InChI Status

Stephen Heller

steve@inchi-trust.org

srheller@nist.gov

www.inchi-trust.org

www.iupac.org/inchi

4/13/2011

InChI TRUST

What chemistry needs is a system or network of people to make good use of the Internet, blogs, and twitter. We need to connect information and ideas and stop protecting them with barriers (which are primarily financial). Only the power of open systems will be able to generate new ideas which will lead to knowledge.

The problem is simple to see, but hard to fix. Why – because there is a lack of integration. There are:

multiple applications
multiple repositories
multiple interfaces and protocols

Missing information, facts, and data wastes times and cost money!   The way to move integration forward is with standards.

But for chemical structures there is a solution….

InChI

“No, no, not another structure standard!!!”

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.

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

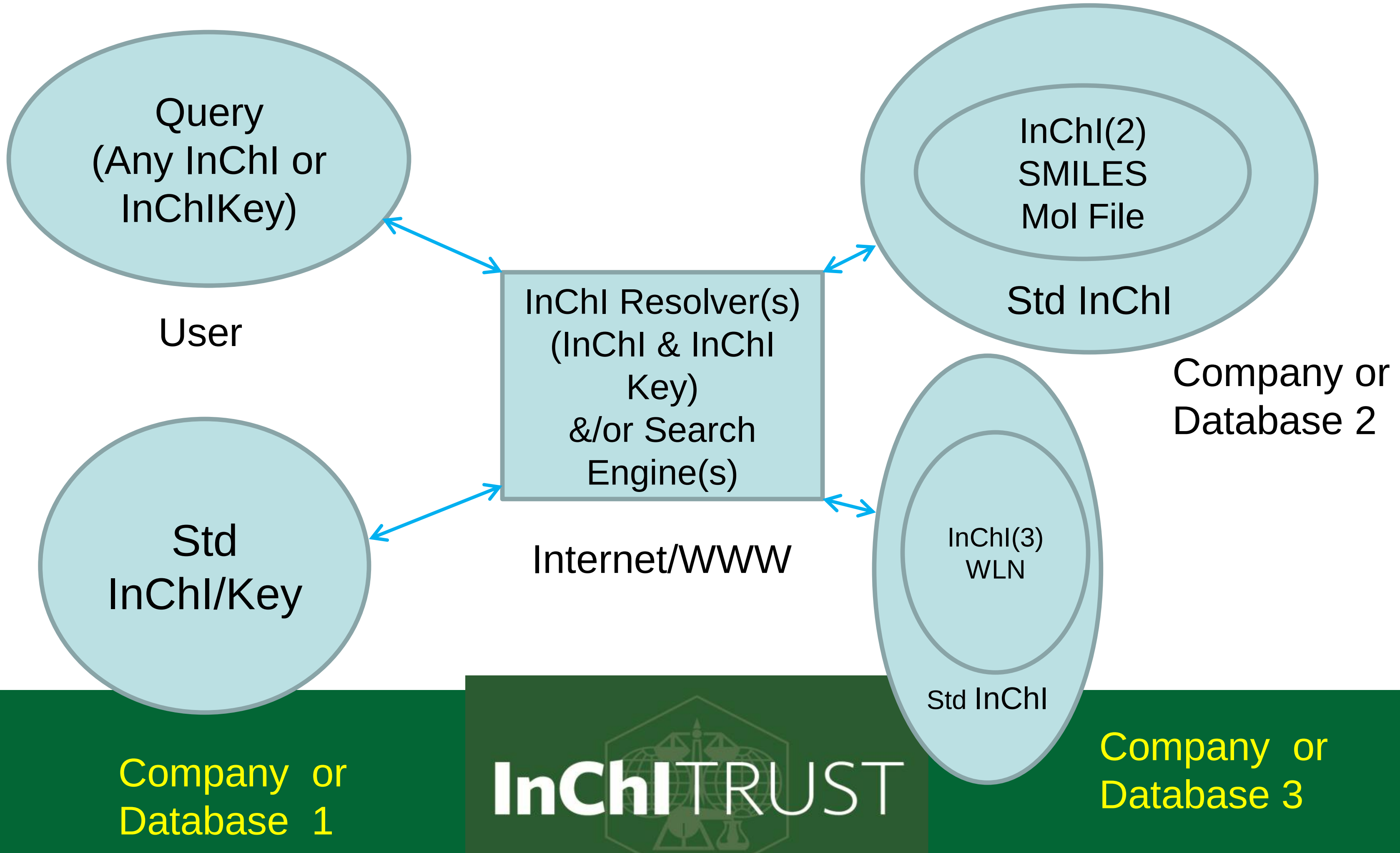
1. Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be LINKED to and combined with other content on the Internet.

