Indexing Patent Chemistry with InChI

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

10/24/2011

Slides are available at http://www.hellers.com/steve/pub-talks/bcn-10-11/frame.htm

What is InChI ?

The IUPAC International Chemical Identifier, or InChI, is a machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner.

InChIs are produced by computer from structures drawn on-screen, and the original structure can be regenerated from an InChI with appropriate software.

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, InChIs are not designed to be read by humans. Or, put another way – never send a human to do a machine's job!

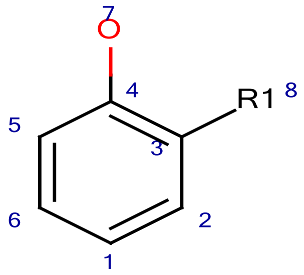
The InChI algorithm is Open Source, freely available, and is a non-proprietary identifier.

InChI for Patents

While the InChI algorithm covers some 99+% of everyday chemistry, one major area not yet handle are the generic structures associated with patents. Under IUPAC guidance a working group was established to enable InChI to represent generically described structures. The result of this effort a standard was defined for classes of compounds or specific structure scaffolds with variable substituents, a Generic InChI identifier would enable indexing of manually curated Markush databases as well.

The vendors participating in this agreed that the focus of Generic InChI should be to store all information for generic or Markush structures and therefore be used as a data exchange format.

A proposed syntax for a Generic InChI string is shown below, using different InChI layers to refer to different sets of generic variables.



…/C6H5O*1/1-2-3(-8)-4(-7)-5-6-1/(H layer)/(variable layer: information about attachment points, etc.)/…

```
Another option is to write out the core InChI with all its
representatives (R1 ...) as if they were atoms:

…/C6H5OR1/…
```

Information about the attachment points could be depicted thus:

v8:R1:3

where the No 8 atom is variable, its label is R1, it is connected to 3rd atom etc.)

InChI is not a replacement for any existing internal structure representations. InChI is IN ADDITION to what one uses internally. Its main value to most organizations is in LINKING information.

InChI links and connects the islands of independent silos of information on chemicals found throughout the Internet.

Why InChI? - Too Many Identifiers (“Standards”)

Structure diagrams

- various conventions
- contain ‘too much’ information

Connection Tables

- MolFiles, SMILES, ROSDAL, …

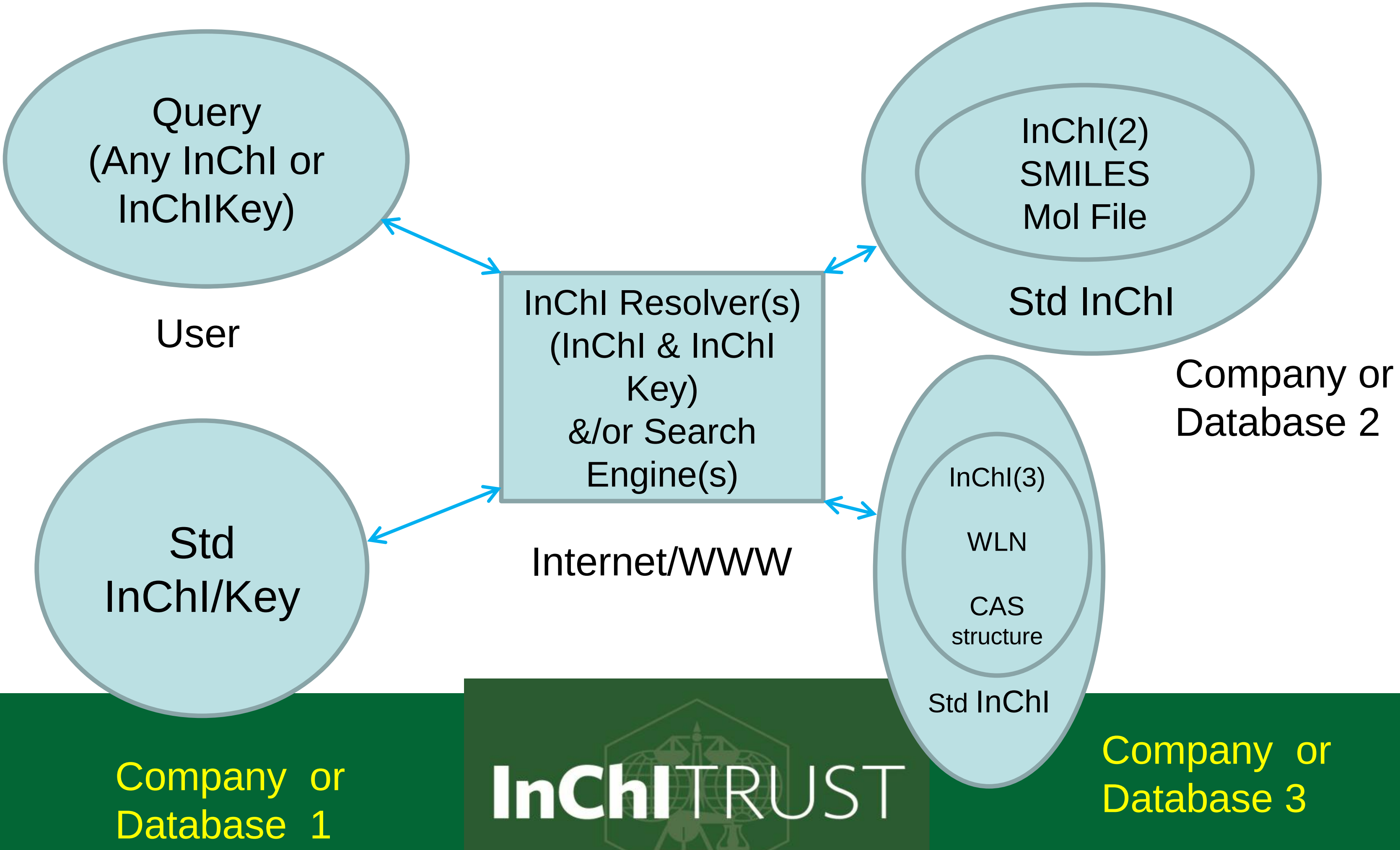
Pronounceable names (and mostly unpronounceable) and mostly complex names

- IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN

(Dumb) Index Numbers

- EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN , EC, ChemSpider ID, REACH, MFCD#, …

The LINKED and Interoperable and Combinable World of InChI and the InChIKey



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

CAS structure

Std InChI

Company or Database 3

Why Use InChI ?

For publishers, database providers, and organizations with one or more databases and with customers and stakeholders needing to access this information, using InChI and the InChIKey gives one an advantage being able to LINK and FIND content from multiple sources from independent silos found all over on the Internet.

It offers everyone the ability to more easily FIND existing information and data by easily being able to integrate, remix, and retell. InChI is a small, but vital, part of new organization models and technologies involving chemicals that will lead to improved efficiencies new discoveries. Combinability increases the value of information and data.

InChI will save time, resources, money – and find information!

How do we know the InChI project is beneficial?

Success is uncoerced adoption

What is in it for the various organizations?

The particular value of InChI is simple. The justification (or perhaps a better way to put it - the return on investment - ROI) is that the groups and their stakeholders can more easily and cost effectively, find the information they need – internally and externally. This will improve quality and the quantity of the results they obtain.

Put very explicitly, today there already are more InChIs in databases and information resources than any other chemical identifier because of two factors. One is that InChIs and InChIKeys are free. The second is that the Internet allows one to find information associated with an InChI and/or an InChIKey.

Who uses/searches InChIs?

InChIs are now found in virtually all major chemical databases, particularly in the very large ones. Databases such as Reaxys (30 million structures), NIH/PubChem (25 million structures), NIH/NCI (60 million structures), and SciFinder (55 million structures) all have InChIs and allow for InChIs as input for a search.

How difficult is it to create an InChI?

Today, all the major structure drawing programs (ChemDraw, MDL/Symyx/Accelrys Draw, ISIS Draw, ChemAxon Marvin Sketch, ACD Labs ChemSketch, Jmol and so on) have incorporated the InChI algorithm in their products, with usually an “InChI” button for generating the InChI.

