The IUPAC InChI project – CSIRO presentation

Stephen Heller
InChI Trust Project Director
steve@inchi-trust.org

Outline

1. Why InChI?
2. How does the InChI work?
3. Future developments for InChI
4. InChI Trust
5. Acknowledgements

Objective

The IUPAC Chemical Identifier (InChI) is an open source, freely available, non-proprietary identifier for well defined chemical substances.

The InChI enables chemical information in electronic data sources (databases, registries, journals and repositories) to be machine readable.

Enabling easier LINKING of, and working with, diverse data and information compilations.

Why use InChI

As the InChI encodes the exact structure of the chemical, the article, database entry, catalogue etc… can be searched as the user may want to search or linked into the wider information space.

It enables new discoveries from existing information and data by easily being able to integrate, remix, and retell.

InChI is a vital part of new business models and technologies involving chemicals that will lead to new discoveries.

Why use InChI

There are different identifiers out there BUT InChI is the only one that is:

- Non-proprietary
- Encodes the exact structure of a chemical
- Is inter-operable and portable

And most importantly is:

- Unique to the structure it describes…

Why use InChI

Re: [CHMINF-L] Inchi and chemical databases

You forwarded this message on 9/15/2010 5:37 PM.

CHEMICAL INFORMATION SOURCES DISCUSSION LIST [CHMINF-L@LISTSERV.INDIANA.EDU] on behalf of Ian A Watson

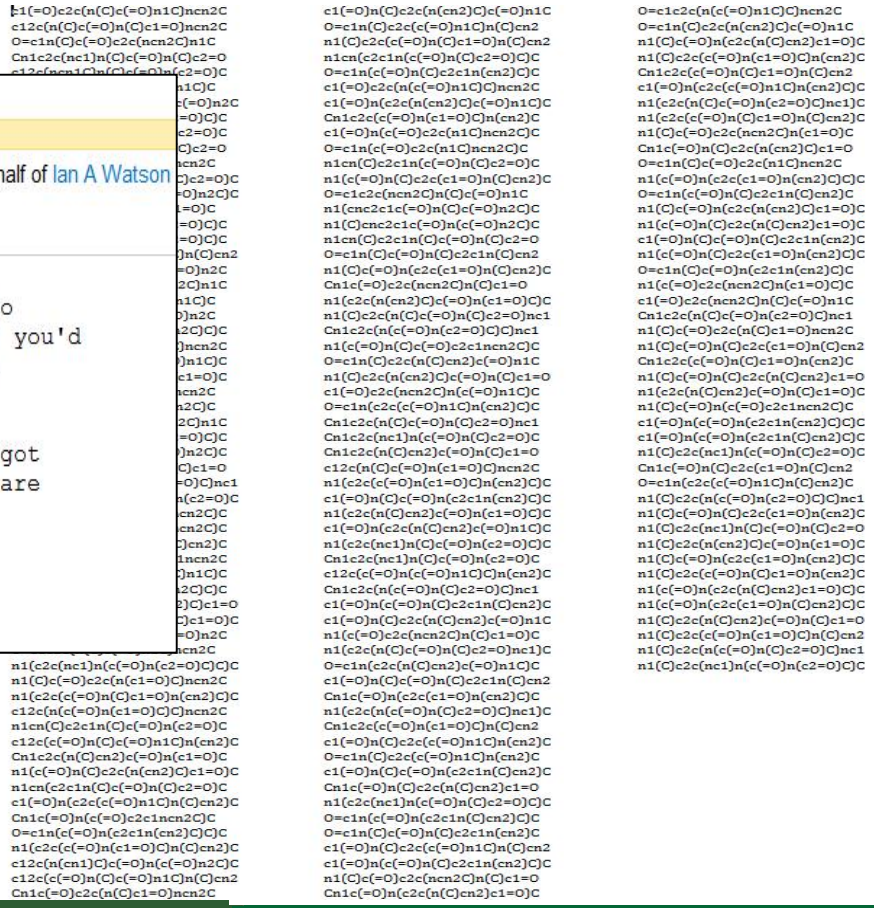
Sent: Wednesday, September 15, 2010 3:24 PM

To: CHMINF-L@LISTSERV.INDIANA.EDU

Interesting example of Caffeine smiles on the web site. I was able to generate 172 different smiles for the Caffeine molecule (email me if you'd like them). Presumably each one of these could be a unique smiles in somebody's implementation.

But when I converted each of those 172 different smiles to InChI, I got the exact same InChI string for each one. That's exactly how things are supposed to work. Nice.

