The IUPAC InChI project

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

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

http://www.inchi-trust.org

The slides from this presentation can be found at:

http://www.hellers.com/steve/pub-talks/

11/30/2010

Disclaimer

These slides were made from 100% recycled electrons.

This will be a well balanced presentation.
I have a chip on both shoulders.

I am professionally irreverent.

Outline

1. Background/History/Objective/Why InChI?
2. InChI Technical Details and Examples
3. InChI Trust
4. Current and Future InChI activities
5. Acknowledgements

Objective

The objective of the IUPAC Chemical Identifier Project is to create a unique label, the IUPAC Chemical Identifier (InChI), which will be an Open Source, freely available, non-proprietary identifier for well defined chemical substances that can be used in printed and electronic data sources thus enabling easier LINKING of and working with diverse data and information compilations.

Why InChI? - Too Many Identifiers

Structure diagrams
- various conventions
- contain ‘too much’ information

Connection Tables
- MolFiles, Smiles, ROSDAL, …

Pronounceable names
- IUPAC, CAS, trivial

Index Numbers
- EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAF

Why Use InChI

For publishers and database providers using InChI gives one a competitive advantage being able to LINK content from multiple sources. It offers users the ability to help in new discoveries from existing information and data by easily being able to integrate, remix, and retell. InChI is a small, but vital, part of new business models and technologies involving chemicals that will lead to new discoveries. Combinability increases the value of information and data.

InChI & Open Access

InChI is the way publishers are improving their product delivery and meeting their US Government mandated Open Access requirements by conveying supplemental information to users.

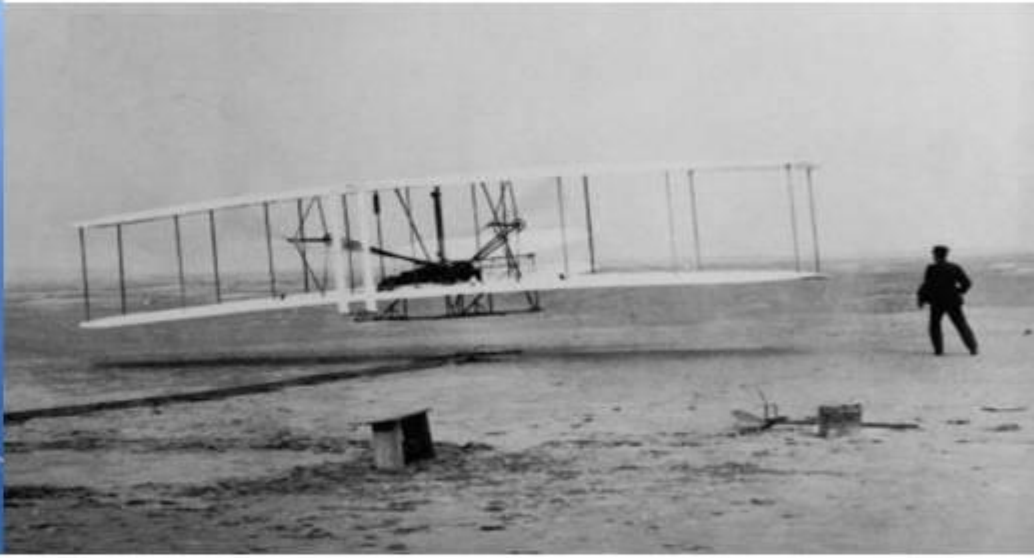
Critical factors for the success of InChI project

1. Technically competent staff

2. Fulfill a real community need

3. Political and Financial Support

The Internet has made the world more homogenous for chemical information and the Open Source InChI/InChIKey is not affected by global boundaries or proprietary chemical structure representations.



InChI Technology

Other Technology

Technical: InChI is a unique representation/identifier for defined chemical structures. Probably marginally better than previous ones. The InChI algorithm was built on the shoulders of giants.
http://en.wikipedia.org/wiki/Graph_theory

Practical: InChI and the related hash-code compressed InChIKey are the only available universal LINKs for in-house and public databases of defined chemical structures. Adoption and use by the vast majority of publishers and database providers assure it will be widely used.

InChI is the worst computer readable structure representation except for all those other forms that have been tried from time to time.

With apologies to Sir Winston Churchill
(House of Commons speech on Nov. 11, 1947)

Why InChI is becoming a success

1. Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be LINKED to and combined with other content on the Internet.

2. InChI is a public domain algorithm that anyone, anywhere can freely use.

How do we know the InChI project is beneficial?

Success is uncoerced adoption

InChI have some advantages over other chemical identifiers developed before:

(1) They are freely useable and non-proprietary.

(2) They allow a more advanced representation of chemical information than other codes (such as the SMILES code).

(3) They are unambiguous, i.e. conversion of chemical structures using standardized algorithms only leads to one InChI.

(4) They are precisely indexed by major search engines such as Google.

However, InChI are not applicable to generic formats often disclosed in patent literature, such as Markush structures, since they were rather designed to represent specific chemical structures and compounds. InChI therefore are not yet useful for comprehensive retrieval of patent literature.

