InChI & the Publication and Information Chain

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 **8/12/2014**

Slides are available at http://www.hellers.com/steve/pub-talks/sfo-8-14.pdf

This is a green talk –

These slides were made from 100% recycled electrons

My clone is giving this lecture.

I am actually somewhere else having a much better time.

What is InChI ?

The IUPAC International Chemical Identifier structure representation standard, or InChI, is a non-proprietary, freely available, 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

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#, …

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

Phil Bourne, Associate Director for Data Science, NIH

What "*is*" the InChI standard?

The InChI standard/rules that are programmed into the algorithm are an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., CH_4 - 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 is plumbing/infrastructure. InChI is an (enabling) tool. It is a means to an end. InChI is a modern enabling technology.

For all but small group of chemists developing it, InChI is not something anyone should want to know about.

All you want to do is use InChI to teach your students and researchers how to find information on the web in publications, databases, and other resources (e.g., chemical catalogs)

InChI is helping chemists to do better work and find/link to the latest information. InChI is the infrastructure foundation that allows for higher productivity

InChI is not a replacement for any existing internal structure representations. (We do not start religious wars.) InChI is in ADDITION to what one uses internally. Its value to student or scientist is in FINDING and LINKING information

Internal:
Your representation (e.g. WLN)
Your formats

External:
Same representation (Standard InChI/InChIKey)
Same format

The problem with too much information on the Internet: **Lack of integration and connection.**

multiple applications
multiple repositories/publishers/journals
multiple interfaces and protocols

We have a tower of Babel.

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.

InChI YouTube 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

How do I create an InChI ?

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 appropriate software (e.g., ACD ChemSketch, ChemAxon Marvin, MDL/Symyx/Accelrys/BIOVIA Draw & PE ChemDraw).

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

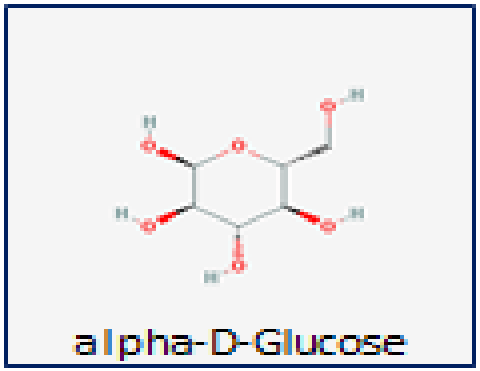
c1(=O)c2c(n(C)c(=O)n1C)ncn2C	c1(=O)n(C)c2c(n(cn2)C)c(=O)n1C	O=c1c2c(n(c(=O)n1C)C)ncn2C
c12c(n(C)c(=O)n(C)c1=O)ncn2C	O=c1n(C)c2c(c(=O)n1C)n(C)cn2	O=c1n(C)c2c(n(cn2)C)c(=O)n1C
O=c1n(C)c(=O)c2c(ncn2C)n1C	n1(C)c2c(c(=O)n(C)c1=O)n(C)cn2	n1(C)c(=O)n(c2c(n(C)cn2)c1=O)C
Cn1c2c(nc1)n(C)c(=O)n(C)c2=O	n1cn(c2c1n(c(=O)n(C)c2=O)C)C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

InChI

172 SMILES representations

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

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 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, Blekko etc.) using the InChI (hash) Key.

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

InChI as a web index for molecules

"We have now discovered, serendipitously, that these InChIs have been comprehensively and accurately indexed by the Google search engine. From preliminary exploration it appears that every known document in which an InChI appears has been indexed and that all are retrievable by standard queries with virtually 100% precision. This means that standard Web-based indexers, without any alteration, are capable of acting as completely precise chemical search engines. Although we have many years of developing chemistry on the web, this was an unexpected and very welcome finding"

Murray-Rust et al. 2004 http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/

Where are InChIs?

PubChem ~ 50 million
ChemSpider ~ 30 million
Reaxys ~ 30 million
PubChem from patents (all sources) ~ 15 million
PubChem journal sources (PubMed + ChEMBL) ~ 1 million
SciFinder ~ 60 million (estimated as input for searches)
Web sources outside the above (no idea)

Chris Southan BioIT 2014 lecture

InChI is an international computer readable standard not just for chemists, but rather has very wide technical and non-technical use for linking and connecting information in many areas of scientific and everyday activities - -

abstracting services
biochemistry
biology/genomics databases
bio-activity databases
books
chemical biology
chemical spills
chemistry databases
clinical trials
company annual reports
drug discovery
drug information
drug overdoses
electronic books
environmental information
food additives
lawsuits
magazines
medicinal chemistry
medical information
medical records
metabolomics
newspapers
patents
packages/bottles/transportation labels/ everyday product labels
pharmacology
scientific journals
toxicology
toxicological information