Ian Watson



InChI TRUST

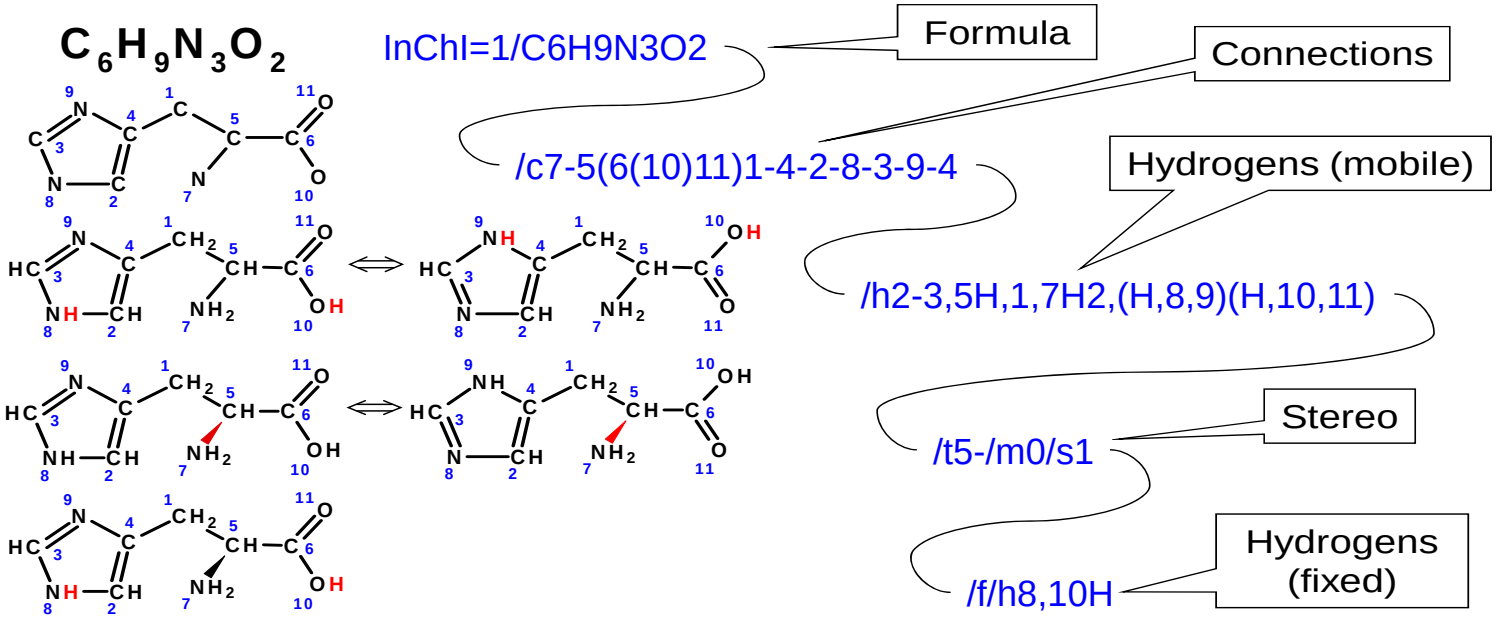
How does the InChI work?

The InChI has a layered structure design with the current layers being:

1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond (Z/E)
    b. tetrahedral (sp3)
5. Tautomers (on or off)
6. Charges are added to end of the string

How does the InChI work?

InChI Layers: L-Histidine



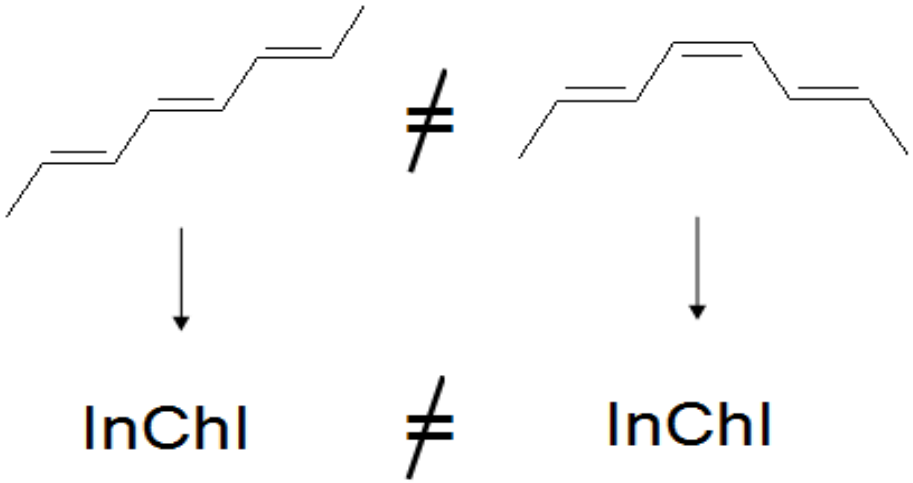
InChI=1/C6H9N3O2/c7-5(6(10)11)1-4-2-8-3-9-4/h2-3,5H,1,7H2,(H,8,9)(H,10,11)/t5-/m0/s1/f/h8,10H