2. InChI is a public domain algorithm that anyone, anywhere can freely use. By giving away the algorithm the project is building trust with the community.

How do we know the InChI project is beneficial?

Success is uncoerced adoption

The LINKED and Interoperable and Combinable World of InChI



Query
(Any InChI or
InChIKey)

User

Std
InChI/Key

Company or
Database 1

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

Internet/WWW

InChI(2)
SMILES
Mol File

Std InChI

Company or
Database 2

InChI(3)
WLN

Std InChI

Company or
Database 3

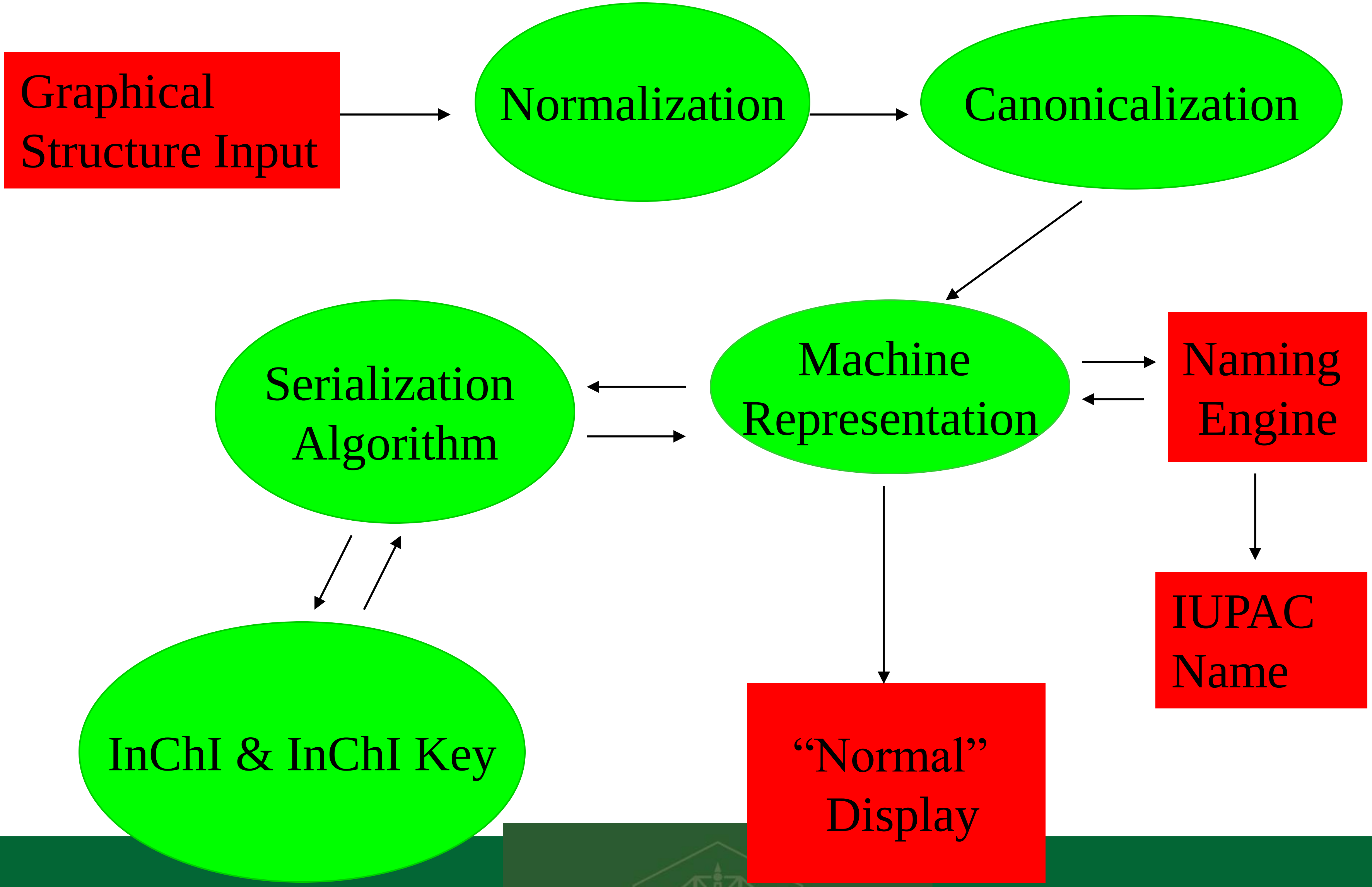
InChI Policy & Culture

