The Status of the IUPAC InChI Project and the InChI Trust

Stephen Heller
(steve@inchi-trust.org)

The slides from this presentation can be found at :

http://www.hellers.com/steve/pub-talks/
(Goslar 2009 link)

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi
and
http://www.inchi-trust.org

The InChI Team

(alphabetical order)

Stephen R. Heller
Alan McNaught
Igor Pletnev
Stephen E. Stein
Dmitrii Tchekhovskoi

This presentation is about InChI and the InChIKey from a view 40,000 feet up. It is the overall vision and direction.

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 Use InChI?

For publishers and database providers it gives one a competitive advantage being able to link content. It offers users the ability to help in discovery of information and data.

"In my view, the most important rule of business in today's integrated and digitized global market, where knowledge and innovation tools are so widely distributed. It's this: Whatever can be done, will be done. The only question is will it be done by you or to you. Just don't think it won’t be done."

By Thomas L. Friedman
Published: December 9, 2008
NY Times

(That is – not if, but when and by whom)

InChI is an agent of change

Critical factors for the success of InChI project

1. Technically competent staff
2. Fulfill a real community need
3. Political and Financial Support

Why InChI is becoming a success

1. Organizations need a structure representation for their content (databases, journals, products, and so on) so that their content can be linked to 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

Initial InChI Goal (Plan A)

– Cover 100% of all chemical found in the literature and in databases

Current InChI Goal (Plan B)

- Cover 99.9% of chemicals found in the literature and in databases.

Bar Codes – not designed to be read by humans

InChI – not designed to be read by humans

The InChI representation and algorithms are not new. They are just a further, well thought out and tested (minor) improvement on graph theory which is some 300 years old. It started with a publication by the Swiss mathematician Euler and has been applied to chemical structures in the mid 20th century.

http://en.wikipedia.org/wiki/Graph_theory

&

http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

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)

InChI layered structure design

The current InChI layers are:

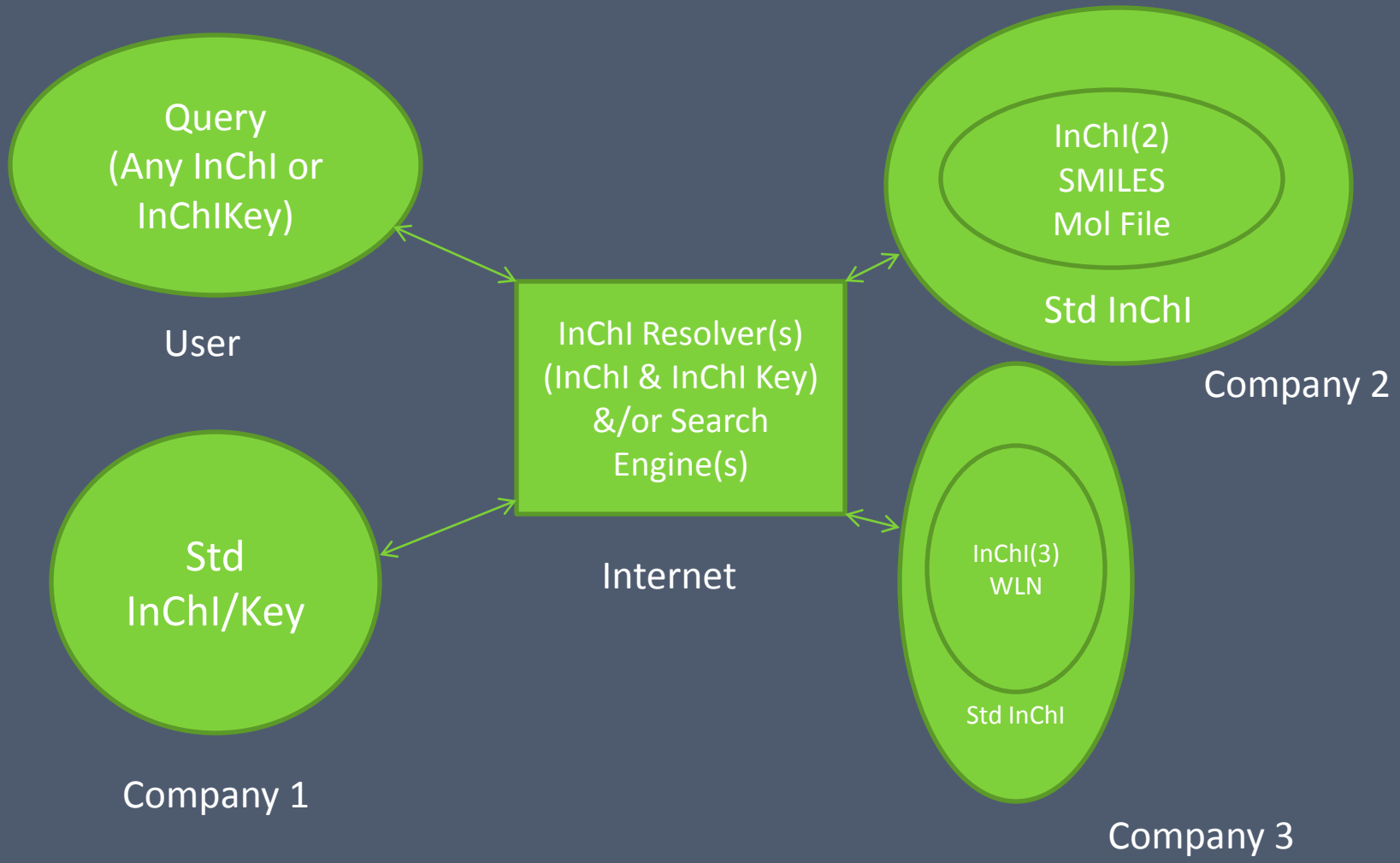
1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond (*Z/E)*
    b. tetrahedral (sp3)
5. Tautomers (on or off)