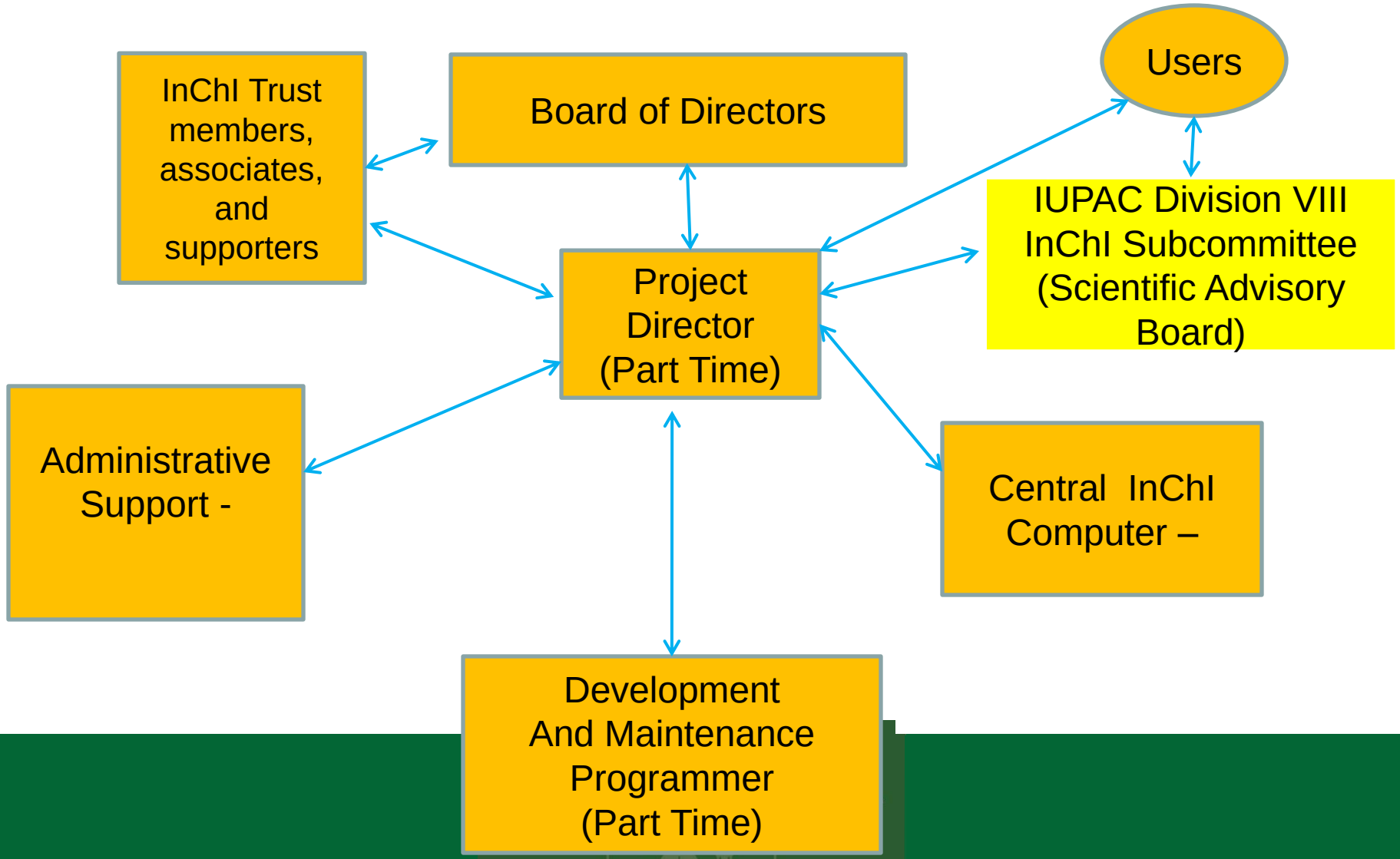
Critical words/phrases for InChI

Link
Addition; not replacement
Algorithm
Synonym
No bureaucracy

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



Total number of Members, Associate Members, and (non paying) Supporters ~60

(Please consider joining !!)

“Evidence indicates that personal relations count a lot in international politics, as in life.”

Reagan at Reykjavik by Ken Adelman

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 senior scientific staff to develop the InChI standard.

Why InChI is a success (#1)

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. InChI provides an excellent ROI (return on investment). InChI increases productivity!