Excerpt taken from:
Full-text prior art and chemical structure searching in e-journals and on the internet – A patent information professional's perspective
World Patent Information, Volume 31, Issue 4, December 2009, Pages 278-284
Maik Annies (Syngenta)

The best way to represent a chemical compound is not by a name or even a database identifier, but by its structure encoded in Structure Data Format (SDF MDL V2000) or the open Chemical Markup Language (CML) format or InChI codes. A few databases already provide the IUPAC/NIST standard of InChI codes or the shorter hashed InChIKey. The new InChIKey resolver services implemented by the Royal Society of Chemistry (RSC) and Chemspider allows to create InChIKeys from molecular structures and a reverse lookup of InChIKeys to obtain the associated known structures from molecular databases. The InChIKey can be used for web based literature search and also for chemical database search and merging of compound lists from multiple sources. Some other databases support the SMILES code for structures. The use of SMILES code is not recommended because multiple vendors create different representations of the SMILES code. Also true canonical (unique) SMILES are vendor specific.

Extracted from:
Kind T, Scholz M, Fiehn O:
How large is the metabolome? A critical analysis of data exchange practices in chemistry.
PLoS One 2009, 4:e5440.

Scientific Articles Mentioning InChI

"Challenges in integrating *Escherichia coli* molecular biology data"
Anália Lourenço, Sónia Carneiro, Miguel Rocha, Eugénio C. Ferreira, Isabel Rocha
Brief. Bioinform. **2010**, published online on 07 November 2010

"EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities"
Kun-Yi Hsin, Hugh P. Morgan, Steven R. Shave, Andrew C. Hinton, Paul Taylor and Malcolm D. Walkinshaw
Nucl. Acids Res. **2010**, published online on 4 November 2010 (Open Access)

"The RCSB Protein Data Bank: redesigned web site and web services"
Peter W. Rose, Bojan Beran, Chunxiao Bi, Wolfgang F. Bluhm, Dimitris Dimitropoulos, David S. Goodsell, Andreas Prlić, Martha Quesada, Gregory B. Quinn, John D. Westbrook, Jasmine Young, Benjamin Yukich, Christine Zardecki, Helen M. Berman and Philip E. Bourne
Nucl. Acids Res. **2010**, published online on 29 October 2010 (Open Access)

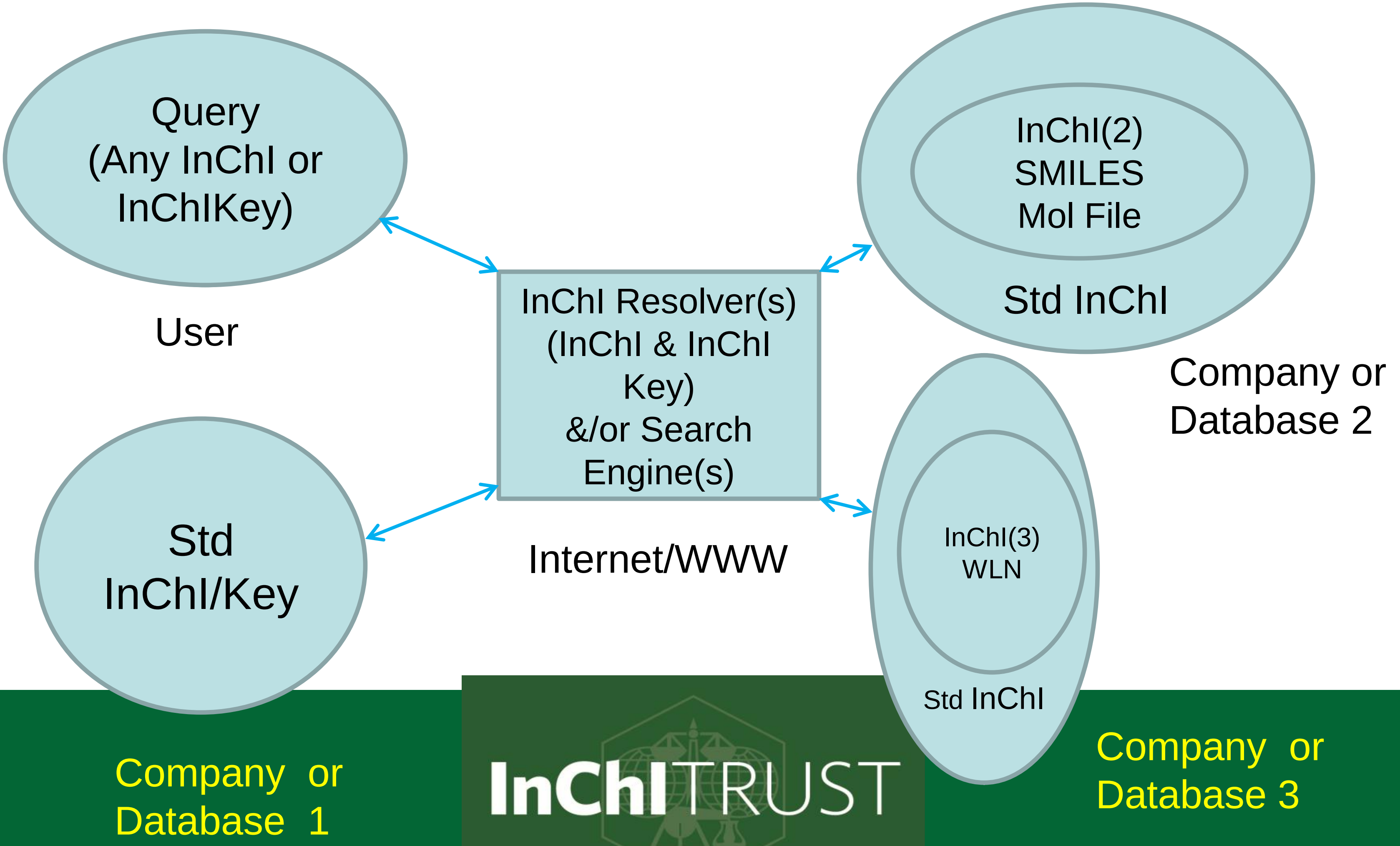
"ChemProt: a disease chemical biology database"
Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak and Tudor I. Oprea
Nucl. Acids Res. **2010**, published online on 8 October 2010 (Open Access)

