matches. Here we report how typical free-text search engines process, index and retrieve such InChI strings (Fig. 1 and the supplementary information).

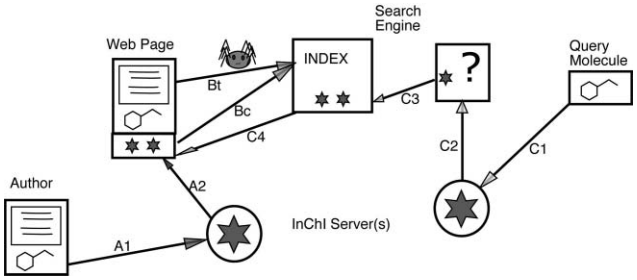

Fig. 1 Molecules translated to text (InChIs) in the Chemical Semantic Web using the sequence: molecules (A1) converted to InChI text are (A2) merged with an article published on the Web. A robot indexes text (Bt) and InChI identifiers (Bc). A normal chemical query (C1) converted to an InChI (C2) and submitted to query engine (C3). The InChI string is located in index (C4) which returns target URL for retrieval.

Because there has hitherto been very little Web deployment of InChIs, we know precisely how many InChIs we have published, and have recorded when search engines have discovered them.[10] Our corpora included 104 crystal structures published under the eCrystals/eBank project.[11] Each structure in CIF format was converted to CML[12] and thence to InChI. XHTML and CML files containing the InChI were published under an OAI interface[13] and retrieved from the Web server. Each InChI string, having no white space, is treated as a single word during the indexing process, and is broken down into smaller tokens defined by delimiting characters such as '/' (solidus), ',' (comma),

† Electronic supplementary information (ESI) available: detailed analysis of the strategies of search engines and their similarities and differences. See http://www.rsc.org/suppdata/ob/b5/b502828k/

Table 1 InChI strings for organic compounds[8]

Molecule	InChI	Comment
CH3C(=O)F	InChI = 1.12Beta/C2H3FO/c1-2(3)4/h1H3	
FCH2C(=O)OH	InChI = 1.12Beta/C2H3FO2/c3-1-2(4)5/h1H2,(H,4,5)	
CH3C(=O)NH2	InChI = 1.12Beta/C2H5NO/c1-2(3)4/h1H3,(H2,3,4)	Note possible tautomerism
$CH3C(=O)O^-$	InChI = 1.12Beta/C2H4O2/c1-2(3)4/h1H3,(H,3,4)/p-1	Ionized proton denoted by p
Lactic acid	InChI = 1.12Beta/C3H6O3/c1-2(4)3(5)6/h1H3,2H,4H,(H,5,6)	Undefined stereocentre
Lactic acid	InChI = 1.12Beta/C3H6O3/c1-2(4)3(5)6/h1H3,2H,4H,(H,5,6)/t2-/m1/s1	Defined stereocentre
Mauveine	InChI = 1.12Beta/C26H22N4/c1-17-8-10-19(11-9-17)28-20-12-13-23-25(15-20)30(21-6-4-3-5-7-21)26-16-22(27)18(2)14-24(26)29-23/h1-2H3,3-16H,(H2,27,28)/p + 1	Cationic component

DOI: 10.1039/b502828k

Table 2 Glossary of terms

Term	Meaning
W3C	World Wide Web Consortium (www.w3.org). An international industry consortium that develops standards for the Web
CML	Chemical Markup Language and XML Schema for Chemistry
GUI	Graphical User Interface, a user interface based on graphics that uses a mouse as well as a keyboard as an input device
MSN	Microsoft Network is an internet service provider and web portal created by Microsoft
XML	eXtensible Markup Language, W3C meta-language for specifying specific markup languages
XHTML	A general purpose display markup language conforming to XML syntax
RSS	Rich Site Summary or RDF Site Summary, an XML format for news and content syndication
SVG	Scalable Vector Graphics, an XML vector graphics format from the W3C for the Web

'-' (hyphen), ':' (colon) and '.' (full stop). A search of the resulting index proceeds with a Boolean 'AND' between all the tokens. Table 2 provides a glossary of related terms.

The results (Table 3) are measured by recall (proportion of pages found) and false positives (retrieval of non-InChIs and the wrong InChI). The corpus of 104 molecules included one pair of diastereomers, and because stereochemistry was not initially included in eBank, there is one inter-InChI collision. This recall and precision is very encouraging, and even InChIs for 10 small molecules (water, methanol, acetic acid, *etc.*) showed 0% non-InChI contamination, as the leading strings are very discriminating. The main limitation is the maximum word length defined by each search engine, which corresponds to about 10 tokens per InChI (with, *e.g.*, Google). Consequently, large isomers (comprising more than 10 tokens) may be co-retrieved. This noise can, of course, be completely filtered out by further searching the retrieved document. Where the configuration of the index/search engine is under our control (as for example the htDig software[14]), the maximum word length can be reset from the default (60 characters) to a much larger value to allow arbitrarily long InChI strings to be uniquely retrieved. In our tests, we set this word length to 255 characters, which covers small and medium sized molecules; if indexing of, *e.g.*, larger (bio)molecules were required, appropriately longer word lengths would have to be set.

A larger collection of 9591 molecules from the current collection of KEGG molecules[15] were converted to CML, InChIs were computed and the results published on a Web server. As of 14 November 2004, Google appeared to have indexed the first 4870 (C00001–C07576); a week later a further 300 molecules had been indexed. Similar gradual increase in the coverage was also noted for the Yahoo and MSN indices. Searches with a subsample of InChIs showed no non-InChI collisions.