InChIKey=HNDVDQJCIGZPNO-QLMCEAFFNA-N InChIKey=HNDVDQJCIGZPNO-YFKPBYRVSA-N

InChI TRUST

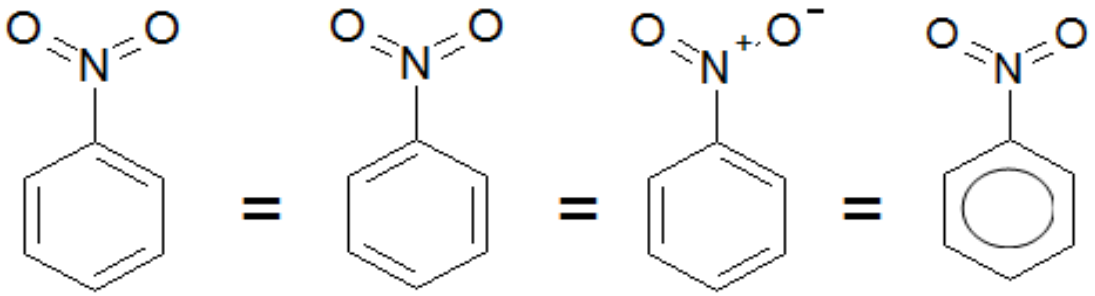
How does the InChI work?

Different geometric or stereo isomers have different identifiers – all distinguishing structural information is included

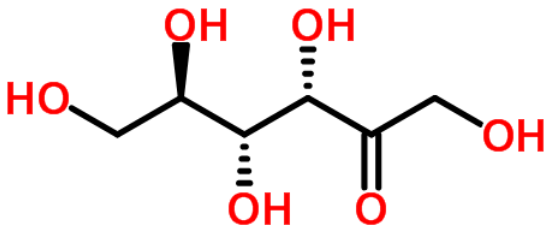


How does the InChI work?

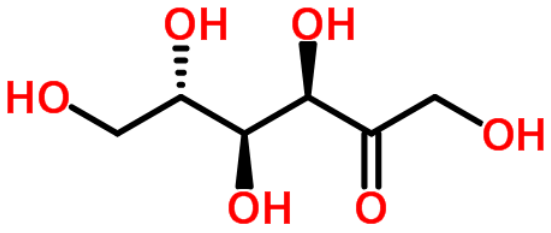
One compound has only ONE InChI



How does the InChI work?



D-Fructose



L-Fructose

D-Fructose (Natural)

InChI=1S/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h3,5-9,11-12H,1-2H2/t3-,5-,6-/m1/s1

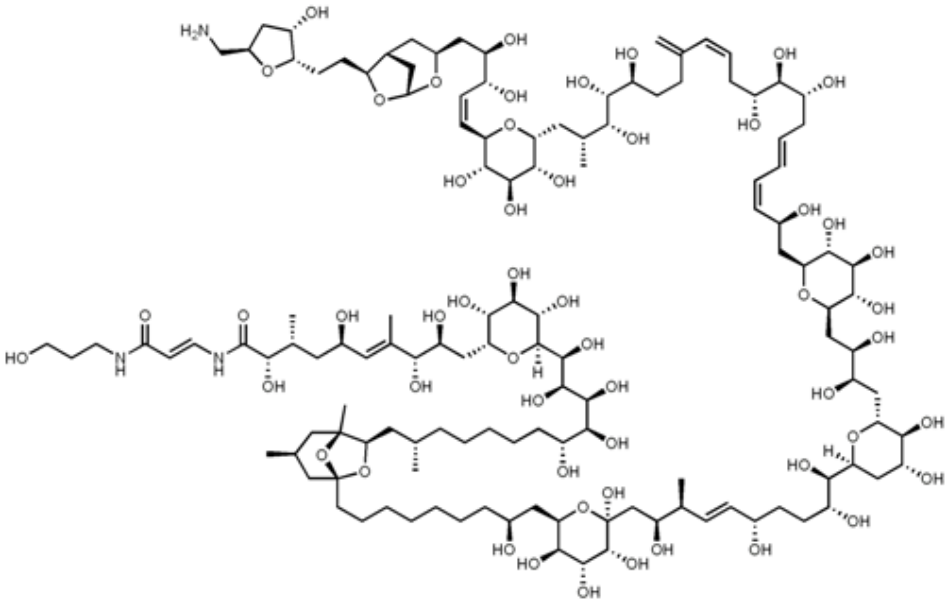
InChIKey: BJHIKXHVCXFQLS-UYFOZJQFSA-N

L-Fructose

InChI=1S/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h3,5-9,11-12H,1-2H2/t3-,5-,6-/m0/s1

InChIKey: BJHIKXHVCXFQLS-FUTKDDECSA-N

How does the InChI work?