"The Chemical Translation Service - a web-based tool to improve standardization of metabolomic reports"
Gert Wohlgemuth, Pradeep Kumar Haldiya, Egon Willighagen, Tobias Kind and Oliver Fiehn
Bioinformatics **2010**, *26(20)*, 2647-2648 (Open Access)

"PathwayAccess: CellDesigner plugins for pathway databases"
John L. Van Hemert and Julie A. Dickerson
Bioinformatics **2010**, *26(18)*, 2345-2346 (Open Access)

"Utopia documents: linking scholarly literature with research data"
T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer and D. Thorne
Bioinformatics **2010**, *26(18)*, i568-i574 (Open Access)

The LINKED and Interoperable and Combinable World of InChI



Query
(Any InChI or
InChIKey)

User

InChI Resolver(s)
(InChI & InChI
Key)
&/or Search
Engine(s)

Internet/WWW

InChI(2)
SMILES
Mol File

Std InChI

Company or
Database 2

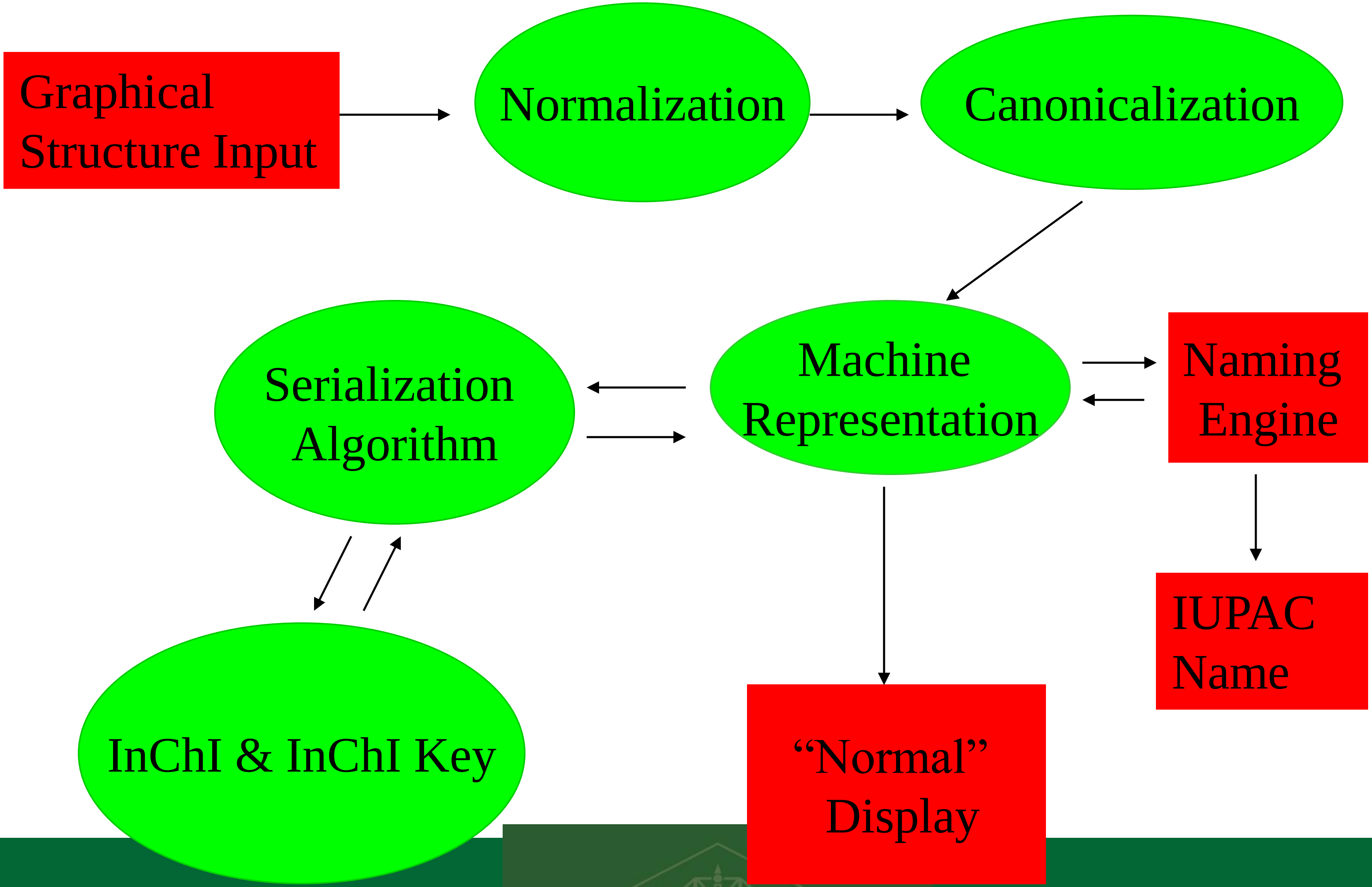
Std
InChI/Key

Company or
Database 1

InChI(3)
WLN

Std InChI

Company or
Database 3



Graphical Structure Input

Normalization

Canonicalization

Machine Representation

Naming Engine

Serialization Algorithm

IUPAC Name

InChI & InChI Key

"Normal" Display

InChI layered structure design

The current InChI layers are:

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)

Charges are added to end of the string

InChI Characteristics

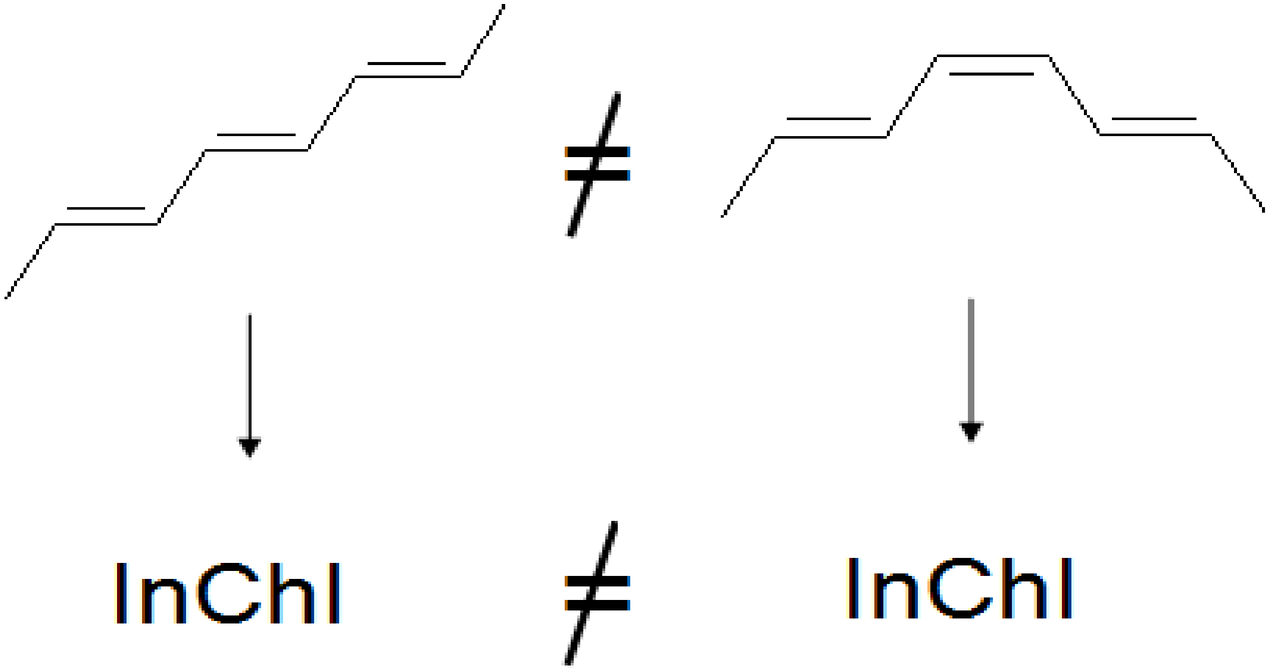
1. Easy to generate (It will use existing software.)