2. InChI is a freely available, Open Source, algorithm that anyone, anywhere can freely use. And they sure use it!

Success is uncoerced adoption

Why is InChI a Success (#2)

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 no one would even know InChI existed.

Combining existing known things is not new. Kary Mullis, 1993 Chemistry Nobel Laureate said his PCR work was "just recombining known things" in a new way.

Why is InChI a Success (#3)

InChI improves productivity by taking existing resources and making them more valuable by being able to easily find them and put them together and analyze/use them more efficiently and effectively.

Again, we are taking advantage of what is called the second machine age *, which includes “recombinant innovation” or mashups.

***The Second Machine Age**
Work, Progress, and Prosperity in a Time of Brilliant Technologies
Authors: Andrew McAfee & Erik Brynjolfsson

Bypassing IUPAC procedures

The usual very, lengthy IUPAC approval process was hijacked and sped up by sending the IUPAC bureaucracy, not a white paper with InChI rules, but rather the coding of these rules which were unreadable and unintelligible C code to non-programmers.

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

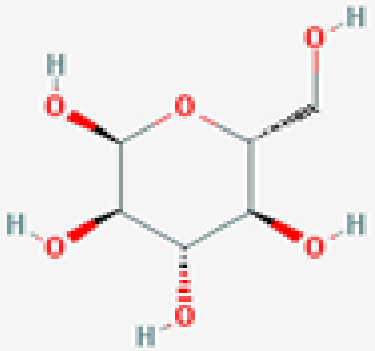
How did InChI succeed?

This project was the perfect "good" storm. The project came about in 1999 when Steve Heller retired and his wife threatened him with divorce unless he found some to do. (Yes, behind every successful project is a woman.) IUPAC discovered that nomenclature was for 20th, not 21st century. NIST, the US standards agency, needed a way to represent and link the structures from its standard property databases. The Internet (web 2.0) was taking off enabling silos and islands of information to be linked and searched if only there was a linking element. Publishers and database producers realized their information would be more valuable (i.e., they could sell more to more people) if only there was a way to link chemical structures from all the diverse resources on the Internet. With no funds to support the project, IUPAC needed the private sector to pay for the short and long term project needs. Lastly, the decentralized structure and hands-off management of the project enabled all the expert egos to be satisfied by putting everyone in charge of what they do best and giving them the final say - allowing for proper, scientific, bottom-up decisions.

InChI is a string

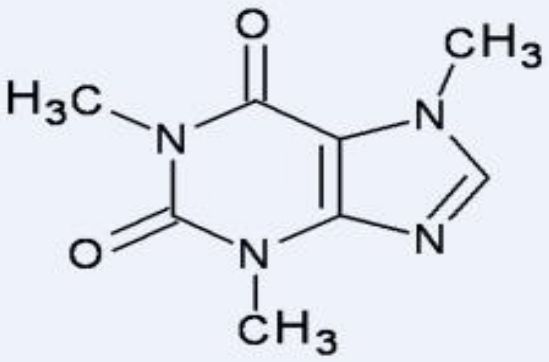
InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

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



alpha-D-Glucose

"layered" line notation



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)

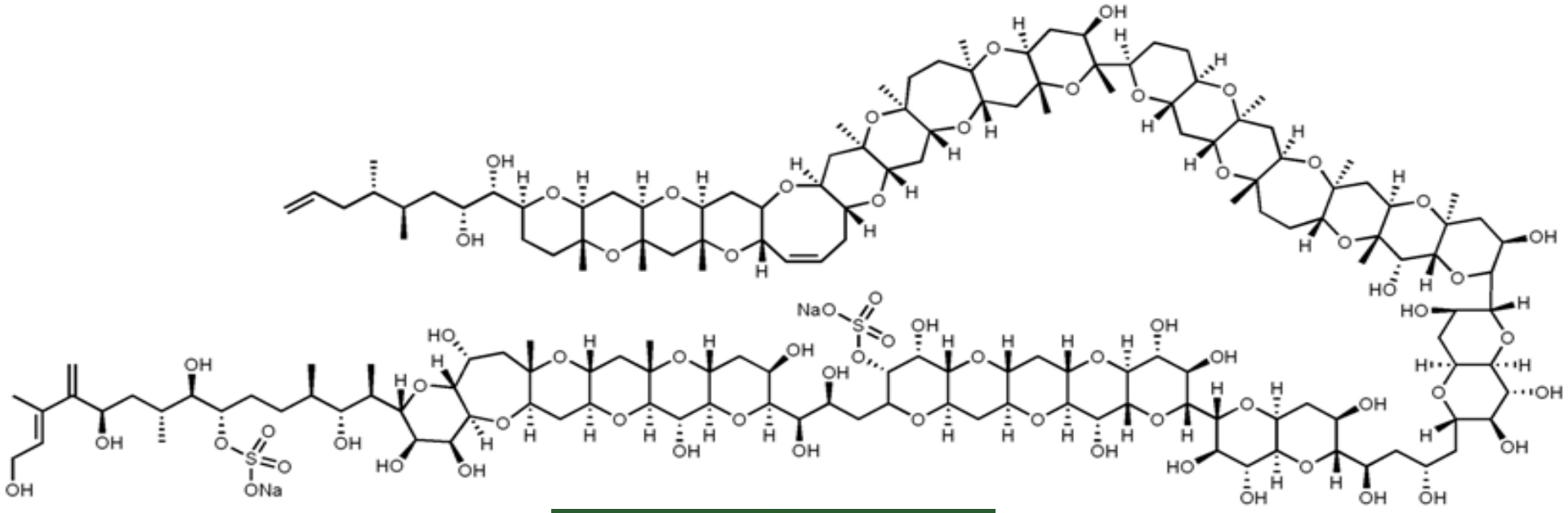
Second block (8 letters)