Palytoxin

- Isolated from Hawaiian soft coral
- One of the most toxic non-peptide substances
- Contains more than 70 stereochemical elements

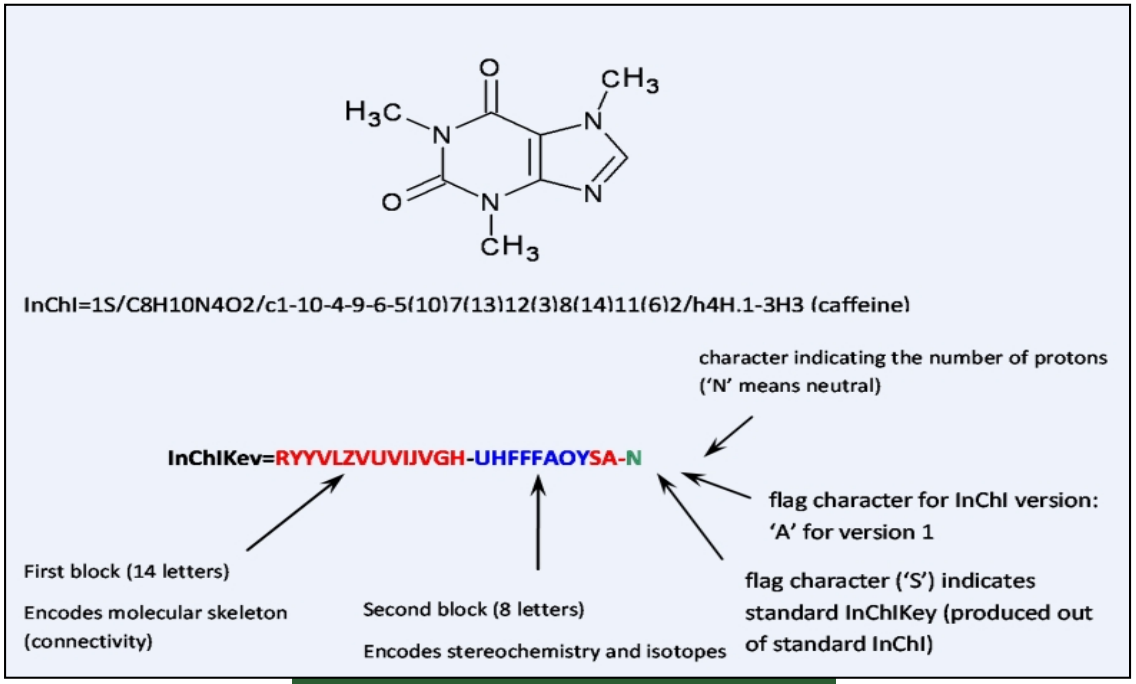
InChI=1S/C129H223N3O54/c1-62(29-33-81(143)108(158)103(153)68(7)47-93-111(161)117(167)110(160)91(180-93)36-35-76(138)82(144)51-73-50-74-53-92(178-73)90(177-74)38-37-89-85(147)52-75(61-130)179-89)23-20-28-78(140)105(155)77(139)26-18-13-16-25-70(135)48-94-112(162)118(168)113(163)97(181-94)55-84(146)83(145)54-95-107(157)87(149)57-96(182-95)106(156)80(142)34-32-69(134)31-30-65(4)88(150)60-129(176)125(174)123(173)115(165)99(184-129)49-71(136)24-15-10-9-11-19-40-128-59-64(3)58-127(8,186-128)100(185-128)44-63(2)22-14-12-17-27-79(141)109(159)116(166)120(170)122(172)124-121(171)119(169)114(164)98(183-124)56-86(148)102(152)66(5)45-72(137)46-67(6)104(154)126(175)132-42-39-101(151)131-41-21-43-133/h13,16,18,20,23,25,30-31,35-36,39,42,45,63-65,67-100,102-125,133-150,152-174,176H,1,9-12,14-15,17,19,21-22,24,26-29,32-34,37-38,40-41,43-44,46-61,130H2,2-8H3,(H,131,151)(H,132,175)/b18-13+,23-20-,25-16-,31-30+,36-35-,42-39+,66-45+/t63-,64-,65-,67+,68+,69+,70+,71-,72-,73-,74+,75-,76+,77+,78+,79+,80+,81-,82+,83+,84+,85+,86-,87+,88-,89+,90+,91+,92-,93+,94-,95+,96-,97+,98+,99+,100+,102+,103+,104-,105-,106+,107-,108+,109-,110+,111-,112-,113+,114-,115-,116-,117-,118+,119+,120+,121-,122-,123+,124-,125+,127+,128-,129-/m0/s1

