InChI & Public Health Laboratories

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 **3/20/2014**

Slides are available at http://www.hellers.com/steve/aphl-3-14.pdf

No animals were harmed in the preparation of this presentation, however a number of WWW sites were hit.

This presentation is made from 100% recycled electrons

Feel free to ask questions anytime –

You can’t interrupt my train of thought - I don’t have one.

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.

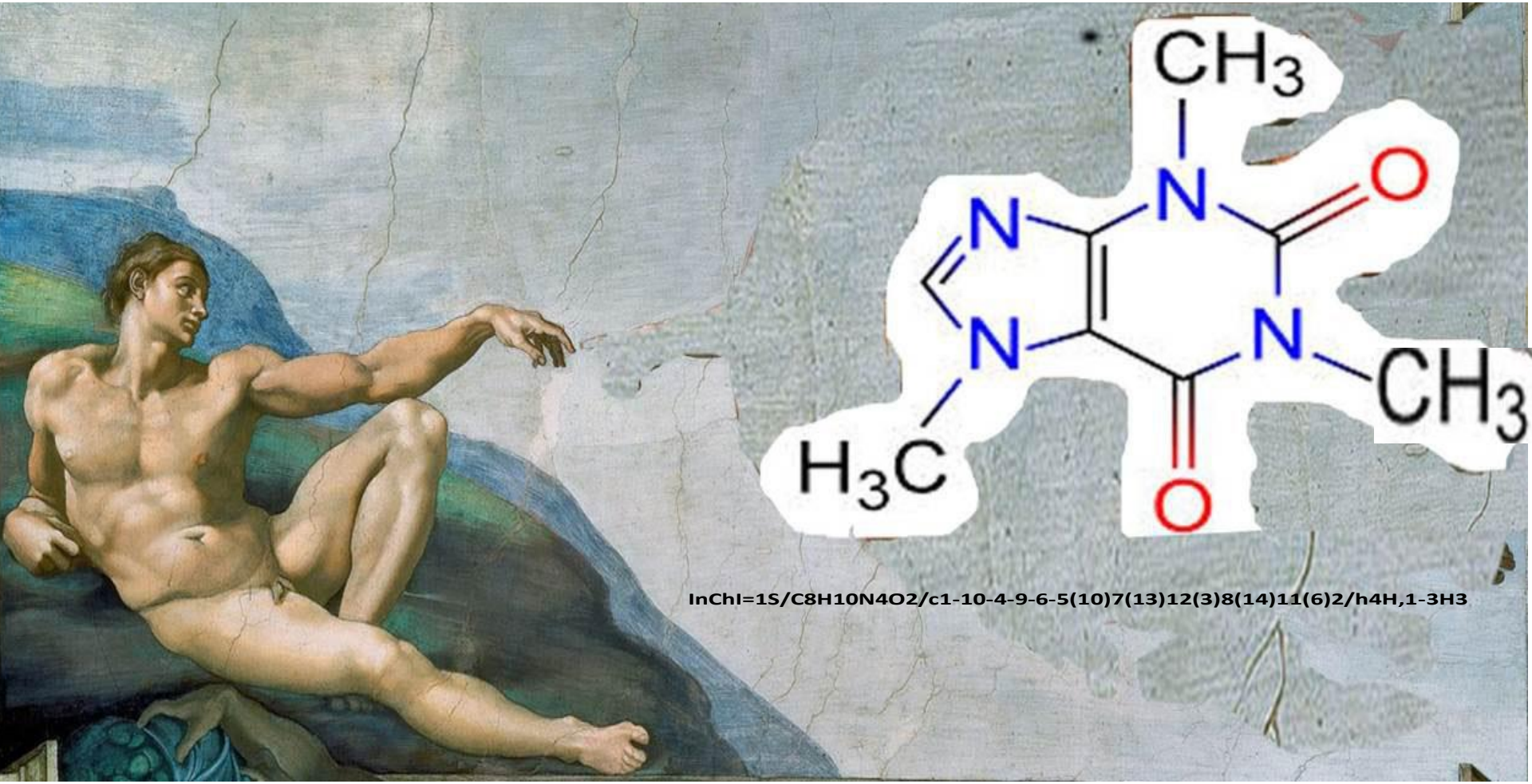
It is a long term project to create and obtain acceptance for an Open Source computer readable chemical structure representation by careful planning, worldwide cooperation and co-opting of chemical structure experts.

InChI is really just a synonym.

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

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.



The creation of InChI

InChI TRUST

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

InChI does not replace any internal, local system pieces. Your language and format remain as is.

But even though we communicate around the world in English, there are still over 2,500 language versions of the Bible.

InChI for external use only. The InChI standard is like universal language -- like English.

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.

What "*is*" the InChI standard?

