The Status of the IUPAC InChI Chemical Structure Standard – Today and the Future

Ray Boucher, Stephen Heller, and Alan McNaught

The main web sites for the IUPAC InChI project are:
http://www.iupac.org/inchi
and
http://www.inchi-trust.org **7/13/2017**

Slides are available at http://www.hellers.com/steve/sao-paulo-7-17.pdf

This is a green talk –

These slides were made from 100% recycled electrons

InChI Project Goal

To link everything about a chemical from many sources with the purpose of creating new information.

Today publishers have both scientific/chemical journals and chemical databases. Before InChI publishers of both forms of information and data were unable to connect and link the chemicals found in all these resources.

What is InChI?

The IUPAC International Chemical Identifier, or InChI, is a non-proprietary, 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 with existing structure drawing software, and the original structure can be regenerated from an InChI with existing structure drawing software.

InChI is really just a synonym.

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

Unique InChI Features

Only IUPAC International structure standard

Only Open Source structure standard

Only structure standard support by a wide majority of publishers, database producers, and chemistry software companies

For those in the audience who don’t want to hear the rest of the talk just pay attention to the next slide.

InChI Videos

1. What on Earth is InChI?

http://www.youtube.com/watch?v=rAnJ5toz26c

2. The Birth of the InChI

http://www.youtube.com/watch?v=X9c0PHXPfso

3. The Googlable InChIKey

http://www.youtube.com/watch?v=UxSNOtv8Rjw

4. InChI and the Islands

http://www.youtube.com/watch?v=qrCqJ0o4jGs

Four Requirements for a Computer Representation Standard

Need
Definition/Specification
Timing/Infrastructure
Acceptance/Use

Need

There was no open source (freely available) standard method to "name" a chemical structure. That is, a method to give a structure an electronic signature – an identifier.

Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be found and LINKED to and combined with other content on the Internet. InChI provides an excellent ROI (return on investment). InChI increases productivity!

Why InChI? - Too Many Good and Excellent Identifiers ("Standards")

Structure diagrams
- various conventions
- contain 'too much' information

Connection Tables/Notations
- MolFiles, SDF, SMILES, SLN, ROSDAL, …

Pronounceable names (and mostly unpronounceable) and mostly complex names
- IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN, ASK, ISO

(Dumb) Index Numbers
EINECS, ELINCS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN, EC, ChemSpider ID, REACH, PubChem CID, BAN, NSC, ASK, KEGG, BP, IND, MARTINDALE, MESH, IT IS, RX-CUI, NDF-RT, ATC, AHPA, USP/NF, UNII, MFCD#, and so on

“Standards are like toothbrushes – everyone has one but no one wants to use someone else's.”

Phil Bourne,
Former Associate Director for Data Science (Big Data), NIH

Definition/Specification

A computer algorithm to insure consistency and reproducibility and to be able to call it a real standard.

What "*is*" the InChI standard?

The InChI standard programmed into the algorithm is an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane). In cases of more complex molecules where there is not agreement among chemists, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the SAME result using it – which is what a standard is!

InChI Characteristics

1. Easy to generate

2. Expressive (it will contain structural information)

3. Unambiguous/Unique

4. Does not require a centralized operation (it can be generated anywhere – can use crowdsourcing/free labor)

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

InChI is for computers

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, and InChI QR codes - InChIs are not designed to be read by humans.

Or, put another way – never send a human to do a machine's job!

Technology is at its best when it is invisible.

How difficult is it to create an InChI?

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

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 November 11, 1947)

Timing &Infrastructure

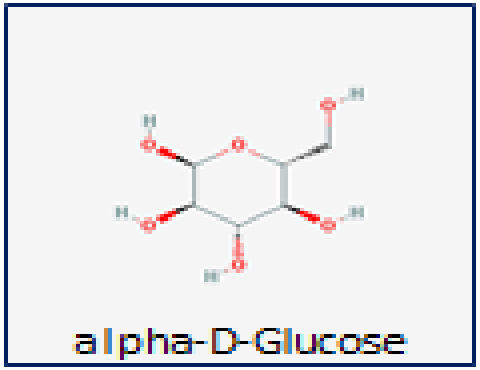
InChI has become a standard only because of the world has changed in the last 20 years.