The InChI Trust

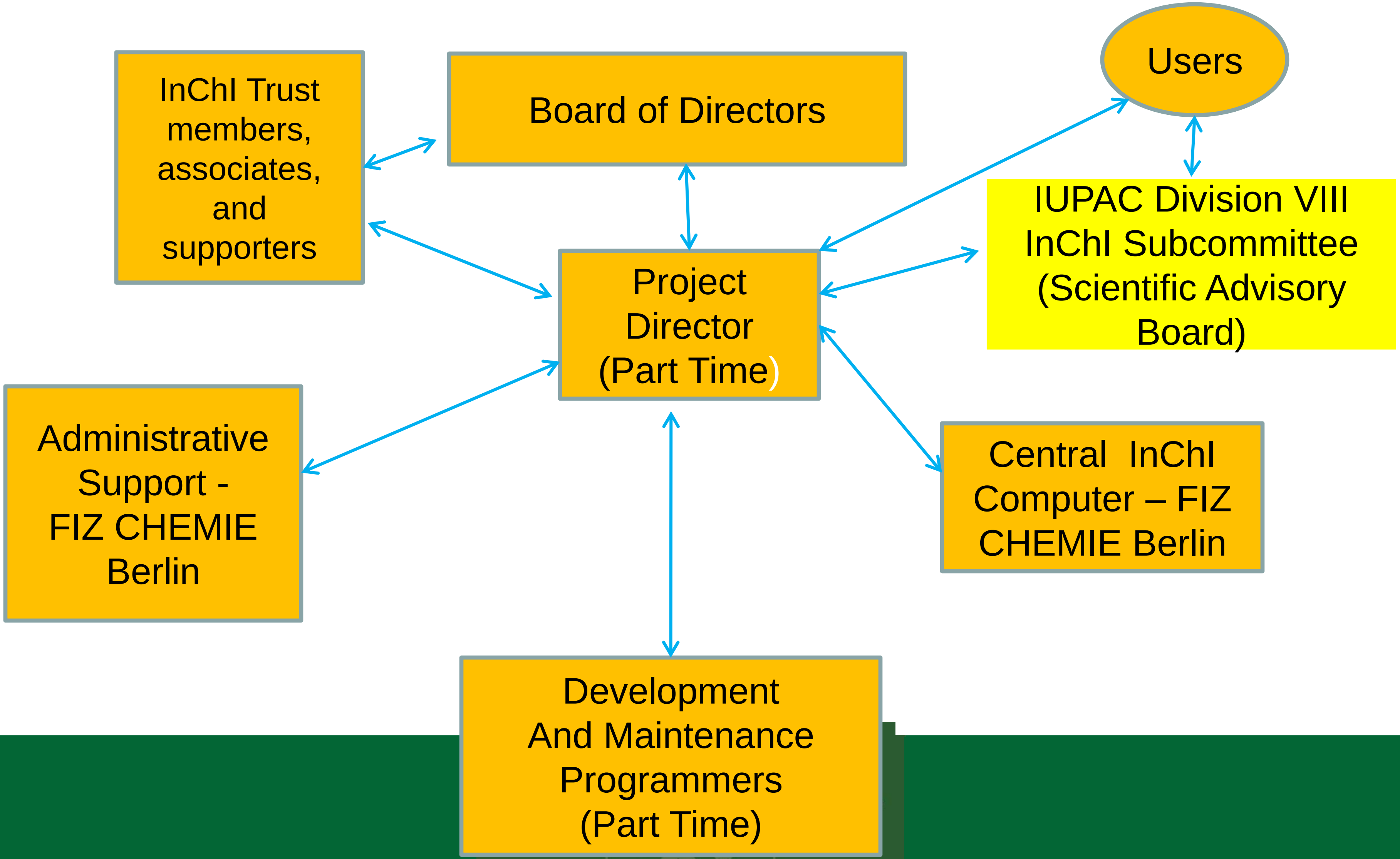
InChI had to be “institutionalized” and turned over to an entity that would ensure its ongoing activities and be acceptable to the community. It was concluded that a not-for-profit organization would best fit the ongoing and future project needs. Thus the decision to create and incorporate the "InChI Trust“ as a UK charity.

The InChI Trust (continued)

As there is no "free lunch", the Trust needs 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
CSIRO
Dialog
Elsevier Properties SA
FIZ CHEMIE
IBM Research
IUPAC
Informa / Taylor & Francis
Mcule
Nature Publishing Group
OpenEye
Royal Society of Chemistry
Springer
Wiley

* includes 1 being processed

16 as of 10/11

Current InChI Trust 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)
ChrisDS Consulting Limited (Chris Southan)
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)
Gesellschaft Deutscher Chemiker e.V. (GDCh), Germany (Wolfram Koch)
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)
KNCV - Dutch Chemical Society (Martin Post)
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)

29 as of 10/2011

InChI Trust Freeloaders

Too numerous to list

QA/QC - InChI Certification Suite

The InChI certification suite is a software package developed and designed to check that your installation of the InChI program has been performed correctly. The programs test your installation against a broad set of structures (which are provided with the Suite) to assure the InChIs and InChIKeys are correct and valid.

Once the certification package is run in-house, the results are sent back to the Trust, an "InChI certified" logo will be sent to person/organization. The InChI Trust certification logo can then be put on the pages of the web site for all users to see.

Unlike other Trust products (software and documentation) the Certification Suite is NOT free. It costs $5,000 per year.

The Future for InChI

Change seems to be occurring faster with the Internet. Even glaciers are moving a lot faster these days due to climate change.

It has been some 12 years since the InChI project was conceived. InChI has finally come of age and into the mainstream of chemical information.

The institutional support of InChI will assure its future.

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.