Do not go outside our circle of competence.

No mission creep.

Staff is not territorial.

InChI Trust is doing well because it really doesn’t require a lot of resources.



Graphical Structure Input

Normalization

Canonicalization

Machine Representation

Serialization Algorithm

Naming Engine

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

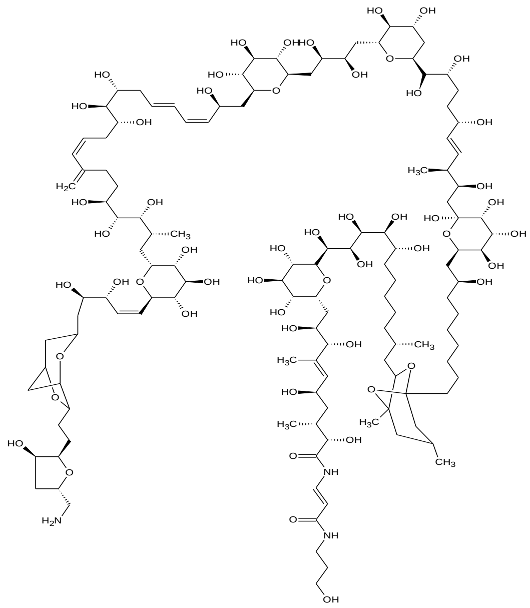
1. Easy to generate (It will use existing software.)

2. Expressive (It will contain structural information.)

3. Unique/Unambiguous

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

Really long InChI (Palytoxin)