Encodes stereochemistry and isotopes

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

InChI for Maitotoxin (from NextmoveSoftware)

InChI=1S/C164H258O68S2.2Na/c1-24-26-65(2)68(5)41-74(168)117(179)85-33-36-152(11)106(203-85)55-109-162(21,231-152)64-161(20)105(210-109)51-89-83(220-161)28-25-27-82-99(199-89)59-157(16)108(202-82)56-107-153(12,230-157)39-38-151(10)112(211-107)61-158(17)111(224-151)54-101(176)163(22,232-158)103-32-31-84-90(204-103)53-110-156(15,219-84)62-113-150(9,223-110)37-34-102-155(14,225-113)63-114-164(23,227-102)147(192)149-159(18,226-114)58-81(175)134(218-149)133-79(173)47-93-136(216-133)120(182)119(181)92(200-93)44-72(166)43-76(170)131-77(171)46-94-137(214-131)122(184)124(186)143(207-94)145-126(188)125(187)144-146(217-145)128(190)139-97(208-144)50-88-87(206-139)49-96-138(205-88)127(189)141(229-234(196,197)198)95(201-96)45-75(169)118(180)132-78(172)48-98-140(215-132)129(191)148-160(19,221-98)60-100-91(209-148)52-104-154(13,222-100)57-80(174)135-142(212-104)123(185)121(183)130(213-135)71(8)115(177)67(4)29-30-86(228-233(193,194)195)116(178)69(6)42-73(167)70(7)66(3)35-40-165;;/h24-25,28,35,65,67-6 9,71-149,165-192H,1,7,26-27,29-34,36-64H2,2-6,8-23H3,(H,193,194,195)(H,196,197,198);;/q;2*+1/p-2/b28-25-,66-35+;;/t65-,67+,68+,69+,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+,101+,102+,103-,104+,105-,106-,107+,108-,109+,110+,111-,112-,113+,114+,115+,116+,117-,118+,119-,120+,121+,122+,123-,124+,125+,126+,127+,128+,129-,130-,131-,132-,133+,134+,135+,136+,137+,138+,139-,140+,141-,142-,143+,144-,145+,146+,147-,148+,149+,150-,151+,152+,153-,154-,155-,156-,157+,158+,159-,160-,161+,162-,163+,164+;;/m0../s1



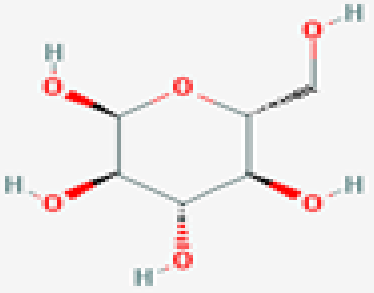
InChIKey is a "hashed" InChI

- 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

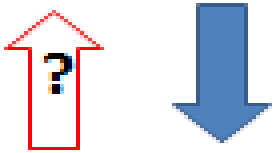
"layered" line notation



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

About 1,370 results (0.28 seconds)

Caffeine | C8H10N4O2 | ChemSpider
www.chemspider.com/Chemical-Structure.2424.html ChemSpider
Structure, properties, spectra, suppliers and links for: Caffeine, 58-08-2. ... ChemSpider - Search and share chemistry ... **RYYVLZVUVIJVGH-UHFFFAOYSA-N**
You've visited this page 3 times. Last visit: 7/31/14

caffeine
webbook.nist.gov/.../cbo... National Institute of Standards and Technology
IUPAC Standard InChIKey: **RYYVLZVUVIJVGH-UHFFFAOYSA-N**; CAS Registry Number: 58-08-2; Chemical structure: C8H10N4O2 This structure is also ...