2. Expressive (It will contain structural information.)

3. Unique/Unambiguous

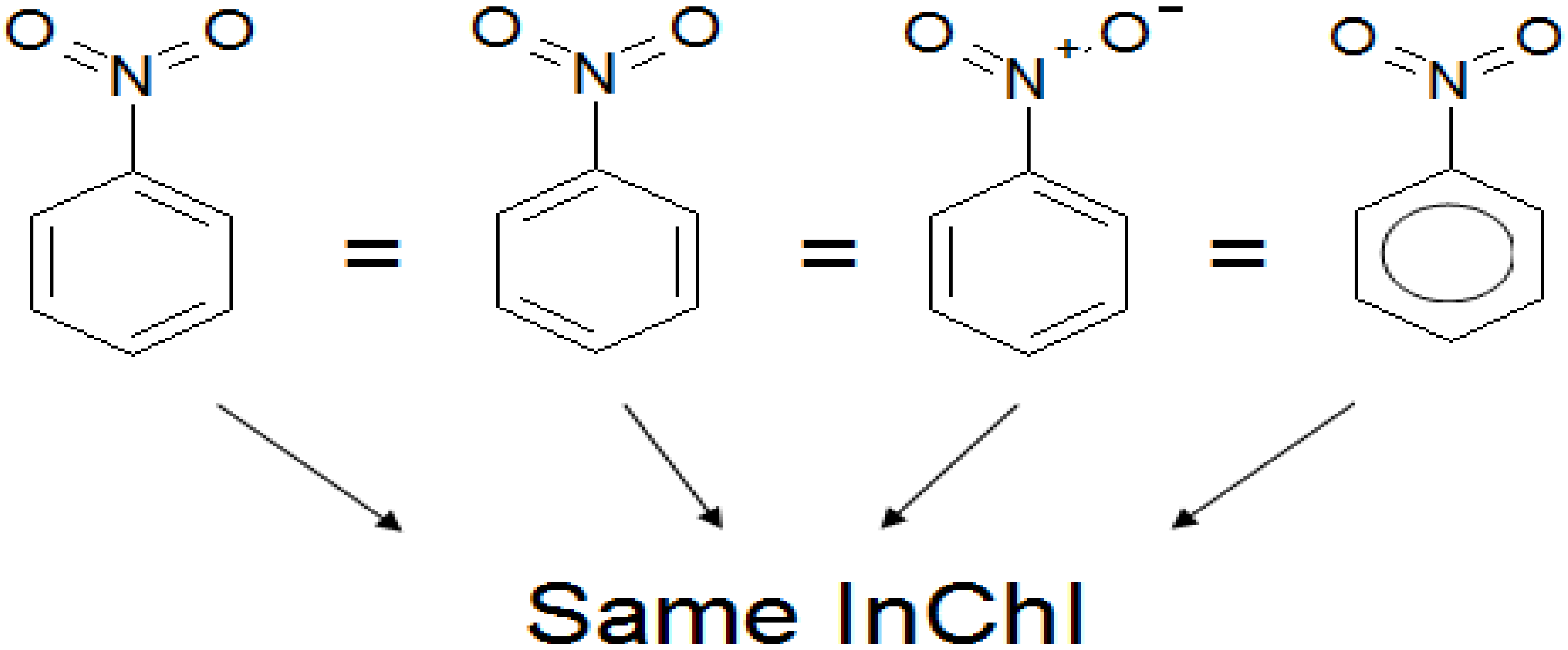
4. Easy to search for structure via Internet search engines (Google, Yahoo, Microsoft Live, etc.) using the InChI (hash) Key.

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

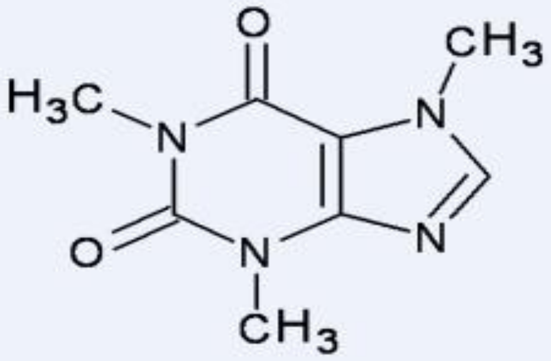


InChI ≠ InChI

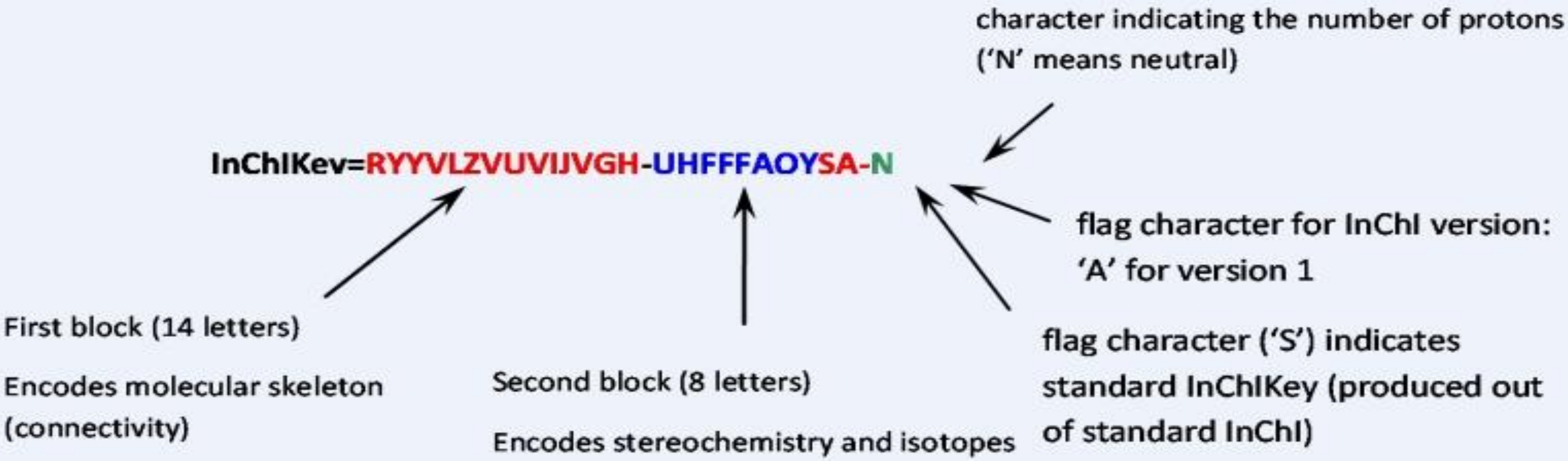
One compound has only ONE InChI



Same InChI



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



InChIKev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

character indicating the number of protons ('N' means neutral)

flag character for InChI version: 'A' for version 1

First block (14 letters)

Encodes molecular skeleton (connectivity)