Palytoxin

Isolated from Hawaiian soft coral

One of the most toxic non-peptide substances

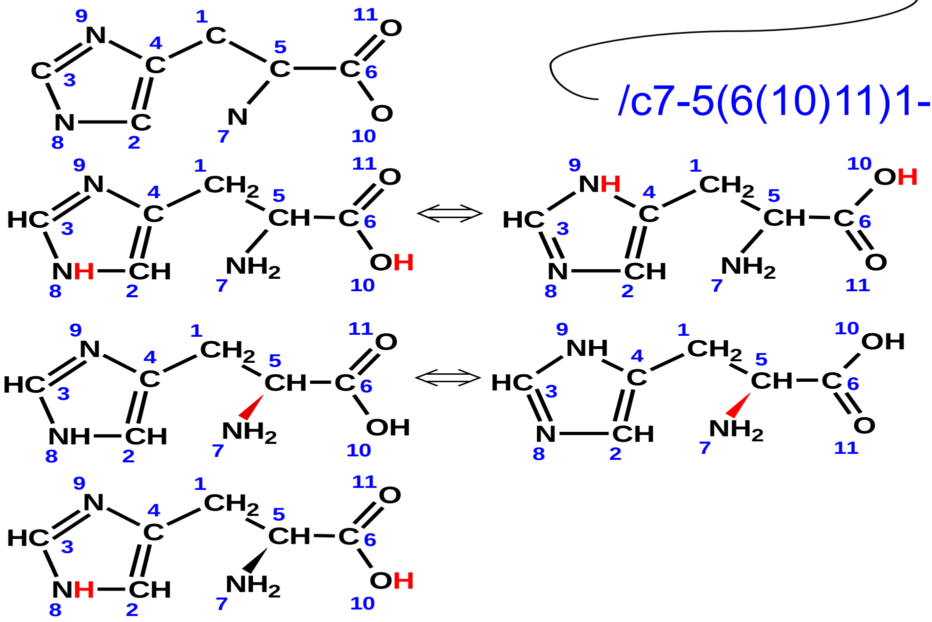
Contains >70 stereochemical elements

```
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(15
7)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(14
1)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-DCBUCRFRSA-N

InChI Layers: L-Histidine

$C_6H_9N_3O_2$



InChI=1/C6H9N3O2

Formula

/c7-5(6(10)11)1-4-2-8-3-9-4

Connections

/h2-3,5H,1,7H2,(H,8,9)(H,10,11)

Hydrogens (mobile)

/t5-/m0/s1

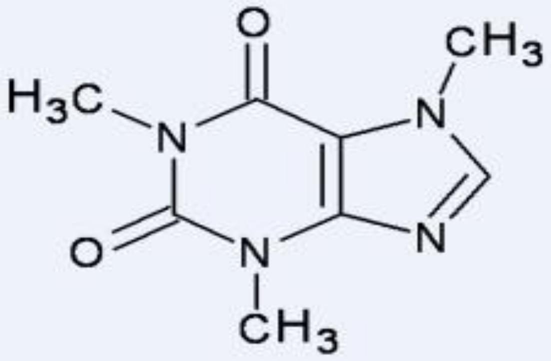
Stereo

/f/h8,10H

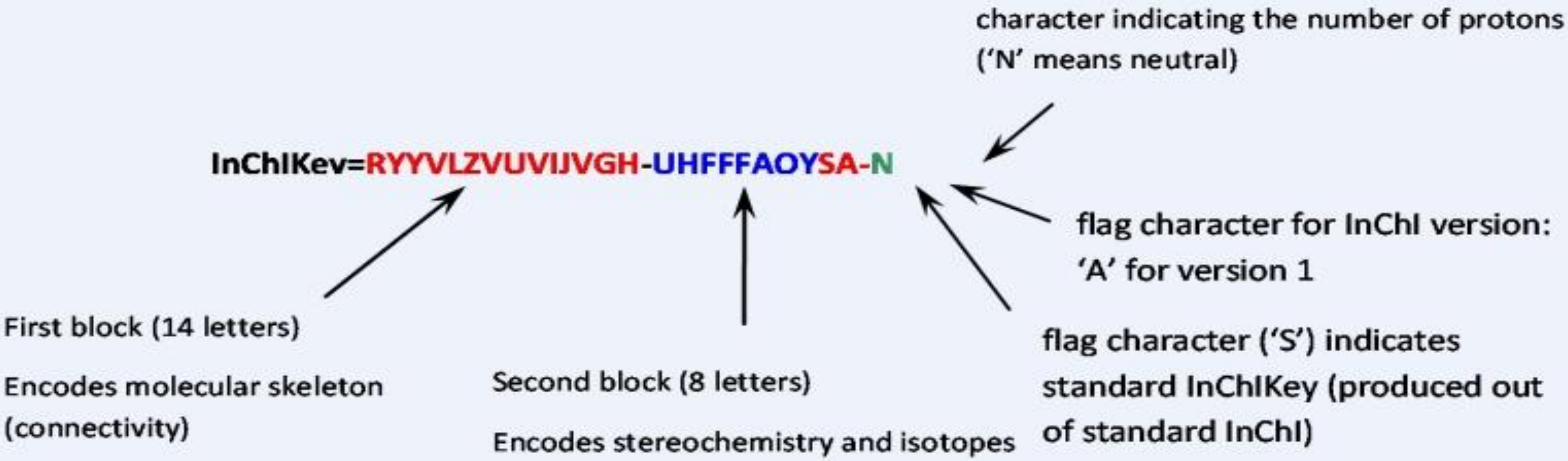
Hydrogens (fixed)

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



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

InChIKev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

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)

Example of using InChI vs. SMILES for actual Chemistry/Science:

Simplified molecular input-line entry system and International Chemical Identifier in the QSAR analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors.
AP Toropova, AA Toropov, E Benfenati, and G Gini
Chem Biol Drug Des, February 26, 2011; .

The simplified molecular input-line entry system (SMILES) and IUPAC International Chemical Identifier (InChI) were examined as representations of the molecular structure for quantitative structure - activity relationships (QSAR), which can be used to predict inhibitory activity of styrylquinoline derivatives against the human immune deficiency