InChIKey: CWODDUGJZSCNGB-HQNRRURTSA-N

InChI TRUST

How does the InChI work?

The InChI can be long but the shorter ‘hashed’ InChIKey retains all the information of the InChI.



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

About 5,780 results (0.36 seconds) Advanced search

Everything
Images
Videos
More

Show search tools

InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H**,1-3H3 ... reveals an inhibitor of **Mre11**-Rad50-Nbs1 complex , Nature Chemical Biology, 2008 ...
www.chemspider.com/InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW - Cached - Similar

Caffeine - Wikipedia, the free encyclopedia
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. InChI key, RYYVLZVUVIJVGH-UHFFFAOYAW. Properties. Molecular formula, C8H10N4O2 ...
en.wikipedia.org/wiki/Caffeine - Cached - Similar

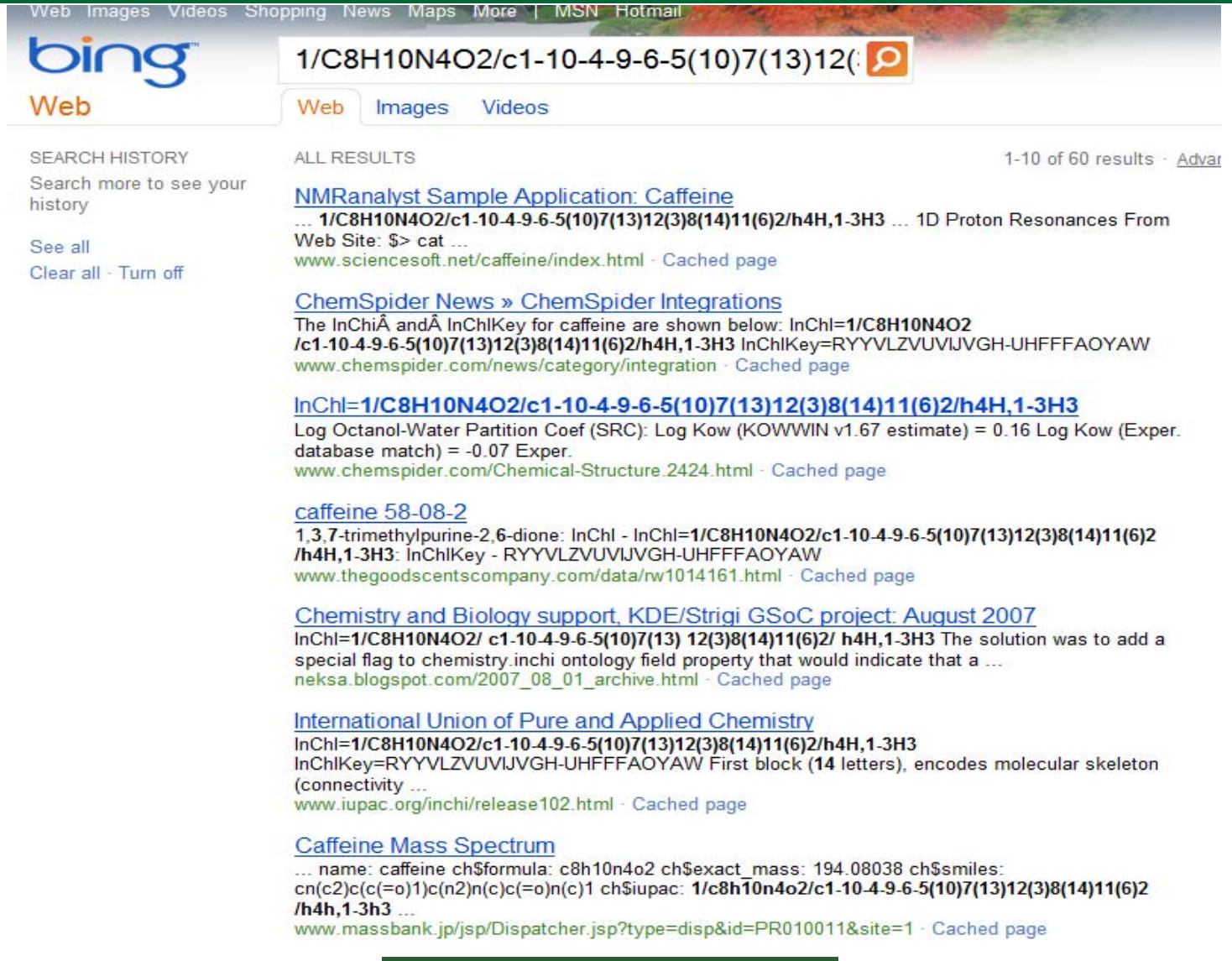
Compound 7 : Moonlighting proteins Hal3 and Vhs3 form a ...
Nov 1, 2009 ... InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3**. InChIKey: RYYVLZVUVIJVGH-UHFFFAOYAW ...
www.nature.com › Journal home › Archive › Article › Full text