Without the Internet, without vast amounts of data and information becoming available in computer readable form for the first time, without Google (and other search engines), without structure drawing programs, and with most chemistry publishers now needing chemical structures in their products, InChI would be yet another interesting graph theory project that died like so many before it.

Without this perfect good storm that created a foundation for InChI, at best, I would be talking to a group a 5-7 people at an IUPAC meeting talk.

What about SMILES as a standard?

C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O)O)O)O)O



- SMILES is a popular line notation
  - But not a published standard
- Every vendor has its own implementation
  - Differences in aromaticity models can lead to structure corruption
- Cannot reliably compare strings
  - Different software packages can make different strings for same structure
- No structure normalization
  - Different structural representations can yield different strings

Slide from Evan Bolton – NIH/PubChem

Re: [CHMINF-L] Inchi and chemical databases

You forwarded this message on 9/15/2010 5:37 PM.

CHEMICAL INFORMATION SOURCES DISCUSSION LIST [CHMINF-L@LISTSERV.INDIANA.EDU] on behalf of Ian A Watson

Sent: Wednesday, September 15, 2010 3:24 PM

To: CHMINF-L@LISTSERV.INDIANA.EDU

Interesting example of Caffeine smiles on the web site. I was able to generate 172 different smiles for the Caffeine molecule (email me if you'd like them). Presumably each one of these could be a unique smiles in somebody's implementation.

But when I converted each of those 172 different smiles to InChI, I got the exact same InChI string for each one. That's exactly how things are supposed to work. Nice.

