Status of the InChI and InChIKey algorithms

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

The slides from this presentation can be found at:

http://www.hellers.com/steve/pub-talks/ (Boston 2010 link)

and at the new InChI Trust web site

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

http://www.inchi-trust.org

Disclaimer

These slides were made from 100% recycled electrons.

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

I will try to be politically correct, and have even take a course in being PC. But, I flunked it – twice.

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 **ONLY** 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?

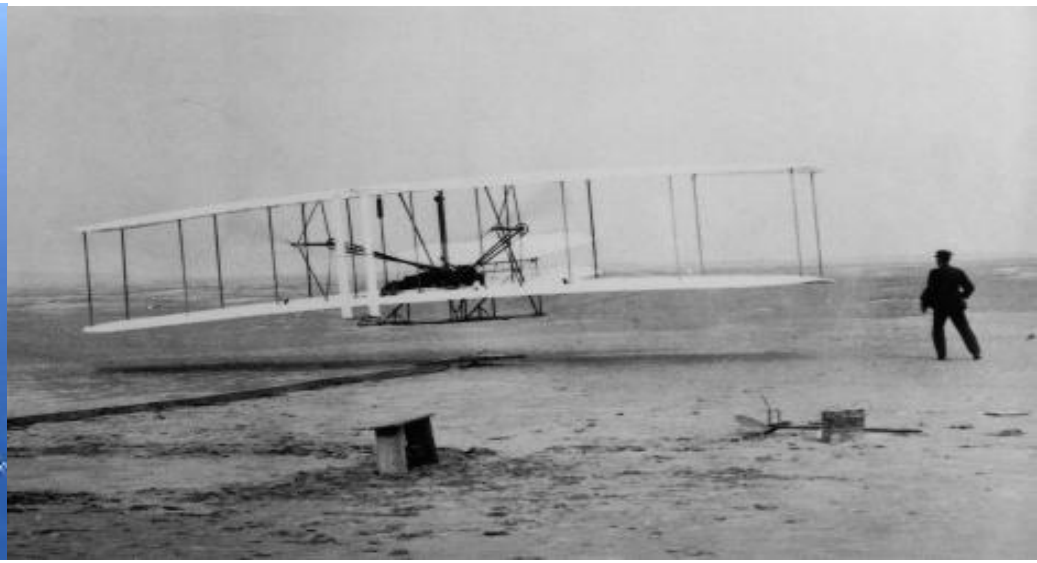
For publishers and database providers using InChI gives one a competitive advantage being able to LINK content from multiple sources. It offers users the ability to help in new discoveries from existing information and data by easily being able to integrate, remix, and retell. Business models that depend on things not changing (closed access and making aggregation and integration difficult) are so 20th century. InChI is a small, but critical and vital, part of new business models and technologies involving chemicals that will lead to new discoveries. Combinability increases the value of information and data.

We all know that in duck hunting you always lead the duck, which is to say, aim where you anticipate the duck to be in flight so you can succeed.

In chemical information the same holds true, if you want to continue to be a leader in the field you need to aim for where the field is heading. NIST, and now the InChI Trust members are doing just that.

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 is an agent of change for those individuals and organizations who have defined chemical structures and want to make their information known and available to the community.



InChI Technology

Other Technology

How long does it take for people/organizations actually use standards?

Title 15 U.S.C. Chapter 6 §(204) 205a - 205l
METRIC CONVERSION LAW
(Pub. L. 94-168, §2, Metric Conversion Act, Dec. 23, 1975)

§ 204. Metric system authorized. - It shall be lawful throughout the United States of America to employ the weights and measures of the metric system; and no contract or dealing, or pleading in any court, shall be deemed invalid or liable to objection because the weights or measures expressed or referred to therein are weights or measures of the metric system. (14 Stat.339, Adopted July 28,1866)

Executive Order 12770 of July 25, 1991
Metric Usage in Federal Government Programs

By the authority vested in me as President by the Constitution and the laws of the United States of America, Including the Metric Conversion Act of 1975, Public Law 94-168 (15 U.S.C.205a et seq.) ("the Metric Conversion Act"),as amended by section 5164 of the Omnibus Trade and Competitiveness Act of 1988, Public Law 100-418 ("the Trade and Competitiveness Act"), and in order to implement the congressional designation of the metric system of measurement as the preferred system of weight sand measures for United States trade and commerce, it is hereby ordered as follows:

Sec 1. Coordination by the Department of Commerce.

The Secretary of Commerce ("Secretary") is designated to direct and coordinate efforts by Federal departments and agencies to implement Government metric usage in accordance with section3 of the Metric Conversion Act (15 U.S.C. 205b), as amended by section5164(b) of the Trade and Competitiveness Act. ….