The InChI standard programmed into the algorithm is a 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 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 is plumbing. InChI is an (enabling) tool. InChI is a modern enabling technology.

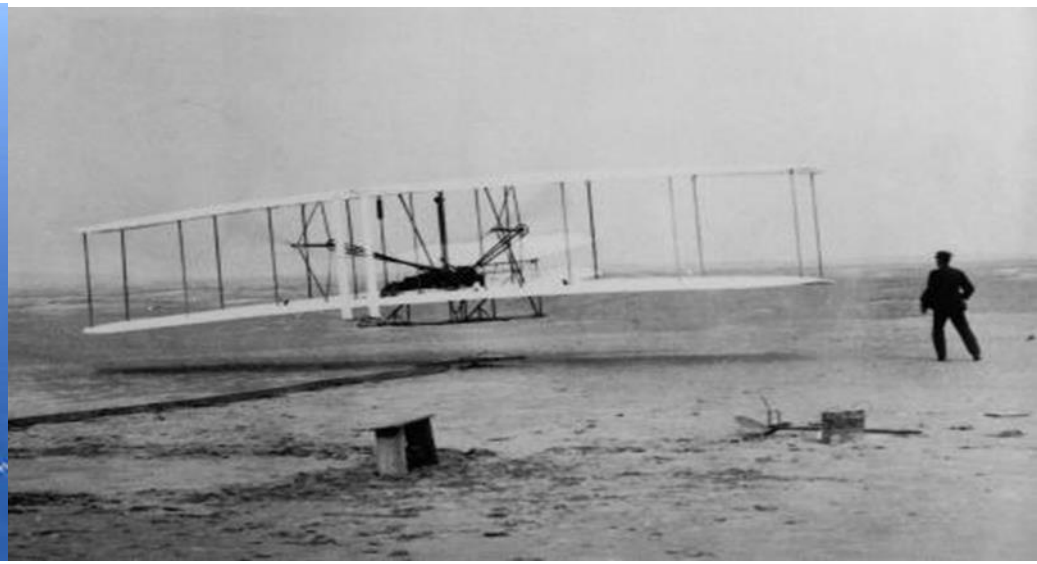
For all but small group of chemists developing it, InChI is not something anyone should want to know about.

All you want to do is use InChI to find information on the web.

InChI is helping scientists to do better work and find/link to the latest information.

Without InChI, finding something on the Internet is like trying to find the bathroom in a house with 1,000,000 unmarked doors

The Internet is like drinking from a fire hydrant; InChI will cut it to a faucet drip.



InChI Technology

Other Technology

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

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

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.

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

multiple applications
multiple repositories
multiple interfaces and protocols

For a chemical structure a (i.e., one) standard will move integration forward.

Missed information is a waster of time. Missed and/or out of date safety, health, and environmental data you need can be lethal.

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

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

Connection Tables/Notations
- MolFiles, SDF, 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#, …

Too many “standards” actually slow things down and make getting to the information you want and need take a lot longer time and effort than it would take with InChI

E Pluribus Unum
Out of many, One

InChI

172 SMILES representations

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	

“Public health laboratories analyze disease agents and other health threats to deliver the answers needed to mount an effective public health response.”

(APHL home page)

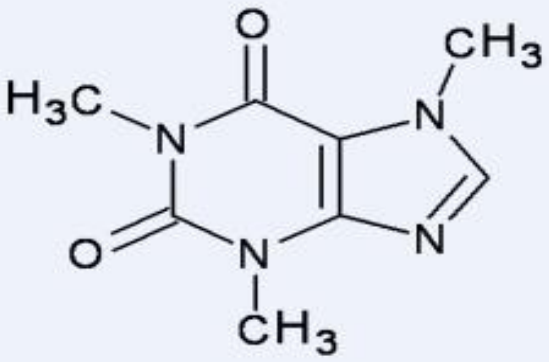
Elk River Chemical Spill

Please note that in some social media and early news reports, the chemical was MISIDENTIFIED as Methylcyclohexanol (CASRN: 25639-42-3). This is NOT the correct chemical.