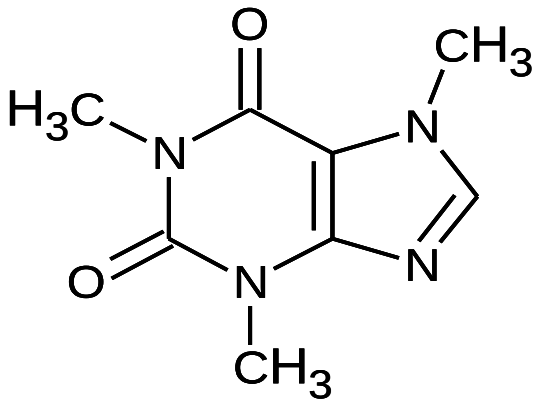
Charges are added to end of the string

The InChI/InChIKey Standard

The nice &/or awful thing about standards is there are so many of them. In September 2008 at the first meeting of the IUPAC Division VIII InChI subcommittee a single standard was chosen (dissidents quietly cremated – which also had the side effect of making the size of the subcommittee more manageable) , which, being a single standard, probably satisfied no one except perhaps Google and Microsoft Live Search. The ONLY purpose of the standard is to allow linking between databases internally or on the web. As noted in the next slide, the InChI standard is NOT a replacement for the way in which any organization represents their structures. The standard InChI is in ADDITION to any existing internal way in which a structure is represented in a database.

The Linked and Interoperable World of InChI





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)

InChIKey=RYYVLZVUVIJVGH-UHFFFAOYSA-N

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

First block (14 letters)