Ian Watson

```
c1(=O)c2c(n(C)c(=O)n1C)ncn2C
c12c(n(C)c(=O)n(C)c1=O)ncn2C
O=c1n(C)c(=O)c2c(ncn2C)n1C
Cn1c2c(nc1)n(C)c(=O)n(C)c2=O
c12c(ncn1C)n(C)c(=O)n(c2=O)C
O=c1c2c(ncn2C)n(c(=O)n1C)C
c12c(n(cn1)C)c(=O)n(C)c(=O)n2C
Cn1c2c(nc1)n(c(=O)n(c2=O)C)C
c12c(ncn1C)n(c(=O)n(C)c2=O)C
c12c(ncn1C)n(C)c(=O)n(C)c2=O
Cn1c(=O)n(C)c(=O)c2c1ncn2C
n1(c2c(nc1)n(C)c(=O)n(C)c2=O)C
c12c(n(C)cn1)c(=O)n(c(=O)n2C)C
Cn1c(=O)c2c(ncn2C)n(c1=O)C
n1cn(C)c2c1n(c(=O)n(c2=O)C)C
n1cn(c2c1n(C)c(=O)n(c2=O)C)C
c12c(c(=O)n(c(=O)n1C)C)n(C)cn2
c1nc2c(n1C)c(=O)n(C)c(=O)n2C
c1(=O)n(C)c(=O)c2c(ncn2C)n1C
O=c1n(c(=O)c2c(ncn2C)n1C)C
Cn1cnc2c1c(=O)n(C)c(=O)n2C
n1(c(=O)n(c(=O)c2c1ncn2C)C)C
c1(=O)n(C)c(=O)c2c(n1C)ncn2C
O=c1n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(n(cn2)C)c(=O)n(c1=O)C
Cn1c(=O)c2c(n(c1=O)C)ncn2C
Cn1cnc2c1c(=O)n(c(=O)n2C)C
c1nc2c(c(=O)n(C)c(=O)n2C)n1C
c12c(ncn1C)n(c(=O)n(c2=O)C)C
c1nc2c(n1C)c(=O)n(c(=O)n2C)C
Cn1c2c(n(cn2)C)c(=O)n(C)c1=O
n1(C)c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c2c(nc1)n(C)c(=O)n(c2=O)C
n1(c(=O)c2c(n(c1=O)C)ncn2C)C
n1(c(=O)c2c(n(C)c1=O)ncn2C)C
Cn1c(=O)n(c2c(c1=O)n(C)cn2)C
n1(C)c(=O)n(C)c2c(=O)n(C)c2=O
c1nc2c(n1C)c(=O)n(C)c(=O)n2C
c1(=O)n(c(=O)c2c(ncn2C)n1C)C
n1(cnc2c1c(=O)n(c(=O)n2C)C)C
n1(C)c(=O)n(C)c2c(n(cn2)C)c1=O
n1(c2c(nc1)n(C)c(=O)n(c2=O)C)C
n1(C)c(=O)c2c(n(c1=O)C)ncn2C
n1(c2c(c(=O)n(C)c1=O)n(cn2)C)C
c12c(n(c(=O)n(c1=O)C)C)ncn2C
n1cn(C)c2c1n(C)c(=O)n(c2=O)C
c12c(c(=O)n(C)c(=O)n1C)n(cn2)C
Cn1c2c(n(C)cn2)c(=O)n(c1=O)C
n1(c(=O)n(C)c2c(n(cn2)C)c1=O)C
n1cn(c2c1n(C)c(=O)n(C)c2=O)C
c1(=O)n(c2c(c(=O)n1C)n(C)cn2)C
Cn1c(=O)n(c(=O)c2c1ncn2C)C
O=c1n(c(=O)c2c(n(C)cn2)C)C)C
n1(c2c(c(=O)n(c1=O)C)n(C)cn2)C
c12c(n(cn1)C)c(=O)n(c(=O)n2C)C
c12c(c(=O)n(C)c(=O)n1C)n(C)cn2
Cn1c(=O)c2c(n(C)c1=O)ncn2C
```

```
c1(=O)n(C)c2c(n(cn2)C)c(=O)n1C
O=c1n(C)c2c(c(=O)n1C)n(C)cn2
n1(C)c2c(c(=O)n(C)c1=O)n(C)cn2
n1cn(c2c1n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(C)c2c1n(cn2)C)C
c1(=O)c2c(n(c(=O)n1C)C)ncn2C
c1(=O)n(c2c(n(cn2)C)c(=O)n1C)C
Cn1c2c(c(=O)n(c1=O)C)n(cn2)C
c1(=O)n(c(=O)c2c(n1C)ncn2C)C
O=c1n(c(=O)c2c(n1C)ncn2C)C
n1cn(C)c2c1n(c(=O)n(C)c2=O)C
n1(c(=O)n(C)c2c(c1=O)n(C)cn2)C
O=c1c2c(ncn2C)n(C)c(=O)n1C