Forwarded from the National Library of Medicine (NLM) Disaster Information Outreach by Librarians list serv

Multiple names for 4-Methylcyclohexanemethanol

- Synonyms: (CID 118193) Total: 30
- 4-Methylcyclohexanemethanol
- (4-Methylcyclohexyl)methanol
- 4-Methyl-1-cyclohexanemethanol
- Cyclohexanemethanol, 4-methyl-
- 1-(Hydroxymethyl)-4-methylcyclohexane
- 34885-03-5
- Cyclohexanemethanol, 4-methyl-, trans-
- ACMC-20amee
- AC1L3MEZ
- SureCN155951
- AC1Q7C7Z
- SureCN3485266
- SureCN8216437
- DSSTox_CID_21813
- DSSTox_RID_79850
- DSSTox_GSID_41813
- cis-4-Methylcyclohexanemethanol
- CTK1B4020
- trans-4-Methylcyclohexanemethanol
- Tox21_301528
- AR-1G3608
- AKOS009158915
- Cyclohexanemethanol, 4-methyl-, cis-
- NCGC00255649-01
- AI3-28423
- CAS-34885-03-5
- FT-0692545
- M1412
- M112072
- 3937-49-3
- OR **InChI**=1S/C8H16O/c1-7-2-4-8(6-9)5-3-7/h7-9H,2-6H2,1H3 **InChIKey:** OSINZLLLLCUKJH-UHFFFAOYSA-N
- https://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?q=nama&cid=118193



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

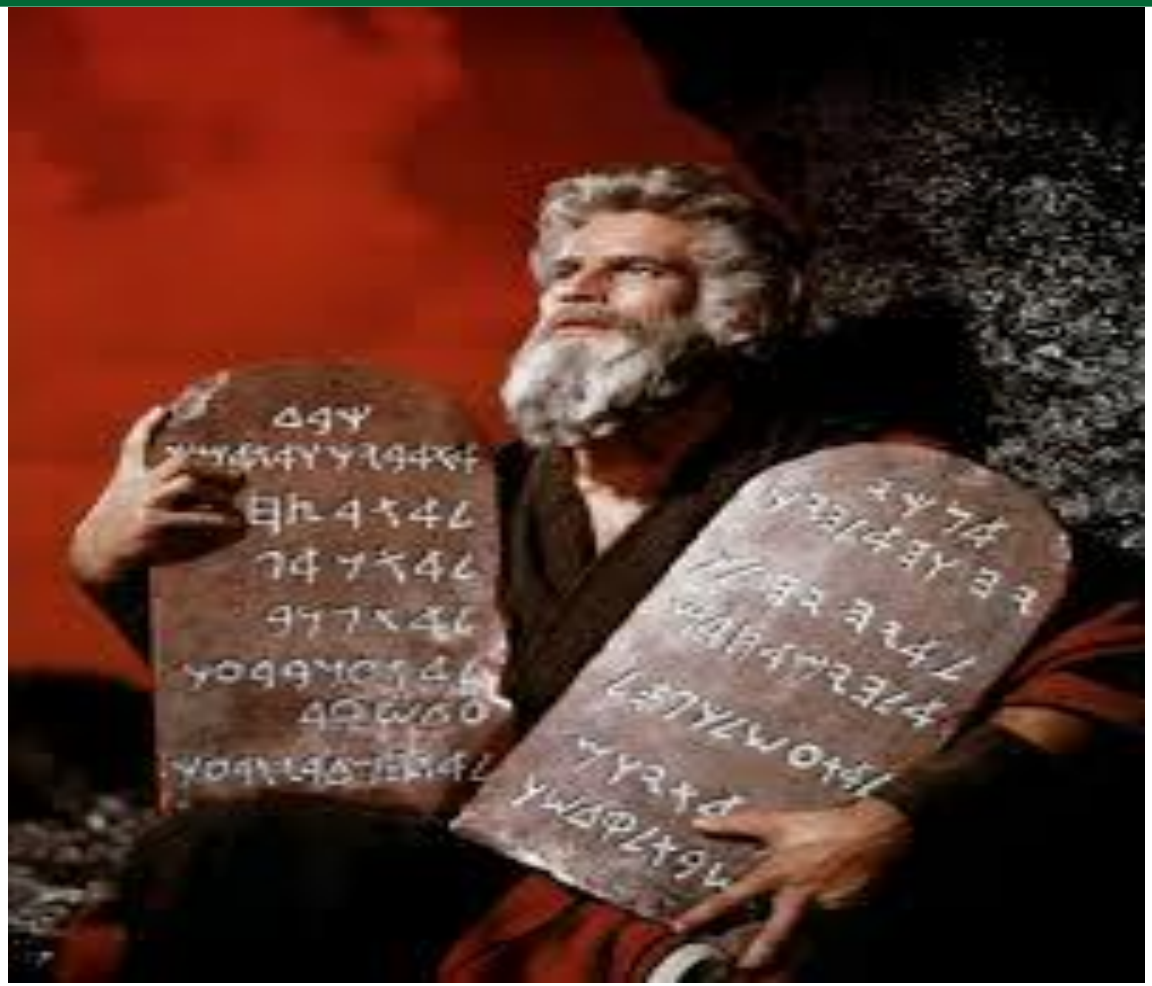
First block (14 letters)

Encodes molecular skeleton (connectivity)

Second block (8 letters)

Encodes stereochemistry and isotopes

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



What about funding ?