caffeine (CHEBI:27732)
Oct 17, 2009 ... InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3**. InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3** ...
www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:**27732** - Cached

InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3**. ... reveals an inhibitor of Mre11-Rad50-Nbs1 complex , Nature Chemical Biology, 2008 ...
mesh.chemspider.com/Chemical-Structure.2424.html - Cached

Caffeine Mass Spectrum
CH$NAME: Caffeine CH$FORMULA: C8H10N4O2 CH$EXACT_MASS: 194.08038 CH$SMILES: ... CH$IUPAC: **1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3** ...
www.massbank.jp/jsp/Dispatcher.jsp?type=disp&id...**1** - Cached - Similar

caffeine 58-08-2
Aug 3, 2010 ... IUPAC Name -, 1,3,7-trimethylpurine-2,6-dione. InChI -, InChI=**1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3** ...
www.thegoodscentscompany.com/data/rw**1014161**.html - Cached - Similar


Web Images Videos Shopping News Maps More | MSN Hotmail
bing
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(
Web
Web Images Videos
SEARCH HISTORY
Search more to see your history
See all
Clear all · Turn off
ALL RESULTS
1-10 of 60 results · Advar
NMRanalyst Sample Application: Caffeine
... 1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 ... 1D Proton Resonances From Web Site: $> cat ...
www.sciencesoft.net/caffeine/index.html · Cached page
ChemSpider News » ChemSpider Integrations
The InChiÂ andÂ InChIKey for caffeine are shown below: InChI=1/C8H10N4O2 /c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW
www.chemspider.com/news/category/integration · Cached page
InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
Log Octanol-Water Partition Coef (SRC): Log Kow (KOWWIN v1.67 estimate) = 0.16 Log Kow (Exper. database match) = -0.07 Exper.
www.chemspider.com/Chemical-Structure.2424.html · Cached page
caffeine 58-08-2
1,3,7-trimethylpurine-2,6-dione: InChI - InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2 /h4H,1-3H3: InChIKey - RYYVLZVUVIJVGH-UHFFFAOYAW
www.thegoodscentscompany.com/data/rw1014161.html · Cached page
Chemistry and Biology support, KDE/Strigi GSoC project: August 2007
InChI=1/C8H10N4O2/ c1-10-4-9-6-5(10)7(13) 12(3)8(14)11(6)2/ h4H,1-3H3 The solution was to add a special flag to chemistry.inchi ontology field property that would indicate that a ...
neksa.blogspot.com/2007_08_01_archive.html · Cached page
International Union of Pure and Applied Chemistry
InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW First block (14 letters), encodes molecular skeleton (connectivity ...
www.iupac.org/inchi/release102.html · Cached page
Caffeine Mass Spectrum
... name: caffeine ch$formula: c8h10n4o2 ch$exact_mass: 194.08038 ch$smiles: cn(c2)c(c(=o)1)c(n2)n(c)c(=o)n(c)1 ch$iupac: 1/c8h10n4o2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2 /h4h,1-3h3 ...
www.massbank.jp/jsp/Dispatcher.jsp?type=disp&id=PR010011&site=1 · Cached page

InChI TRUST

CROSSREF LABS

You have been warned....

Home

INCHI LOOKUP

The idea is to create a mechanism that would allow CrossRef publishers to record InChIs in their submitted CrossRef metadata. This, in turn, would allow us to provide a service that would allow users to:

- Lookup the published articles that mention a particular InChI.
- Lookup the InChIs mentioned in a published article.

Similar services could, conceivably, be provided for other types of semantic metadata.

