

# How the NIH/EPA/NIST Mass Spec Database Created InChI

Stephen Heller

The main web sites for the IUPAC InChI project are:

<http://www.iupac.org/inchi>

and

<http://www.inchi-trust.org>

4/24/2018

Slides are available at <http://www.hellers.com/steve/NIST-4-18.pdf>

**This is a green talk –**

**These slides were made from  
100% recycled electrons**

**It would nice if one could look back now and write about how the need for an EI mass spectral database was well documented, how a careful study was done on what needed to be accomplished, and the well-implemented plan that followed. Certainly the results and success of the library and the many problems and chemical structures of unknowns it has solved could be logically explained in this way. Alas, the true history is not so clean and neat, however it is a nice story about many nice, friendly and highly competent people in different organizations who worked together to create and produce a world-class product still used today in ten's of thousands of analytical chemistry labs around the world.**

**The mass spectrometry database story starts back in the late 1960's. At the National Institutes of Health (NIH), the director of the Division of Computer Research and Technology (DCRT), Arnold (Scotty) Pratt, was looking for projects to show the scientific staff at NIH that computers could be more than calculators. (At the time there was good reason for this. An NIH Nobel Laureate once commented to me that if he had sufficient funds to purchase a computer to model protein structures he would rather use the money instead to purchase additional laboratory reagents.)**

**Henry (Hank) Fales, an eminent organic chemist and mass spectroscopist, and someone always willing to try something new agreed to test any computer based analysis tool that was developed. He asked one of his lab chemists, George W. A. (Bill) Milne, to work in testing the system.**

**The first task I had was to find a suitable digital data collection. Fales convinced Klaus Biemann at MIT to allow NIH to use his database on magnetic tape of some 8000 electron ionization (EI) mass spectra. Biemann and his students (Harry Hertz who later came to NIST and is now retired), and Ron Hites had been developing batch search software and was delighted to hear NIH was looking into interactive data searching**

**In 1970-1971 in a period of about a year, assisted by a very talented DCRT programmer, Richard Feldmann, I wrote a time-sharing program, in the FORTRAN programming language, for the DCRT PDP-10 timesharing computer, to search the database provided by Biemann.**

**What happened next is typical of Hank Fales. When Fales traveled and gave lectures, as well as when chemists from throughout the world came to visit his lab, he would add a little about the Mass Spectral Search System (MSSS) to his lecture or lab tour and show off the system. One day I received a telephone call from Fales who told me that a few (more like dozens and dozens) of Fales' friends really liked the MSSS he talked about and showed them. Fales and his 'few' friends wanted me to find some way to let them use the MSSS. Time sharing was just in its infancy then, with no internet and no large computer networks, no cloud - just 110 baud modems. The MSSS was opened to the public as an experimental project.**

Since it was not the mission of the NIH to provide service to the many people dialing into the DCRT PDP-10, a collaboration was established with the Mass Spectrometry Data Centre (MSDC) in Aldermaston England. The MSDC then converted the system to a commercial (General Electric) timesharing system and soon hundreds of chemists throughout the world were using the library and search system. In 1974 there were 60 accounts with an average of 2 new users each week, by 1976 there were 175 labs using the system. The average number of searches per day grew from 10 searches/day in 1973 to 100 searches/day in 1975. The MSSS later moved to the ADP/Cyphernet system, and then, a few years later, became part of the larger NIH/EPA Chemical Information Systems (CIS), which is no longer available today.



**In 1973 I moved to EPA and continued to work with EPA labs that had been given a Congressional mandate for environmental analysis. Working primarily with two EPA scientists, Bill Budde and John McGuire, the MSSS, and most importantly, the mass spectral library became the mainstay of EPA's environmental analysis program. Analytical methods were developed which required library searching of major unknown compounds against a comprehensive library. When good matches were found they were reported as “tentatively identified compounds”. With EPA having a very practical need for mass spectral search and analysis, considerable funds were provided to increase the size and quality of the library. The initial library and software work was contracted and some quality control work was started, including adding CAS Registry Numbers, which EPA obtained under a contract from CAS.**