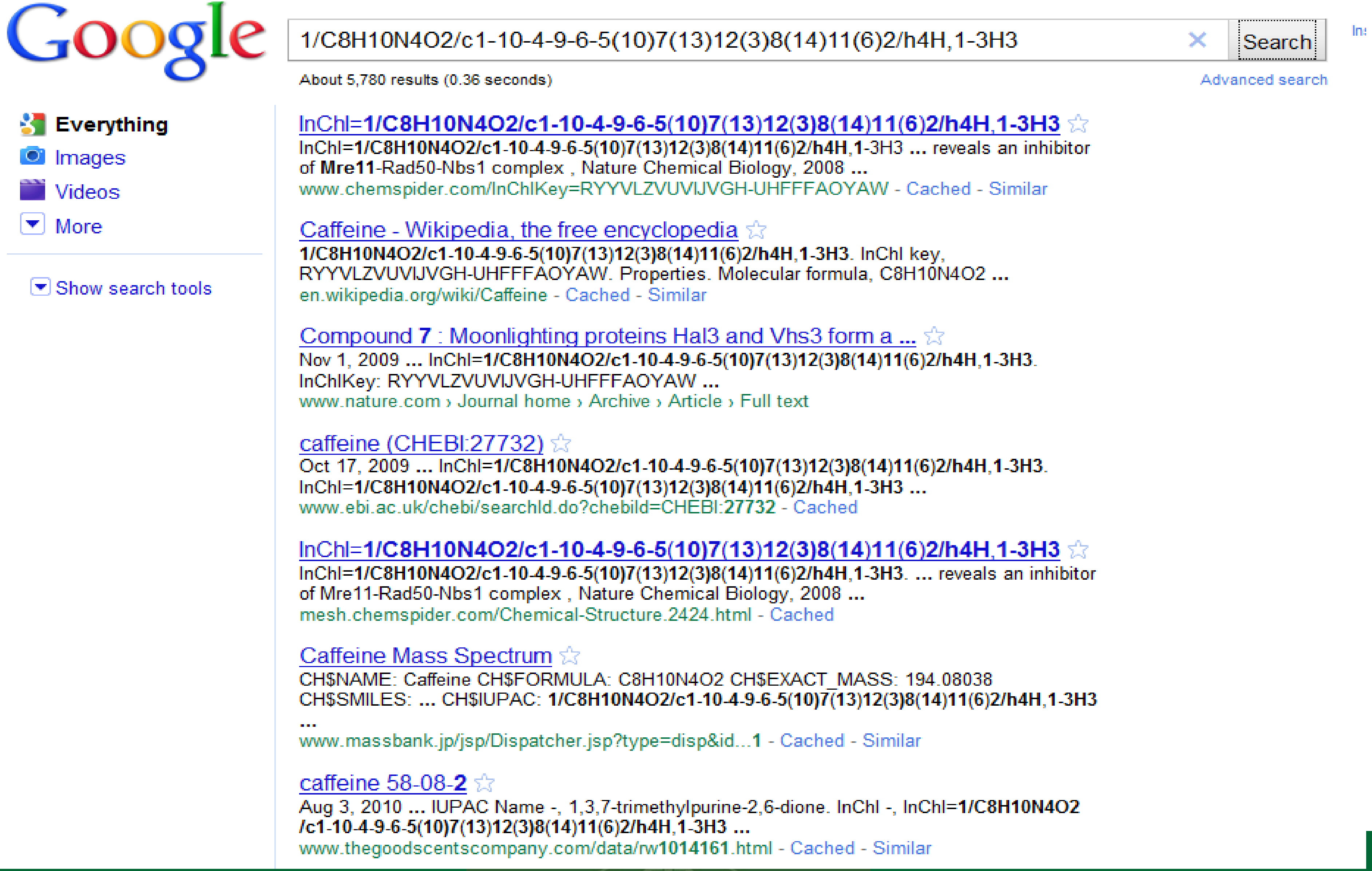
Second block (8 letters)

Encodes stereochemistry and isotopes

flag character ('S') indicates standard InChIKey (produced out of standard InChI)

Bar Codes – not designed to be read by humans

InChI – not designed to be read by humans. It is designed to be read by search engines and computer programs



About 5,780 results (0.36 seconds)