n1cnc2c1c(=O)n(C)c(=O)n2C)C
n1(C)cnc2c1c(=O)n(C)c(=O)n2C)C
n1cn(C)c2c1n(C)c(=O)n(C)c2=O
O=c1n(C)c(=O)n(C)c2c1n(C)cn2
n1(C)c(=O)n(c2c(c1=O)n(C)cn2)C
Cn1c(=O)c2c(ncn2C)n(C)c1=O
n1(c2c(n(cn2)C)c(=O)n(c1=O)C)C
n1(C)c2c(n(C)c(=O)n(c2=O)C)C)nc1
Cn1c2c(n(c(=O)n(c2=O)C)C)nc1
n1(c(=O)n(C)c(=O)c2c1ncn2C)C
O=c1n(C)c2c(n(C)cn2)c(=O)n1C
n1(C)c2c(n(cn2)C)c(=O)n(C)c1=O
c1(=O)c2c(ncn2C)n(c(=O)n1C)C
O=c1n(c2c(c(=O)n1C)n(C)cn2)C
Cn1c2c(n(C)c(=O)n(C)c2=O)nc1
Cn1c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c2c(n(C)cn2)c(=O)n(C)c1=O
c12c(n(C)c(=O)n(c1=O)C)ncn2C
n1(c2c(c(=O)n(c1=O)C)n(cn2)C)C
c1(=O)n(C)c(=O)n(c2c1n(cn2)C)C
n1(c2c(n(C)cn2)c(=O)n(c1=O)C)C
c1(=O)n(c2c(n(C)cn2)c(=O)n1C)C
n1(c2c(nc1)n(C)c(=O)n(c2=O)C)C
Cn1c2c(nc1)n(C)c(=O)n(c2=O)C
c12c(c(=O)n(c(=O)n1C)C)n(cn2)C
Cn1c2c(n(c(=O)n(C)c2=O)C)nc1
c1(=O)n(c(=O)n(C)c2c1n(C)cn2)C
c1(=O)n(C)c2c(n(C)cn2)c(=O)n1C
n1(c(=O)c2c(ncn2C)n(C)c1=O)C
n1(C)c2c(n(C)c(=O)n1C)n(C)cn2
n1(C)c2c(nc1)n(C)c(=O)n(c2=O)C
n1(C)c2c(n(cn2)C)c(=O)n(c1=O)C
n1(C)c(=O)n(C)c2c(c1=O)n(C)cn2
n1(C)c2c(c(=O)n(C)c1=O)n(cn2)C
n1(c(=O)n(c2c(n(C)cn2)c1=O)C)C
n1(c(=O)n(c2c(c1=O)n(C)cn2)C)C
n1(C)c2c(n(C)cn2)c(=O)n(C)c1=O
n1(C)c2c(c(=O)n(C)c1=O)C)n(C)cn2
n1(C)c2c(n(c(=O)n(C)c2=O)C)nc1
n1(C)c2c(nc1)n(c(=O)n(c2=O)C)C
```

```
O=c1c2c(n(c(=O)n1C)C)ncn2C
O=c1n(C)c2c(n(cn2)C)c(=O)n1C
n1(C)c(=O)n(c2c(n(C)cn2)c1=O)C
n1(C)c2c(c(=O)n(c1=O)C)n(cn2)C
Cn1c2c(c(=O)n(C)c1=O)n(C)cn2
c1(=O)n(c2c(c(=O)n1C)n(cn2)C)C
n1(c2c(n(C)c(=O)n(c2=O)C)nc1)C
n1(c2c(c(=O)n(C)c1=O)n(C)cn2)C
n1(C)c(=O)c2c(ncn2C)n(c1=O)C
Cn1c(=O)n(C)c2c(n(cn2)C)c1=O
O=c1n(C)c(=O)c2c(n1C)ncn2C
n1(c(=O)n(c2c(c1=O)n(cn2)C)C)C
O=c1n(c(=O)n(C)c2c1n(C)cn2)C
n1(C)c(=O)n(c2c(n(cn2)C)c1=O)C
n1(c(=O)n(C)c2c(n(C)cn2)c1=O)C
c1(=O)n(C)c(=O)n(C)c2c1n(cn2)C
n1(c(=O)n(c2c(c1=O)n(cn2)C)C)C
O=c1n(C)c(=O)n(c2c1n(cn2)C)C
n1(c(=O)c2c(ncn2C)n(c1=O)C)C
c1(=O)n(C)c(=O)n(C)c2c1n(cn2)C
O=c1n(C)c(=O)n(c2c1n(cn2)C)C
n1(C)c(=O)n(c2c(n(C)cn2)c1=O)C
n1(c(=O)n(C)c2c(n(C)cn2)c1=O)C
c1(=O)n(C)c(=O)n(C)c2c1n(cn2)C
n1(c(=O)n(c2c(n(C)cn2)c1=O)C)C
O=c1n(c(=O)n(C)c2c1n(C)cn2)C
n1(C)c2c(n(C)cn2)c(=O)n(C)c1=O
n1(C)c2c(c(=O)n(C)c1=O)C)n(C)cn2
n1(C)c2c(n(c(=O)n(C)c2=O)C)nc1
n1(C)c2c(nc1)n(c(=O)n(c2=O)C)C
```