The following is a demonstrator of what an DOI2InChI lookup service might look like. Please note that the XML representation of the results is very basic and is not best-practice for linked-data.

The demonstrator currently only holds DOIs and InChIs for a few publishers.

A summary of the contents of the database can be found on the status page
http://inchi.crossref.org/status

A list of all the CrossRef DOIs that contain InChIs can be seen here:
http://inchi.crossref.org/dois

A list of all the InChIs that have been registered with CrossRef can be seen here:
http://inchi.crossref.org/inchis

The system provides the following API calls:

Return all the DOIs that have been registered with a given InChI
http://inchi.crossref.org/dois/InChI=1S/C4H6O2/c1-3-6-4(2)5/h3H,1H2,2H3

Return all the InChIs that have been registered for a given DOI
http://inchi.crossref.org/inchis/10.1038/nchem.215

InChI TRUST

Future development

There are working groups looking at InChI extensions for:

Markush	(results expected 2011)
Polymers/Mixtures	(results expected 2011)
InChI Resolver protocols	(results expected 2011)
Organometallics	(results expected 2012)
Electronic States	(results expected 2012)
RInChI –InChI for Reactions	(results expected 2012)

Future development

Transrutherfordium elements

Electronic states, including transition states & excited states

Work with IUCr for 3D information

Proteins, peptides & biopolymers

Mac supported version

Java version

VS2010 .NET compilation support

Members of the Trust can add to this list and, through the board, prioritise development priorities

The InChI Trust

- The InChI Trust, a not-for-profit organisation to expand and develop the InChI open source chemical structure representation algorithm.

- Membership of the InChI Trust helps support, shape and direct the Trust's ongoing development and maintenance of the InChI algorithm.

The InChI Trust - mission statement

The InChI Trust develops and supports the non-proprietary IUPAC InChI standard and promotes its uses to the scientific community.

The Trust's goal is to enable the interlinking and combining of chemical, biological and related information, using unique machine-readable chemical structure representations to facilitate and expedite new scientific discoveries.

The InChI Trust - IUPAC

In **March 2010** the Trust signed a heads of agreement with IUPAC setting out the conditions under which both the InChI Trust and IUPAC intend to further develop and maintain the InChI.

This was a vital step for both the Trust and the InChI.

The InChI Trust - progress

2009 – InChI Trust set up, first Board meeting in Sept

March 2010 – reports received on InChI recoding options and test suites from Digital Chemistry

April 2010 – www.inchi-trust.org was launched

June 2010 – v1.03 of the InChI software was released

July 2010 – InChI Trust contracted GGA Software Services LLC to develop software programs to test and certify that InChIs and InChIKeys are valid

Aug 2010 – first AGM

The InChI Trust - progress

2010 –InChI Trust provided funds for working groups examining Markush structures and polymers