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

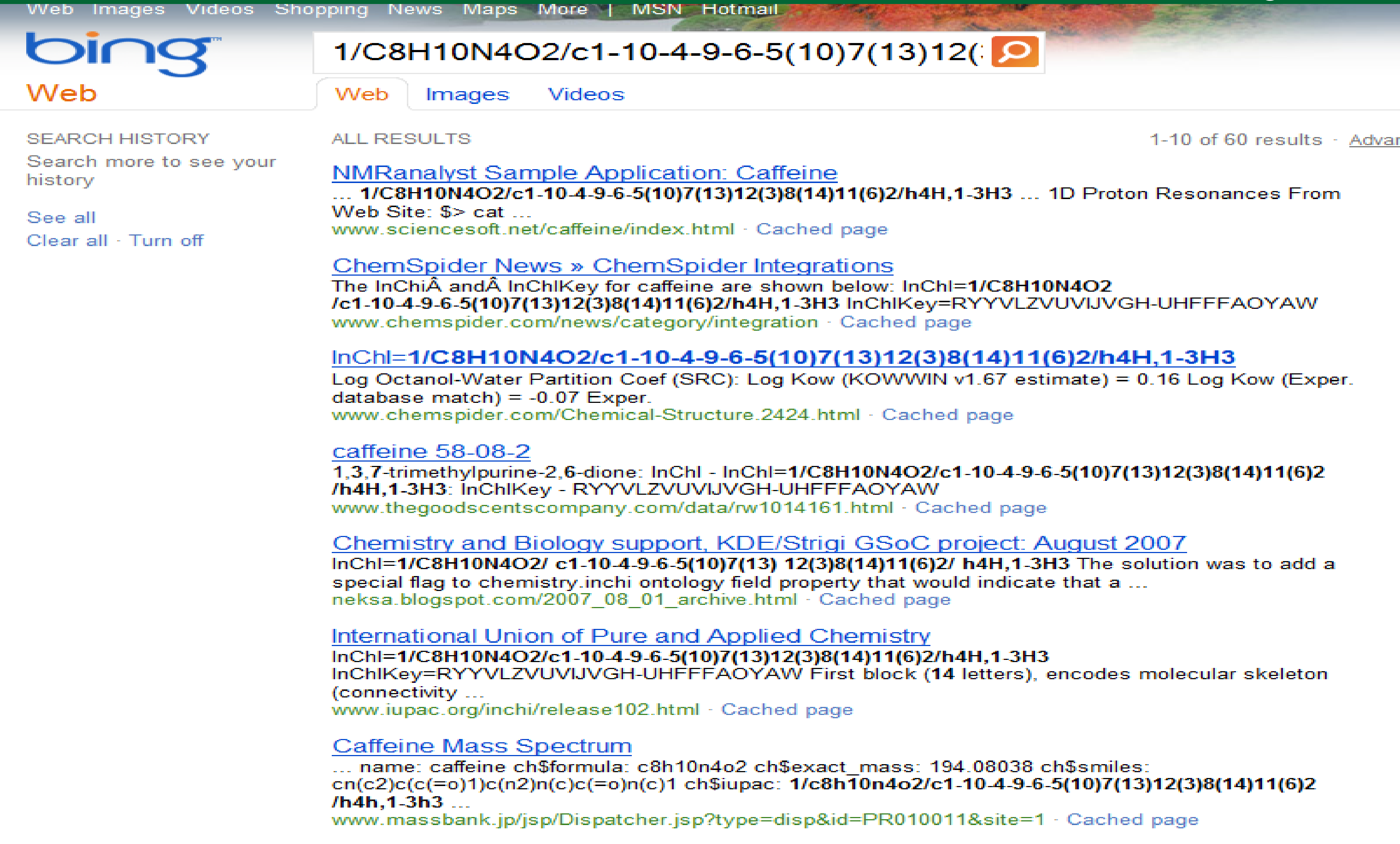
Compound 7 : Moonlighting proteins Hal3 and Vhs3 form a ...
Nov 1, 2009 ... InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3. InChIKey: RYYVLZVUVIJVGH-UHFFFAOYAW ...
www.nature.com › Journal home › Archive › Article › Full text

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/rw1014161.html - Cached - Similar


Web Images Videos Shopping News Maps More | MSN Hotmail
bing
Web
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(
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

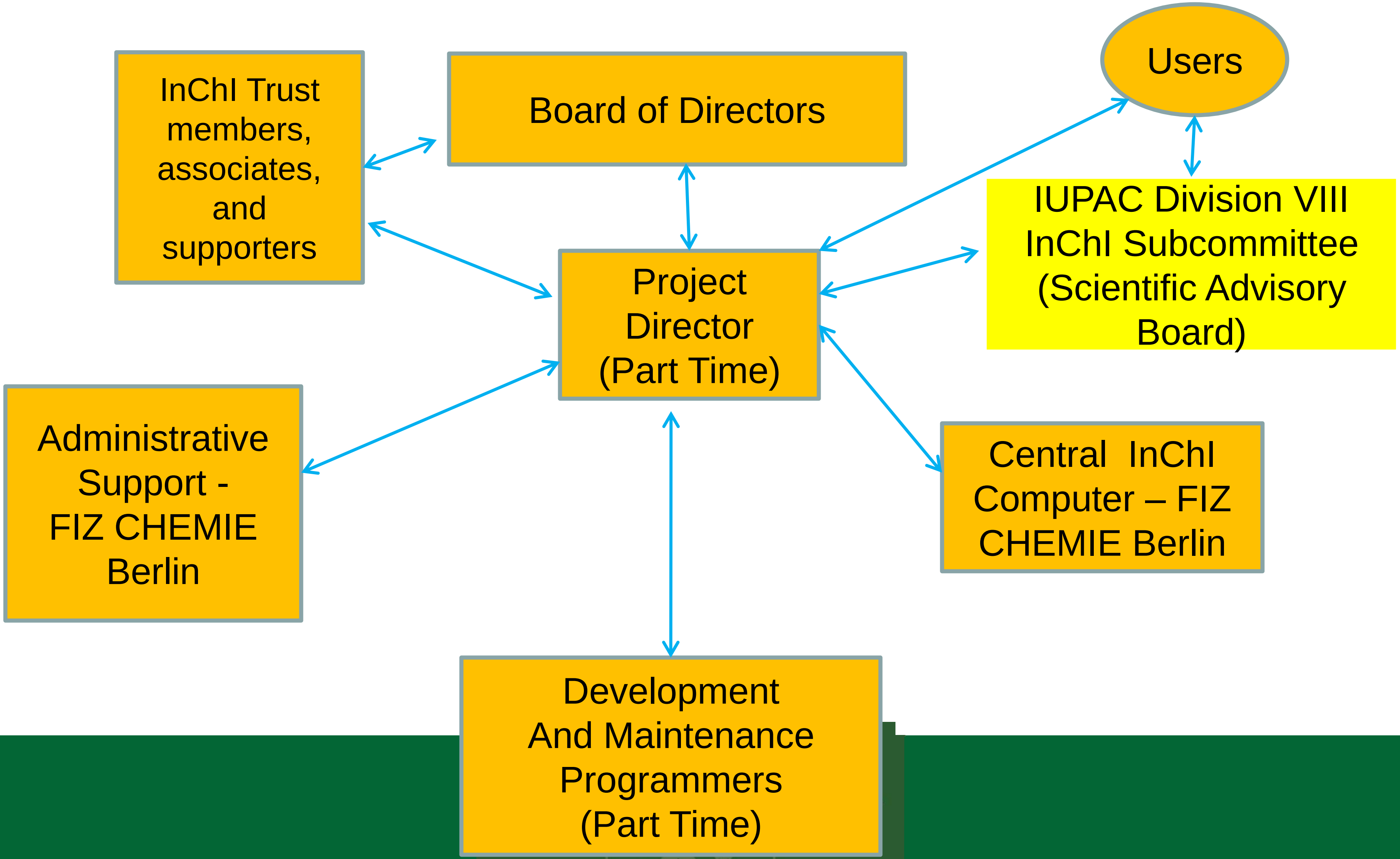
... name: caffeine ch$formula: c8h10n4o2 ch$exact_mass: 194.08038 ch$smiles: cn(c2)c(c(=o)1)c(n2)n(c)c(=o)n(c)1 ch$iupac: 1/c8h10n4o2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2 /h4h,1-3h3 ...