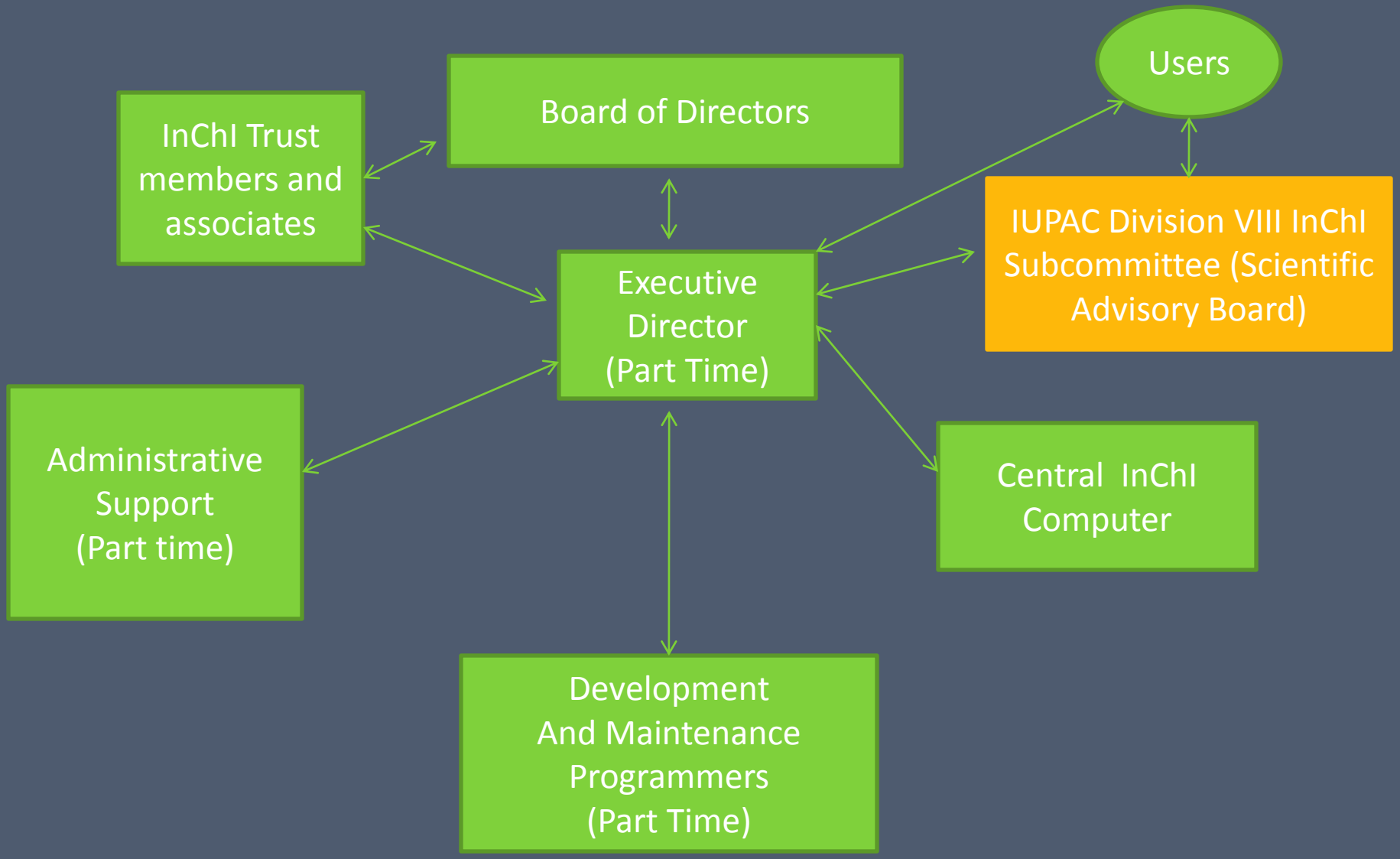
Encodes molecular skeleton (connectivity)

Second block (8 letters)

Encodes stereochemistry and isotopes

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

InChI Trust Organization



InChI Trust membership

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 numerous meetings, emails, and discussions, 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" in the UK. As there is no "free lunch", the Trust will need resources to continue to operate. Membership in the InChI Trust requires annual fees or 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. Membership will also provide InChI Trust "certification" of the InChIs and InChIKeys in a member's database. 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.

Summary

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

Current InChI Trust Members

ACD Labs
ChemAxon
Elsevier
FIZ Chemie – Berlin
Informa/Taylor & Francis
John Wiley & Sons
Nature
OpenEye
Royal Society of Chemistry
Symyx
Thomson-Reuters

11/2/09

Acknowledgements

Steve Bachrach, Colin Batchelor, Ted Becker, Jost Bohlen, Evan Bolton, Pieter Bolman, Steve Bryant, Harry Collier, Alice Cooper, Rene Deplanque, Ron Dunn, Jonathan Goodman, Guenter Grethe, Richard Kidd, Beda Kosata, Peter Linstrom, David Lipman, Gary Mallard, Randy Marcinko, Alan McNaught, Bill Milne, Miloslav Nic, Carmen Nitsche, Igor Pletnev, Josep Prous, Rich Roberts, Peter Murray-Rust, Henry Rzepa, Steve Stein, Peter Shepherd, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, and Tony Williams.