Current InChI Status

At present, practically speaking, InChI can handle simple organic molecules, which turns out to cover 99%+ of what people deal with every day. If it did not the every day needs of chemists and information specialists then the usage of InChI would not be as great as it is.

Large Databases with InChIs/InChIKeys

EBI UniChem –144 million
NIH/NCI – 110 million
NIH/PubChem - 91 million
EBI UniChem –144 million
RSC/ChemSpider – 34 million
Elsevier/Reaxys – 30 million
IUPAC – 0 million

Why is InChI a Success

InChI is able to put things together in a new way. We took IUPAC, the Internet, Open Source software, crowdsourcing (SourceForge), Graph theory, existing representation algorithms, digitized data available on the web, and search engines, combines them, and created a very valuable tool.

InChI only works because of new technology. Without these factors above, for all practical purposes, no one would even know InChI existed.

Success is uncoerced adoption

InChI is not a replacement for any existing internal structure representations. InChI is in ADDITION to what one uses internally. Its value to chemists is in FINDING and LINKING information

InChI Staff and Collaborators

The InChI project has had the unusual perfect "good storm" of cooperation and support. It is a truly international project with programming in Moscow, computers in the cloud, incorporated in the UK, and a project director in the USA. Collaborators from over a dozen countries, from academia, Pharma, publishers, and the chemical information industry, have all offered, and continue to offer, senior scientific staff to develop the InChI standard.

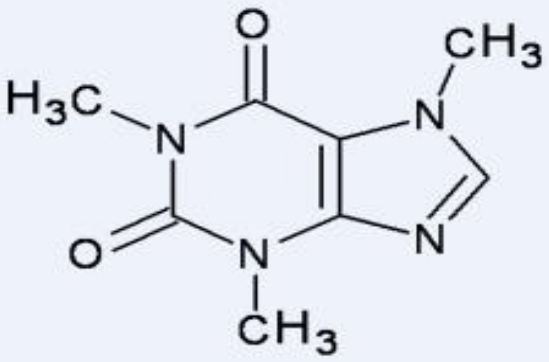
InChI layered structure design

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 Algorithm normalizes chemical representation and includes a "standardized" InChI, and the 'hashed' form called the InChIKey



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)

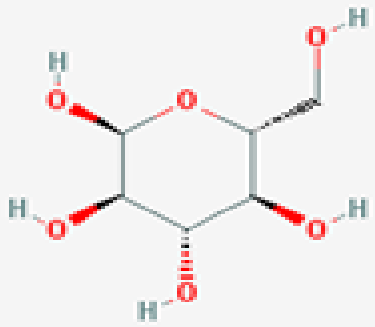
Encodes stereochemistry and isotopes

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

InChI is a string

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

Version/Type
Chemical formula
Connectivity
Charge/Proton
Stereochemical
Other (e.g., Isotopic)



alpha-D-Glucose

"layered" line notation

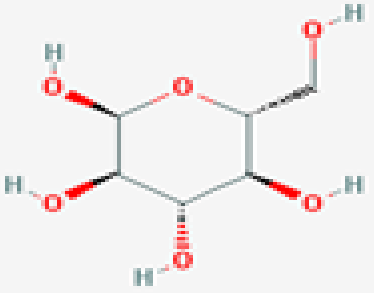
InChIKey is a "hashed" InChI

- Search engine friendly InChI
- May allow for 'secure' lookup of a chemical

WQZGKKKJIJFFOK-DVKNGEFBSA-N

Chemical formula
Connectivity
Stereochemical
Other (e.g., Isotopic)
Type
Version
Charge/Proton

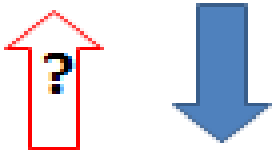
"layered" line notation



alpha-D-Glucose

InChIKey can be a 'secret'