By 1975 there were some 30,000 spectra in the library. While EPA was the driving force behind the rapid expansion and use of the library, it did not have the mission to provide data to the public. In 1978, the National Bureau of Standards – NBS (now known as the National Institute of Standards and Technology – NIST) agreed to print a five volume collection of some 25,500 different mass spectra along with their chemical name, synonyms, and chemical structure. The entire print run of some 1500 copies of the “red books” were sold out quickly, though the books were never reprinted. Through a unique congressional authorization to copyright data collections and recover costs, the Standard Reference Data Act , NBS/NIST did have the mission of providing reference data to the public. NBS/NIST agreed in 1980 to assume responsibility for the dissemination of the library. Under the leadership of David Lide and Lew Gevantman of the Office of Standard Reference Data (OSRD), the library was expanded, and distributed to instrument manufacturers on 9-track magnetic tape.

**Jumping forward a decade or two, at NIST, owing to CAS refusing to continue to contract for supplying CAS registry Numbers in the late 1990's, the chemical structure of a compound replaced the CAS registry number as the unique "key" for identifying the compound.**

**For this purpose, chemical structure processing software was developed at NIST to enable compound "registration" (finding spectra for the same compound). This enabled the immediate inclusion of well over 10,000 compounds held in the archive for which the CAS registry numbers were unavailable. However this was not an ideal situation.,**

**Thus, after over 20 years of working with the mass spec database, the refusal of CAS to supply CAS Registry numbers, coupled with my retirement from the US Government Steve Stein and I came up with an outline of a plan to develop a more rational chemical registration system for the database. At the same time International Union of Pure and Applied Chemistry (IUPAC), the international standards organization for chemistry finally realized that chemical names was not the future for representing chemical structures. (The IUPAC stone age of printed information was replaced with the computer age.) The first draft of the plan was prepared in November 1999, a few months prior to the IUPAC meeting**

Date: Mon, 15 Nov 1999 18:48:30 -0500 (EST)  
From: Stephen R. Heller<srheller@cliff.nal.usda.gov>  
To: stein <sstein@enh.nist.gov>  
Subject: Re: A strawman proposal

**Steve-**

**First rough draft. Let's talk tomorrow about it.**

**Steve**

-----  
**11/15/99**

### **An IUPAC Chemical Registry System**

**In response to the upcoming March 2000 IUPAC meeting -  
Representations of Molecular Structure: Nomenclature and its Alternatives  
- I would like to propose the creation of an IUPAC public domain chemical  
registry system.**

...

**Over the past decade with the ever-increasing reliance on computer processing by chemists, it became evident to Becker and McNaught at IUPAC that this organization should find better ways of handling nomenclature was done in the past. In particular it was felt by many that while IUPAC had stressed conventional chemical names/nomenclature in the 20th century, continued progress into the 21st century required new, computer-driven approaches to the problem of chemical identification**

**Under the leadership of two senior IUPAC officials, Ted Becker (NIH) and Alan McNaught (RSC), IUPAC convened a meeting in March 2000 in Washington DC to look into the matter of chemical structure representation. The IUPAC Strategy Roundtable meeting was called “Representations of Molecular Structure: Nomenclature and its Alternatives”. It brought together 41 participants from 10 countries including experts in organic, inorganic, biochemical, and macromolecular nomenclature; users of nomenclature in academia, industry, the patent, international trade, health and safety communities; journal editors and publishers; database providers; and software vendors.**

**At the meeting in March 2000 Steve the elder and Steve the younger presented a proposal to IUPAC, which extended one developed in the fall of 1999. The initial proposal from November 1999 was widely circulated with the chemical information and chemical structure representation community via e-mail. The proposal presented at the March 2000 meeting was incorporated considerable improvements from this feedback from chemists in the USA, Europe, and Asia.**

**At the end of the March 2000 meeting Bill Town proposed that the new program be called IUPAC Chemical Identifier Project (ICHIP). NIST offered to provide staff to create and program this chemical identifier.**

**And the rest is history.**



**The aim of the IUPAC Chemical Identifier Project (IChIP) is to establish a unique label, the IUPAC Chemical Identifier (IChI), which would be a non-proprietary identifier for chemical substances that could be used in printed and electronic data sources thus enabling easier linking of diverse data compilations and unambiguous identification of chemical substances.**