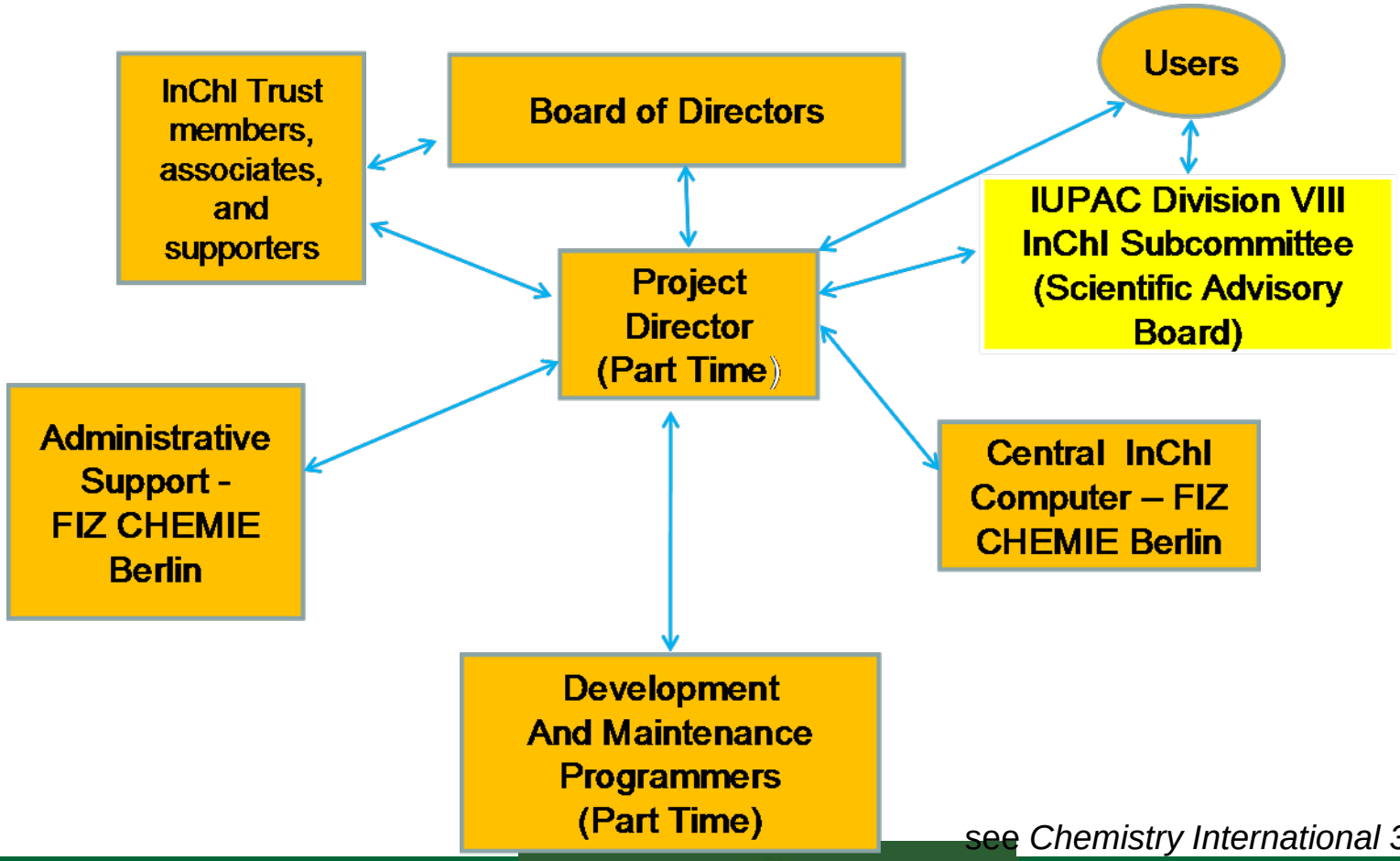
2010 – licensing and trademark issues addressed

2011 – budgeted funds to develop support for polymers & mixtures, Markush (following working group recommendations) and to fund full documentation project

2011 – further working group funding with IUPAC

2011 – funding for an ‘InChI community’ event around working group meetings

InChI Trust - organization



see *Chemistry International 32, 16-17 (2010)*

Current InChI Trust members

ACD/Labs
ChemAxon
Elsevier
FIZ CHEMIE – Berlin
IBM
Informa/Taylor & Francis
IUPAC
John Wiley & Sons
Microsoft
Nature Publishing Group
OpenEye
Royal Society of Chemistry (RSC)
Springer
Symyx
Thomson-Reuters

Current InChI Trust supporters

CalTech, CA, USA
Chem21, IL, USA
Indiana University, IN, USA
Leadscope, OH, USA
National Chemical Laboratory, Pune, India
National Institute of Chemistry, Ljubljana, Slovenia
SharePoint, WA, USA
Trinity University, TX ,USA
Unilever Centre for Molecular Science Informatics, Cambridge UK
University of Applied Science, Gelsenkirchen , Germany
University of California – Riverside, CA, USA
University of California – San Francisco, CA, USA
University of North Carolina, NC, USA
University of the West Indies, Mona, Jamaica
Xemistry GmbH, Germany

Why join the InChI Trust?

- Influence and decide priorities for developments of the InChI standard
- Regular bulletins on progress
- Advice on implementation
- Access to developed tools
- Involvement in the community of other InChI users, developers and implementers

Why the InChI Trust needs you

Help us to develop the InChI for the future

- To expand the standard
- To widen applications and use
- To have you influence the priorities
- To link chemistry publications and resources across the web

Acknowledgements

Primarily members for the IUPAC InChI subcommittee and associated InChI working groups:

Steve Bachrach, Colin Batchelor, John Barnard ,Evan Bolton, Steve Boyer, Steve Bryant, Szabolcs Csepregi ,Rene Deplanque, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau , Marc Nicklaus, Carmen Nitsche, Matthias Nolte , Igor Pletnev, Josep Prous, Hinnerk Rey, Ulrich Roessler, Roger Schenck , Martin Schmidt, Steve Stein, Peter Shepherd, Markus Sitzmann, Chris Steinbeck, Keith Taylor, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, Tony Williams, Andrey Yerin.

Special Acknowledgement:

Ted Becke & Alan McNaught for their vision and leadership of the future of IUPAC nomenclature.