The Secretary shall report to the President annually regardingt he progress made in implementing this order.

Critical factors for the success of InChI project

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

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.**

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)

Initial InChI Goal (Plan A)
– Cover 100% of all chemical found in the literature and in databases and other sources.

Current InChI Goal (Plan B)
- Cover 99.9% of chemicals found in the literature and in databases and other sources.

Why – Because InChI covers ONLY well defined chemical structures (e.g., few polymer structures are known. Polymer chemists describe polymers by their properties, not structure.)

Bar Codes – not designed to be read by humans

InChI – not designed to be read by humans

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

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.

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.
...

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:

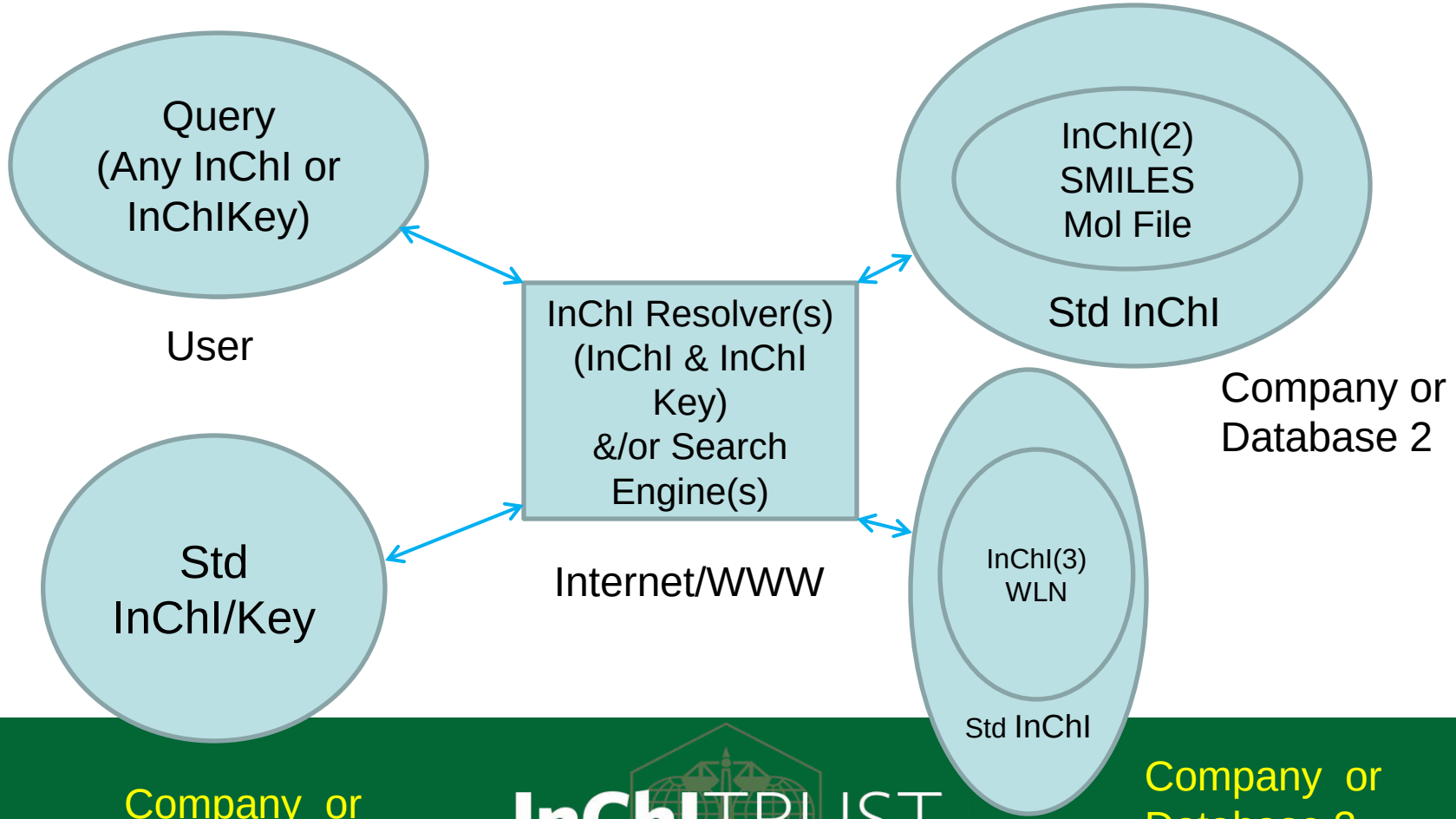
Return all the DOIs that have been registered with a given InChI
http://inchi.crossref.org/dois/InChI=1S/C4H6O2/c1-3-6-4(2)5/h3H,1H2,2H3