**IChI is not a registry system. It does not depend on the existence of a database of unique substance records to establish the next available sequence number for any new chemical substance being assigned an IChI. It will be based on a set of IUPAC structure conventions, and rules for normalization and canonicalization of an input structure representation to establish the unique label. It will thus enable an automatic conversion of a graphical representation of a chemical substance into the unique IChI label which can be created independently of any organization anywhere in the world and which could be built into any chemical structure drawing program and created from any existing collection of chemical structures.**

**The initial name for the structure representation was the IUPAC Chemical Identifier (ICHI). During the initial development by NIST, NIST staff suggested the name be changed to the IUPAC-NIST Chemical Identifier – INChI. By the time the NIST lawyers did not having the NIST name associated with a standard, INChI had become the accepted acronym so the name was changed to the InChI – The International Chemical Identifier.**

# InChI Project Goal

**To link everything (data and information) about a chemical from many sources with the purpose of creating new information and perhaps knowledge.**

**But before you can share and use data and information do you need to find.**

**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. The InChI standard enables competition. The InChI standard will work only with community involvement and engagement.**

# 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](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**

# 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>

# How did we get here?

**1999:** Steve Heller initiated a proposal at NIST for a public domain structure representation standard for the NIST databases

**2000:** Decided that InChI would be an IUPAC initiative

**2001:** The IUPAC Chemical Identifier project began

**2005:** Version 1 was launched

**2009:** Standard versions of InChI and the InChIKey were released, which took the original algorithm with its many variable parameters and fixed them so that interoperability between databases and resources with InChIs could be achieved

**2009:** The UK based InChI Trust was formed

**2011:** Version 1.04 released

**2017:** Version 1.05 of the InChI, along with version 1.00 of Reaction InChI (RInChI)



# Four Requirements for a Computer Representation Standard

**Need**  
**Definition/Specification**  
**Timing/Infrastructure**  
**Acceptance/Use**

**Actually there is a 5<sup>th</sup> requirement for a standard. First rate staff to create, define, program, deliver the standard. Without Steve Stein and Dmitrii Tchekovski there would be no InChI algorithm. And without Alan McNaught and Ted Becker, getting the standard through the IUPAC approval process, InChI would only be used at NIST for the mass spec and WebBook databases.**

**It has taken a dedicated, highly competent team of chemists with a vision of the future to get this all to work. And a team with each player fitting perfectly together with an unbelievable lack of personality clashes. It was and still is a perfect “good storm” of people and needs. It also took the Internet to be there for the many databases and resources that are now linked together with InChI.**

**Being at the right place, at the right time, and having the right very smart people was and is the unique key to the success of the project.**

**“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**

[www.pistoiaalliance.org/](http://www.pistoiaalliance.org/)

**We are a not-for-profit alliance of life science companies, vendors, publishers, and academics that work together to lower barriers to innovation in R&D.**

**Allotrope Foundation: Data Standard**

<https://www.allotrope.org/>

**Revolutionizing the way we acquire, share and gain insights from scientific data, through a community and the framework for standardization and linked data.**

**CODATA, The Committee on Data for Science and Technology**  
[www.codata.org/](http://www.codata.org/)

**CODATA works also to advance the interoperability and the usability of such data: research data should be intelligently open or FAIR. By promoting the policy, technological and cultural changes that are essential to make research data more widely available and more usable, CODATA helps advance ICSU's mission.**

**RDA | Research Data Sharing without barriers**  
<https://www.rd-alliance.org/>

**The Research Data Alliance (RDA) is an international organization focused on the development of infrastructure and community activities aimed to reduce barriers to data sharing and exchange, and promote the acceleration of data driven innovation worldwide.**

**IUPAC CPCDS - Committee on Publications and Cheminformatics Data Standards**  
To advise IUPAC on all aspects of the design and implementation of publications and data-sharing, including computerized databases of all sorts, and to promote the compatibility of the electronic transmission, storage, and management of digital content through the development of standards for the creation of a consistent and interoperable global framework for human and machine-readable chemical information.

## **FAIR**

**<https://www.force11.org/group/fairgroup/fairprinciples>**

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Here, we describe FAIR - a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable. The FAIR principles have now been published.

# 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)**

# 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 active**

**RSC/ChemSpider – 59 million**

**Elsevier/Reaxys – 30 million**

**IUPAC – 0 million**

**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 and Germany, 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.