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



WQZGKKKJIJFFOK-DVKNGEFBSA-N

There is no chemical information in an InChIKey … if you do not know the InChI, you cannot convert the InChIKey back into a chemical structure

Slide from Evan Bolton/NIH/PubChem

Search Engines can use InChIKey

They can use InChI too! .. but your mileage may vary

https://www.google.com/search?q=BSYNRYMUTXBXSQ-UHFFFAOYSA-N&rlz=1C1CHFX_enUS4

Google YHVFECVVGNXFKO-UHFFFAOYSA-N

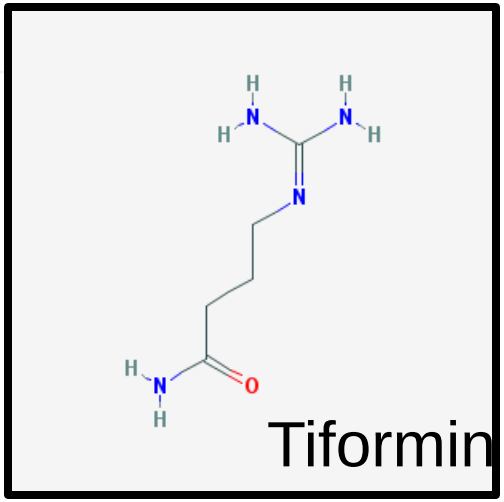
Web Maps Shopping Images News More Search tools

About 100 results (0.32 seconds)

ChemIDplus - 4210-97-3 - YHVFECVVGNXFKO ...
chem.sis.nlm.nih.gov/.../4210... United States National Library of Medicine
4210-97-3 - **YHVFECVVGNXFKO-UHFFFAOYSA-N** - Tiformin [INN:BAN] - Similar structures search, synonyms, formulas, resource links, and other chemical ...

tiformin - PubChem
pubchem.ncbi.nlm.nih.gov › ... › PubChem PubChem
Structure, classification, information, physical and chemical properties for ... Molecular Weight: 144.17498 InChIKey: **YHVFECVVGNXFKO-UHFFFAOYSA-N**.

Compound Name and Classification - Compound Report Card
https://www.ebi.ac.uk/.../index.../1477675 European Bioinformatics Institute
... InChI, InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4, ... Download InChI. Standard InChI Key, **YHVFECVVGNXFKO-UHFFFAOYSA-N** ...