Don’t give up - Moses was once a basket case

Limitations of InChI

InChI does not handle undefined structures –

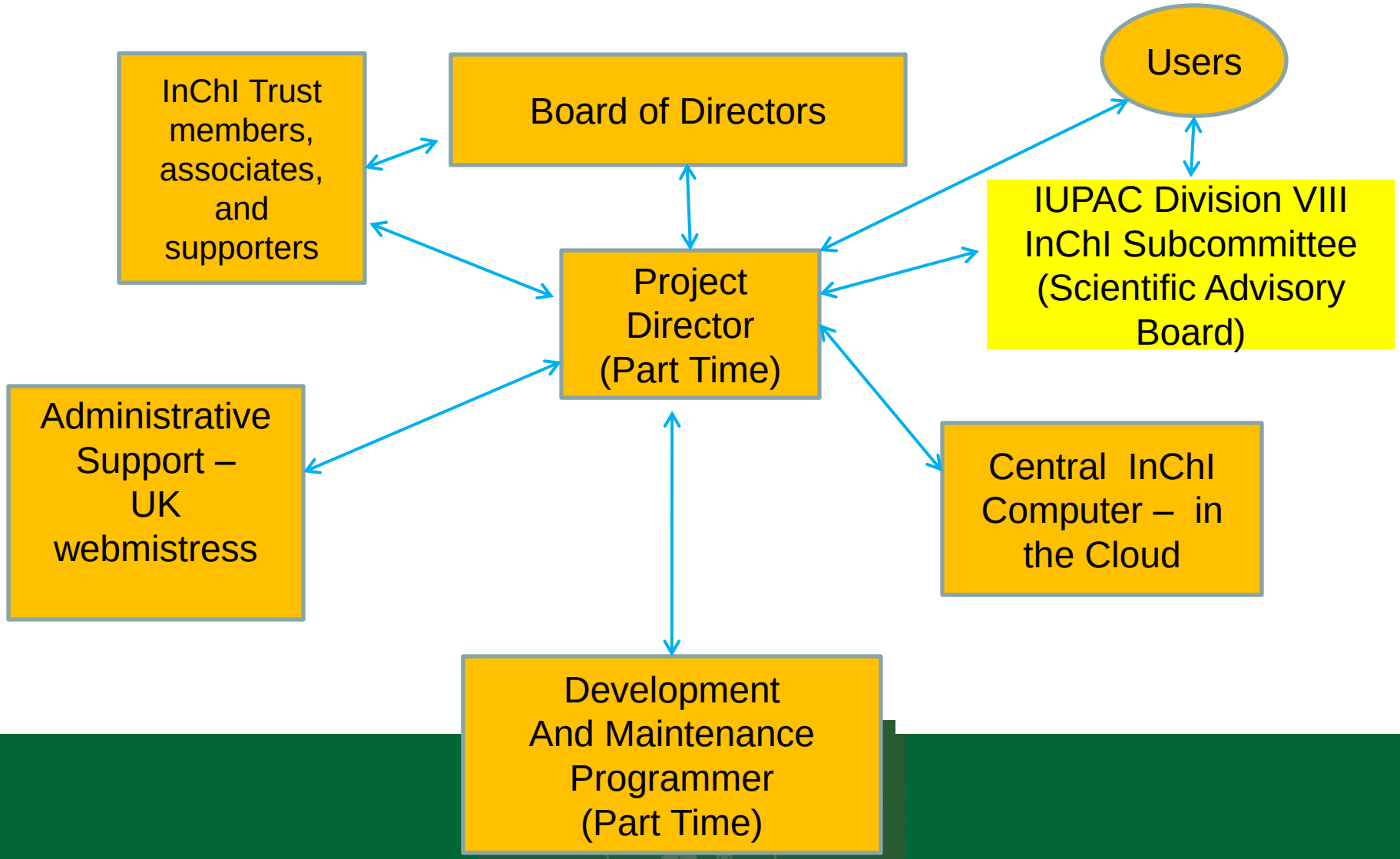
regular gas/summer or winter
balsamic vinegar
vegetable oil
partially hydrogenated oil

The InChI Trust

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 - >55.

The InChI Trust would like APHL members to join as (moral/non-paying) supporters.

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.

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

2. InChI is a public domain algorithm that anyone, anywhere can freely use. And they sure use it!

Success is uncoerced adoption

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

Current IUPAC Working Groups & Projects

In Progress:
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 when funded)
New API

Started/To be started in 2013/2014:
Electronic/Excited States
QR codes for InChI
InChI teaching/educational materials
Large Molecules/Biopolymers/Macromolecules
Material Science (MGI – Materials Genome Initiative)
Inorganics
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

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, 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.

InChI TRUST

I'd prefer not to be known as the late Steve Heller, so I will stop now.