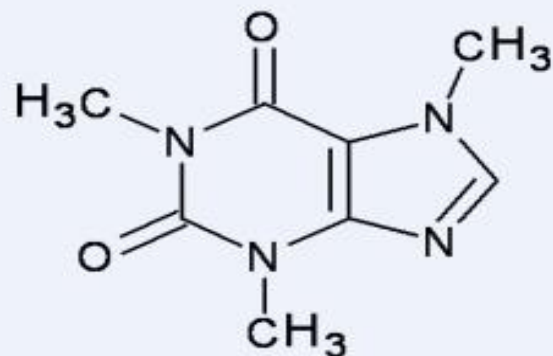
# InChI layered structure design

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 (*sp*<sup>3</sup>)
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)

InChIKey=**RYYVLZVUVIJVGH-UHFFFAOYSA-N**

character indicating the number of protons  
(‘N’ means neutral)

flag character for InChI version:  
‘A’ for version 1

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

First block (14 letters)

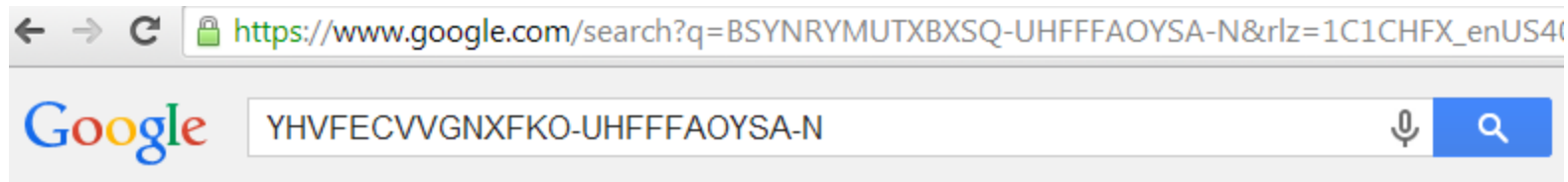
Encodes molecular skeleton  
(connectivity)

Second block (8 letters)

Encodes stereochemistry and isotopes

# Search Engines can use InChIKey

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



Web Maps Shopping Images News More Search tools

About 100 results (0.32 seconds)

## ChemIDplus - 4210-97-3 - YHVFECVVGXFKO ...

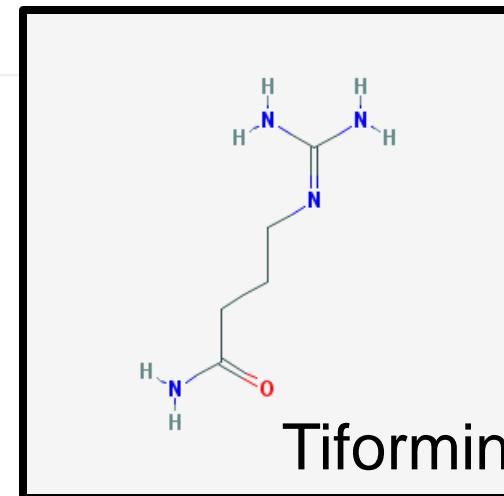
chem.sis.nlm.nih.gov/.../4210... United States National Library of Medicine  
4210-97-3 - YHVFECVVGXFKO-UHFFFAOYSA-N - Tiformin [INN:BAN] - Similar structures search, synonyms, formulas, resource links, and other chemical ...

## tiformin - PubChem

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



**And now to final question of  
how we have been able to  
keep InChI going after NIST  
finished programming the  
needs it had for NIST  
databases.**