www.massbank.jp/jsp/Dispatcher.jsp?type=disp&id=PR010011&site=1 · Cached page

The InChI Trust

With the needs of NIST fulfilled with respect to what capabilities of an InChI are required for NIST databases, and since IUPAC is fundamentally and culturally a volunteer organization, there needs to be a way to continue development of InChI, and maintain the InChI algorithm. As a result of it was concluded that a not-for-profit organization would best fit the project needs. Thus the decision to create and incorporate the "InChI Trust" . As there is no "free lunch", the Trust will need resources to continue to operate. Membership in the InChI Trust requires annual dues. The income from these revenues will be used exclusively for InChI development, maintenance, and educational activities associated with the project. Membership will entitle a member to influence the direction, priority, and speed of further Trust activities. Those organizations which do not join the InChI Trust will still have free access to the InChI algorithms but will not participate in any decision-making or direction -setting activities.

InChI Trust Organization



InChI Trust members, associates, and supporters

Board of Directors

Users

IUPAC Division VIII InChI Subcommittee (Scientific Advisory Board)

Project Director (Part Time)

Administrative Support - FIZ CHEMIE Berlin

Central InChI Computer – FIZ CHEMIE Berlin

Development And Maintenance Programmers (Part Time)

Current InChI Trust Members

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

16 as of 11/16/2010

Current InChI Trust Supporters

American Chemical Society Division of Chemical Information (CINF)
CalTech, CA, USA
Chem21, IL, USA
Indiana University, IN, USA
Leadscope, Columbus, OH, USA
National Chemical Laboratory, Pune, India
National Institute of Chemistry, Ljubljana, Slovenia
NextMove Software, USA
Open Babel Project, USA
SharePoint, WA, USA
Trinity University, TX ,USA
Technical University – Vienna, Austria
Unilever Centre for Molecular Science Informatics, Cambridge, UK
University of Applied Science, Gelsenkirchen, Germany
University of California - Davis
University of California – Riverside
University of California – San Francisco
University of North Carolina, NC, USA
University of Paderborn, Germany
University of the West Indies, Mona, Jamaica
Xemistry GmbH, Germany

21 as of 9/22/2010

Current IUPAC Working Groups

Markush
Polymers/Mixtures
Organometallics
InChI Resolver
Electronic States
RInChI –InChI for Reactions

Possible Future Enhancements

1. Transrutherfordium elements
2. Electronic States, including Transition states and Excited states.
3. Work with IUCr for 3D information
4. Proteins, Peptides & Biopolymers
5. Mac supported version
6. Java version
7. VS2010 .NET compilation support

Trust Funded Activities

GGA Software (St. Petersburg) has just completed an InChI Trust project for an InChI certification program which will allow those with InChI's in their products to post the InChI certification logo indicating the InChI's in that database have been checked and certified by the Trust.

A contract for documentation is currently being evaluated and considerable additional documentation should be available by the spring of 2011. We may even consider a YouTube video as part of the documentation.

The Future for InChI

InChI has become mainstream for publishers, databases providers, and software developers. Over the next 5-10 years, publishers will use data mining to create both better abstracts, useful indexing, and concept terms. Search engines will be able to search for appropriate text and structures and direct users to the original (fee or free/Open Access/Open Data) sources.

Summary

If you are not part of the solution; you are part of the precipitate.

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

Very special acknowledgement to Gideon Babe Ruth Howard for the prologue