Our final test involved adapting one of the "Molecules-of-the-Month" sites.[16] Here, a single InChI string (corresponding to mauveine) was embedded into a variety of document types comprising both XML, (XHTML, CML, RSS, SVG)[2] and non-XML (MDL Molfile, Acrobat) types, and indexed using the ChemDig variant,[17] using a maximum word length of 255 characters. This allowed completely specific retrieval of the string representing (the cationic component) of mauveine for each document containing it. In particular, the token for protonation count (p + 1) (and, *e.g.*, stereochemical information), which comes at the end of the InChI and which may be truncated during indexing by the public search engines, was reliably retained using ChemDig/htDig. The htDig engine can also be set to concurrently index all the above sites. The process takes around 10 min for, *e.g.*, 10,000 documents containing InChI molecule descriptors, suggesting a global trawl for, *e.g.* 5000, sites is entirely feasible.

How do other chemical descriptors compare in Web search retrieval? CAS registry numbers (tested with caffeine "58-08-02" and acetic acid "64-19-7") showed 20% non-CAS collisions (*e.g.*, ringtones, football scores) in AltaVista and Yahoo and only about 70% recall if the string "CAS" was included to reduce this. We found 7 syntactic variants of unique SMILES strings[18] and hence these have a low recall for any given structure.

We conclude that InChI-based text searching can be used for precise matches, or more fuzzy searches when the auxilliary InChI layers are omitted. Chemical substructures of a molecule do not map to InChIs which are substrings of the parent InChI. However we have prototyped algorithms for creating sub-InChIs which can be used for text-based substructure searching and are investigating how they can be deployed.

Even without the active collaboration of the search engine companies, the use of InChIs and CML provides a powerful base for indexing the molecular web. Many search engines preferentially index XML documents (*e.g.*, CML) rather than legacy (*e.g.*, MDL molfile). Since a growing number of publishers submit articles (even when non-Open) to engines such as Google™ for indexing, any InChIs included in text will also be indexed. If publishers allow, or encourage, such inclusion then a high proportion of published molecules will be indexed at source, without corruption, and almost immediately on publication. If search engines collaborate by recognising InChIs and indexing them completely (*i.e.*, by increasing the maximum word length), then precision and recall will be total.

We therefore urge authors and publishers to include InChIs in the full text and supplemental data of manuscripts. Lists of InChIs on departmental or publisher web sites would also be indexed and can be linked to citations, thus exposing the research to much higher visibility. The effort has no monetary cost and is considerably less time-consuming than creating structure diagrams and analytical material, and we contend it as a remarkably effective way of enhancing the world's chemical knowledge.

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.

We thank Steve Stein and Dmitrii Tchechovskoi for early versions of the InChI software, Geoff Hutchison for advice on htDig and the DTI/EPSRC eScience program for support.

References

1 G. Gkoutos, P. Murray-Rust, H. S. Rzepa and M. Wright, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1124; for reviews covering a ten year period, see:; H. S. Rzepa, B. J. Whitaker and M. J. Winter, *J. Chem. Soc., Chem. Commun.*, 1994, 1907; P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell and Y. Zhang, *Org. Biomol. Chem.*, 2004, **2**, 3192–3203.

2 For documentation see http://www.w3.org/2001/sw/.

3 See http://citeseer.ist.psu.edu/.

4 See http://scholar.google.com/.

5 See http://pear.cs.umbc.edu/swoogle/.

6 S. E. Stein, S. R. Heller, and D. Tchekhovski, *An Open Standard for Chemical Structure Representation—The IUPAC Chemical Identifier*, 2003, Nimes International Chemical Information Conference Proceedings, pp 131–143; S. E. Stein, S. R. Heller and D. V. Tchekhovskoi, *Abstracts of Papers, 222nd ACS National Meeting*, Chicago, IL, August 26–30 2001, CINF-005. See http://www.iupac.org/projects/2000/2000-025-1-800.html for information on how to acquire the code.

7 The continuing InChI development will address polymers, transition states, mixtures and reactions.

8 We have compiled an extensive FAQ and tutorials which address many of the new concepts in InChI at http://wwmm.ch.cam.ac.uk/inchfaq/index.html.

9 Available as free services at http://wwmm.ch.cam.ac.uk/gridsphere/gridsphere?cid=generateinchi&JavaScript=enabled.

10 P. Murray-Rust, H. S. Rzepa and Y. Zhang, *W3C Workshop on Semantic Web for Life Sciences, 27–28 October 2004*, Cambridge, MA. See http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/.

11 See http://ecrystals.chem.soton.ac.uk/.

12 P. Murray-Rust and H. S. Rzepa, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 757–72; http://cml.sourceforge.net/ and http://www.ukoln.ac.uk/projects/ebank-uk/schemas/.

13 S. J. Coles, *eScience All Hands Meeting*, Nottingham, UK, 31 Aug–3 Sep 2004, London, UK, EPSRC, ISBN:1904425216, 37 pp (reprint available at http://eprints.soton.ac.uk/9103/). See also http://www.ukoln.ac.uk/projects/ebank-uk/.

14 See http://htdig.sourceforge.net/.

15 M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, *Nucleic Acids Res.*, 2002, **30**, 42–46 and earlier references cited therein. See also http://www.genome.jp/kegg/ligand.html.

16 See http://www.ch.ic.ac.uk/motm/.

17 G. V. Gkoutos, C. Leach and H. S. Rzepa, *New. J. Chem.*, 2002, 656–666. See http://www.ch.ic.ac.uk/rzepa/inchi/ for access to the InChI index/search.

18 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.