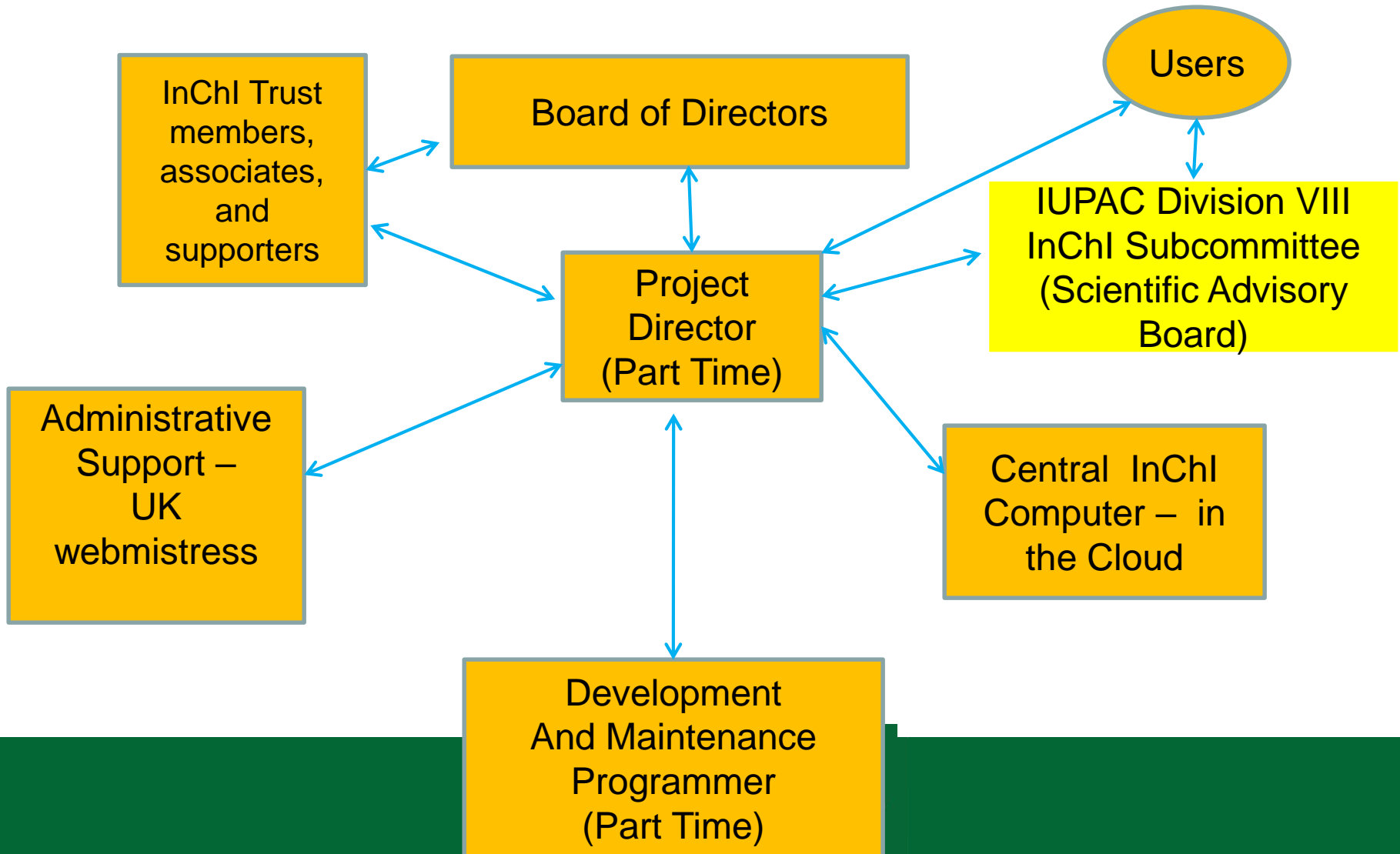
**What about funding ?**

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

# 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 be 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







# 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 with polymers released – 2017  
Markush (contract to be signed when funded)  
RInChI – InChI for Reactions released i- 2017

## 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

## Enhancements in InChI version 1.05

- support for chemical elements numbers 113-118;
- experimental support of InChI/InChIKey for simple regular single-strand polymers;
- experimental support of large molecules containing up to 32767 atoms;
- ability to read necessary for large molecule input files in Molfile V3000 format;
- provisional support for extended features of Molfile V3000;
- significant updating of the InChI API Library; in particular, a novel API procedure for direct conversion of Molfile input to InChI and a new set of API procedures for both low and high-level operations (InChI extensible interface, IXA);
- significant modification of the source code in order to ensure multi-thread execution safety of the InChI Library;
- several minor bug fixes/changes;
- several convenience options added to the inchi-1 executable.