InChI/InChIKey Use and Utility

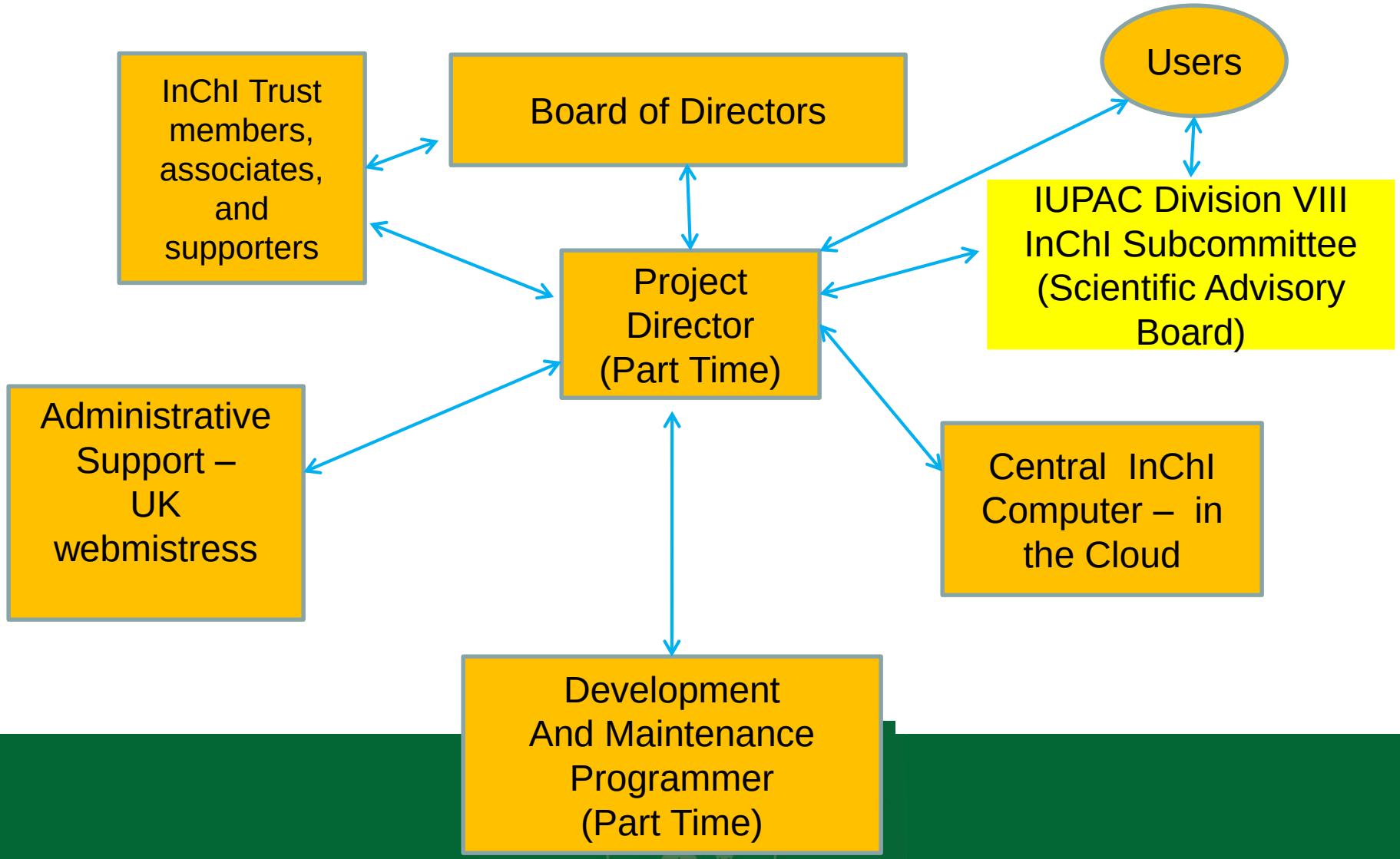
- InChI
  - Enabler of data exchange
  - Provides chemical structure normalization

- InChIKey
  - Compact form for structure lookup
  - Allows “secret” chemical information exchange

The InChI Trust

To function and succeed, InChI had to become personality independent. InChI had to be “institutionalized”. If the work of this project was to be enduring it needed to 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.

InChI Trust Organization



Project Director
(as performed by Steve Heller)

The project Director oversees all aspects of the project. The IUPAC InChI subcommittee working groups defining the standards, the programming of these standards, lecturing on InChI, organizing meetings and workings on InChI.

In other words Steve is like Mark Twain's Tom Sawyer, talking people into doing the real work - like flying to Sao Paulo to give this talk.

InChI characteristics

Consensus
Technical competence
Political and technical cooperation
Precompetitive collaboration – publishers, databases, software
No competition with commercial products
No mission creep
IUPAC blessing/endorsement & rapid IUPAC acceptance
Excellent understanding of what the Internet and how it can be effectively used in Chemical Information

Vision of the future

Current IUPAC Working Groups & Projects

Completed:

Revised FAQ's from Cambridge- Nick Day/Peter Murray-Rust
InChI Certification Suite
Version 1.05 released – 2017
Markush (contract to be signed when funded)
Polymers
RInChI – InChI for Reactions
New API

Started/To be started

Mixtures
InChI Resolver
QR codes for InChI
InChI teaching/educational materials
Large Molecules/Biopolymers/Macromolecules
Inorganics
Positional Isomers
Redesign of Handling of Tautomerism

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.

Keep Calm and Use InChI

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, Ray Boucher, Steve Boyer, Steve Bryant, Szabolcs Csepregi , Rene Deplanque, Gary Mallard, 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.