virus type 1 (HIV-1). Optimal SMILES-based descriptors give a best model with n= 26, r(2) =0.6330, q(2) =0.5812, s=0.502, F=41 (training set) n= 10, r(2) =0.7493, r(2) (pred) =0.6235, R(m) (2) =0.537, s=0.541, F=24 (validation set). Optimal InChI-based descriptors give a best model with n= 26, r(2) =0.8673, q(2) =0.8456, s=0.302, F=157 (training set); n= 10, r(2) =0.8562, r(2) (pred) =0.7715, R(m) (2) =0.819, s=0.329, F=48 (validation set). **Thus, the InChI-based model is preferable**. The described SMILES-based and InChI-based approaches have been checked with five random splits into the training and test sets.

Members and Associates:

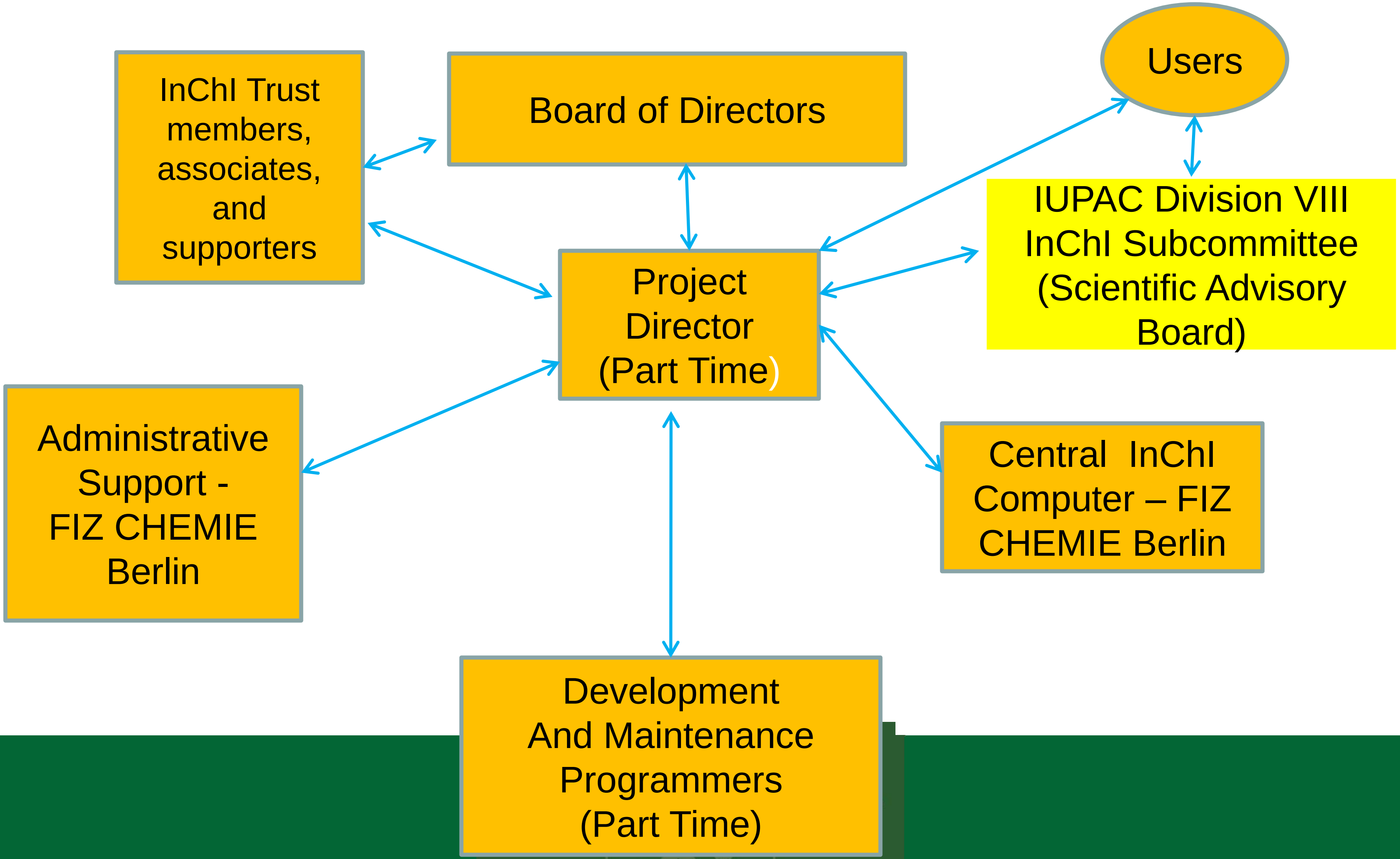
Accelrys
ACD Labs
ChemAxon
Dialog
Elsevier
FIZ CHEMIE Berlin
IBM Research
Informa/Taylor & Francis
IUPAC
John Wiley & Sons
Nature
OpenEye
Royal Society of Chemistry
Springer

3/11

Supporters:

American Chemical Society Division of Chemical Information (CINF) (Carmen Nitsche)
Caltech Library Services, Pasadena, CA, USA (Dana Roth)
Chemistry Department, University of California, Riverside, CA, USA (Chris Reed)
Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, NC, USA (Alex Tropsha)
ETH Zürich, Chemistry Biology Pharmacy Information Center, Switzerland (Martin Brändle)
Faculty of Science, University of Paderborn, Germany (Gregor Fels)
Imperial College London, UK (Henry Rzepa)
Institute for Chemoinformatics and Bioinformatics, University of Applied Sciences Gelsenkirchen, Recklinghausen Section, Germany (Achim Zielesny)