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

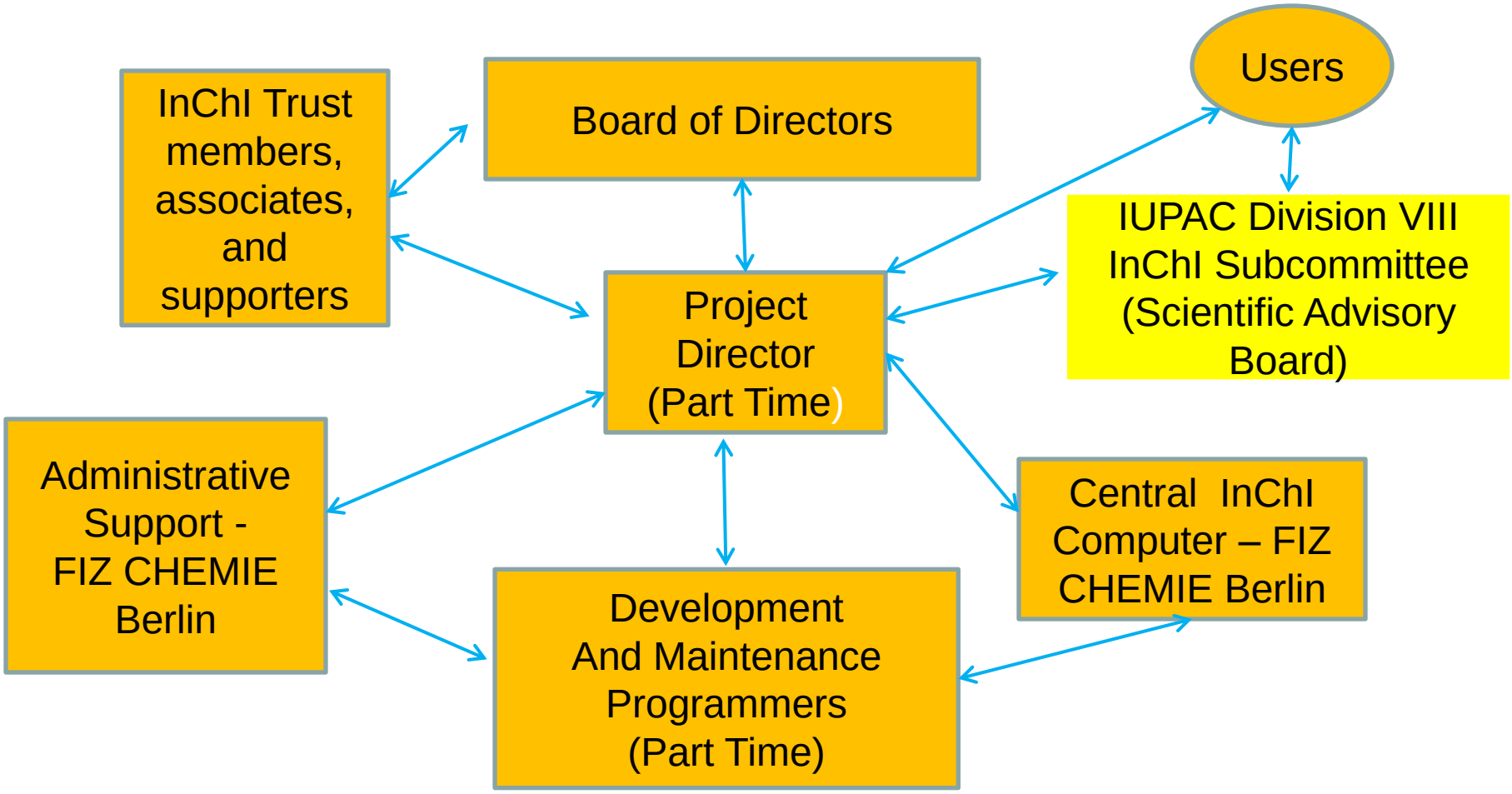
The InChI/InChIKey Standard

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 Bing search engines. 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 and Combinable World of InChI



InChI Trust Organization



InChI Trust Projects

Current:

Feasibility Study (Digital Chemistry)
QC/Certification Software (SciTouch)
IUPAC Working Groups

Future:

Documentation
Additional capabilities

IUPAC Working Groups

Markush

Polymers/Mixtures

Organometalics

InChI Resolver

Electronic States

RInChI – InChI for Reactions

Future Trust Projects/Issues

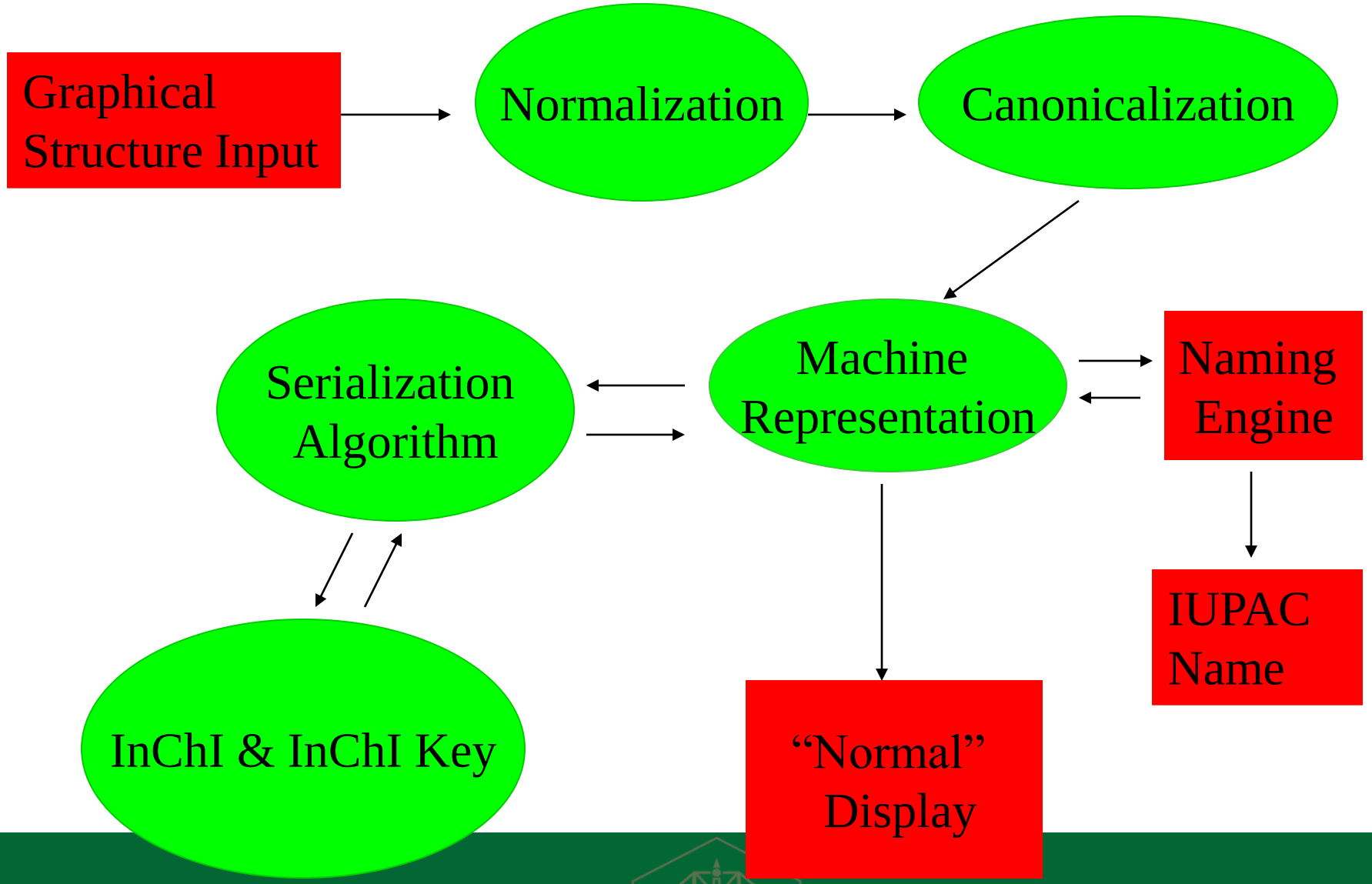
Suggestions and requests from the community:

Java version

Mac support

VS2010 .NET compilation support

Type of Open Source Licensing


Graphical Structure Input
Normalization
Canonicalization
Serialization Algorithm
Machine Representation
Naming Engine
IUPAC Name
InChI & InChI Key
"Normal" Display

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 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. These organizations will also need to pay a fee for certification.

Current 14 InChI Trust Members

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

8/16/2010

Current InChI Trust Supporters

National Institute of Chemistry – Ljubljana
Trinity University, San Antonio
University of North Carolina – Chapel Hill
Unilever Centre for Molecular Sciences Informatics, University of Cambridge

8/16/2010

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.

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, Peter Murray-Rust, 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.