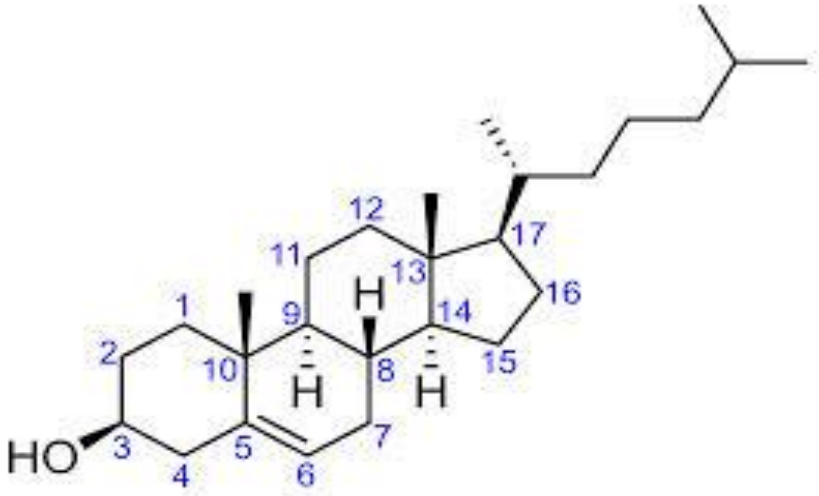
RYYVLZVUVIJVGH-UHFFFAOYSA-N - BRENDA
enzyme-information.info/php/ligand_flatfile.php4?brenda_ligand_id...
Information on enzyme ligand caffeine (8183) - InchiKey: **RYYVLZVUVIJVGH-UHFFFAOYSA-N**.

RYYVLZVUVIJVGH-UHFFFAOYSA-N - PubChem ...
www.ncbi.nlm.nih.gov/pcco... National Center for Biotechnology Information
1. caffeine; Methyltheobromine; Guaranine ... MW: 194.190600 g/mol MF: C8H10N4O2 IUPAC name: 1,3,7-trimethylpurine-2,6-dione CID: 2519.

ChemIDplus - 58-08-2 - RYYVLZVUVIJVGH-UHFFFAOYSA ...
chem.sis.nlm.nih.gov/.../58-0... United States National Library of Medicine
58-08-2 - **RYYVLZVUVIJVGH-UHFFFAOYSA-N** - Caffeine [USP:BAN:JAN] - Similar structures search, synonyms, formulas, resource links, and other chemical ...

Structure Searching with InChI

An accidental feature of the InChIKey, discovered by an InChI collaborator (Nicko Goncharoff working on structures in a patent database), is that the first 14 characters can be used to search for structures with the same skeleton.



Google search for the InChIKey for Cholesterol:
HVYWMOMLDIMFJA-DPAQBDIFSA-N
698 hits

Google search for the InChIKey for Cholesterol Skeleton
HVYWMOMLDIMFJA
1470 hits

InChI TRUST

QA/QC - InChI Certification Suite

The InChI certification suite is a software package 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. Only this way is it possible to know that the InChIs have been generated properly and consistently.

Unlike other Trust products (software and documentation) the Certification Suite is NOT free, except to members and supporters who use for non-commercial activities. It costs $5,000 per year.

So far InChI has taken all the low hanging fruit/structures (small organic molecules) and created a valuable tool. The next steps to expand InChI to handle more complex chemical structures is underway.

Current IUPAC Working Groups & Projects

In Progress/Almost Final:
Organometallics & InChI Resolver

Completed:
Revised FAQ's from Cambridge- Nick Day/Peter Murray-Rust
InChI Certification Suite
Version 1.04 released – 9/11
Markush (contract to be signed when funded)
Polymers/Mixtures
RInChI – InChI for Reactions (contract to be signed in fall 2014)
New API
InChI Videos

Started/To be started in 2013/2014:
Electronic/Excited States
QR codes for InChI
InChI teaching/educational materials
Large Molecules/Biopolymers/Macromolecules/Proteins/Peptides, Enzymes
Positional Isomers
Crystal/3D structures
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.

Summary

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

Keep Calm and Use InChI

InChI world domination is proceeding on schedule.

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, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Gary Mallard , 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.

InChI TRUST

Have any questions?

If you think of a question later, email me:

steve@inchi-trust.org