International Union of Crystallography (Peter Strickland)
Leadscope, Columbus, OH, USA (Michael Conley)
Ludwig-Maximilians-Universität München, Munich, Germany (Thomas Engel)
National Center for Biomedical Ontology, Stanford University, CA, USA (Mark Musen)
National Chemical Laboratory, Pune, India (Muthukumarasamy Karthikeyan)
National Institute of Chemistry, Ljubljana, Slovenia (Dusanka Janezic)
NextMove Software, Santa Fe, NM, USA (Roger Sayle)
Open Babel (Noel O'Boyle)
SciencePoint, Redmond, WA, USA (Rudy Potenzone)
Technical University of Vienna, Austria (Ulrich Jordis)
The Chem21 Group, Inc., Lake Forest, IL, USA (Tony Hopfinger)
Trinity University, San Antonio, TX, USA (Steven Bachrach)
Unilever Centre for Molecular Science Informatics, Cambridge University, UK (Robert Glen)
University of California, Davis, Genome Center, CA, USA (Oliver Fiehn)
University of California, San Francisco, CA, USA (John Irwin)
University of Indiana, Bloomington, IN, USA (David Wild)
University of the West Indies, Mona Campus, Jamaica (Robert Lancashire)
Xemistry GmbH, Königstein, Germany (Wolf-Dietrich Ihlenfeldt)

InChI Trust Organization



Users

InChI Trust members, associates, and supporters

Board of Directors

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 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
8. Integrate with Microsoft Chem4Word

The Future

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.

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, Randy Marcinko, 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.