# Mixtures - MInChI

*Is the mixtures InChI to be a representation or an identifier?*

Mixture InChI may be described as a collection of parent compounds in a substance

Implementations may consider further information layers (provenance, sample)

Handle all concentrations as ranges –uncertainties, isomers

Explicit inclusion of isomers is preferred (to a reasonable limit)

Enhanced stereochemistry is very important

Tautomers are also important, so is using standard InChIs

Mixtures InChI may offer accommodations for convertible isomers

Use /n layer to notate repeated InChI units (e.g., tautomers) no structures,

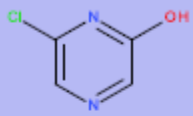
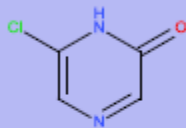
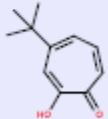
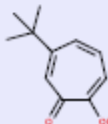
AND/OR logic within isomer groups

Other needs: salts, solvates, states, forms, conditions (?)process (?)

Workflow for file input and conversion

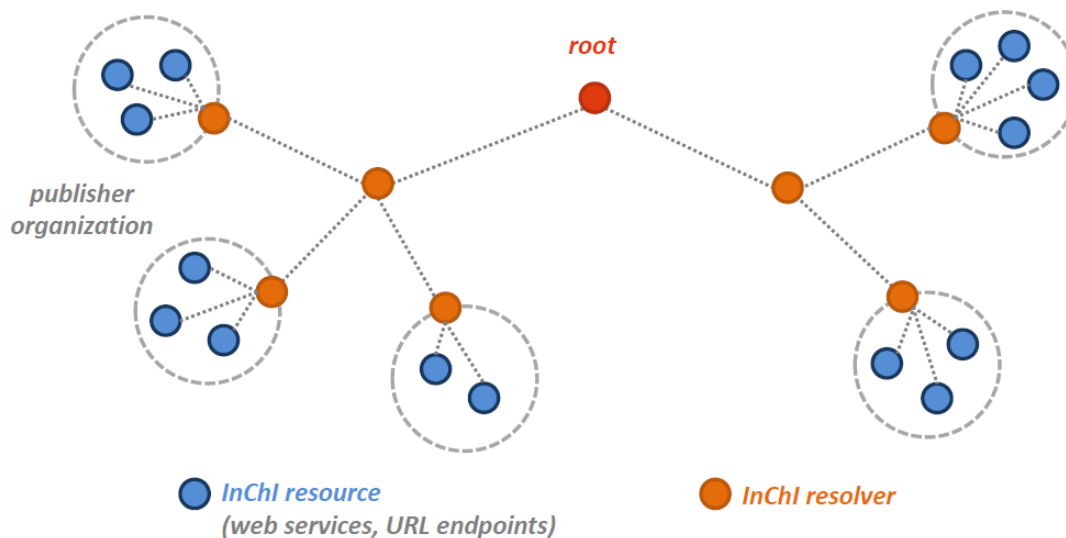
# Mixtures - MInChI

Tautomers

Example	InChI	MInChI
	InChI=1S/C4H3ClN2O/c5-3-1-6-2-4(8)7-3/h1-2H,(H,7,8)	MInChI=0.00.1S/C4H3ClN2O/c5-3-1-6-2-4(8)7-3/h1-2H,(H,7,8)/n{1&1}
	InChI=1S/C4H3ClN2O/c5-3-1-6-2-4(8)7-3/h1-2H,(H,7,8)	
	InChI=1S/C11H14O2/c1-11(2,3)8-5-4-6-9(12)10(13)7-8/h4-7H,1-3H3,(H,12,13)	MInChI=0.001S/C11H14O2/c1-11(2,3)8-5-4-6-9(12)10(13)7-8/h4-7H,1-3H3,(H,12,13) /n{1&1}
	InChI=1S/C11H14O2/c1-11(2,3)8-5-4-6-9(12)10(13)7-8/h4-7H,1-3H3,(H,12,13)	

# Resolver

GOAL: making InChI resources on the Web findable, linkable and browsable by a common, unified protocol



<http://www.inchi-resolver.org/>

# QR Codes

## InChI QR Code: Standard Form



IUPAC-InChI

[database]/[type]/[identifier]

[Source URL]/InChIKey/[Key]

1. Institution incorporates into local app
2. Vendor generated and linked to their database
3. Public version (QRInChI.org for re-direction e.g. to info.identifiers.org)





# 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.**