

Developing Databases and Standards in Chemistry

Stephen Heller

NIST

&

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>

8/23/2016

Slides are available at <http://www.hellers.com/steve/ACS-PHL-8-16.pdf>



This is a green talk –

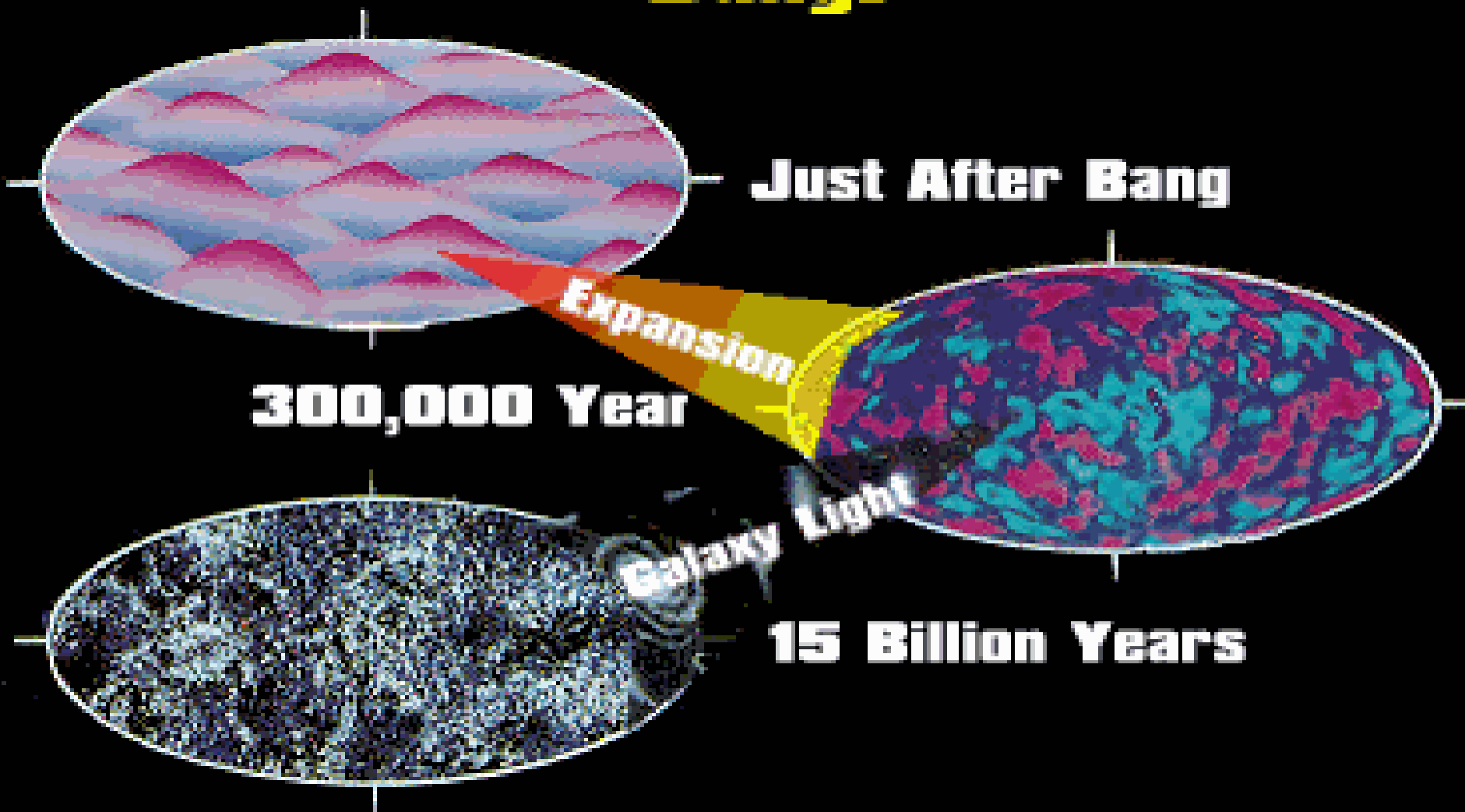
**These slides were made from
100% recycled electrons**

3/25/16 Email invitation from Evan Bolton to speak at this Symposium

“In your talk, you would address the
beginning ...

Your talk will be first and helps to set
the tone for the whole symposium”

Bang!



**Considering Evan gave me
only 30 minutes I will have to
skip the first few billion years
of databases and standards.**



Data, Databases, and connecting information – My 5+ decades of experiences

My secret to how I got here:

Luck, luck, luck – and it takes a village

(If the problem was just technology someone would have solved it already. The real short and long term problem is always non-technical – cultural and political.)

Dislike chemistry lab work

Be at the right place at the right time with the right people

Work with supportive people – “work with your friends and forget your enemies” (Morris Yaguda, EPA)

Plan for the exit day – plan for who will take it next

PS. Looking back it may seem like it was planned and there was a vision – but that was not the case – I just fell into it..

Be at the right place at the right time and work with supportive people

Be an undergraduate at SUNY Stony Brook when there was no graduate program/students

Be at NIH and collaborate with forward, positive thinking Hank Fales & Bill Milne on mass spectral data.

Be at EPA when they just started using Mass Spec to identify pollutants with a forward thinking non-scientist/great manager Morris Yaguda

Be at NIST with Steve Stein when CAS stopped providing Registry Numbers to the NIST Mass Spec database

Be retiring just when Ted Becker and Alan McNaught thought IUPAC needed to move into the 21st century of chemical structure representation.

Join the NCBI PubChem Advisory Board with Steve Bryant and Evan Bolton

**Before I go to the next slide I
want to prepare you all for
the fact that it will be a
chemistry slide.**

The Reaction of Trialkyl Phosphites with Aliphatic Aldehydes. P^{31} and H^1 Nuclear Magnetic Resonance Spectra of Tetraoxyalkyl Phosphoranes^{1,2}

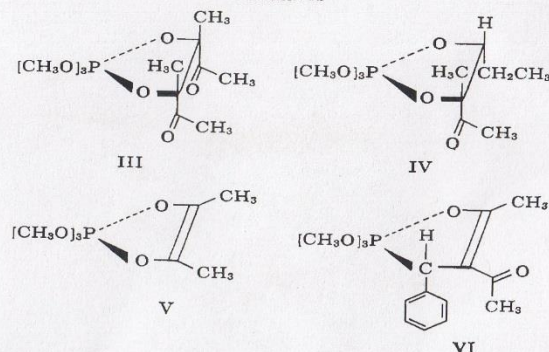
Sir:

We wish to report the isolation and characterization of a tetraoxyalkyl phosphorane(I) from the reaction of 3 moles of anhydrous propionaldehyde with 1 mole of trimethyl phosphite at 20°. The 2:1 adduct

(1) (a) F. Ramirez, A. V. Patwardhan, N. B. Desai, N. Ramanathan, and C. V. Greco, *J. Am. Chem. Soc.*, **85**, 3056 (1963); (b) F. Ramirez, N. Ramanathan, and N. B. Desai, *ibid.*, **85**, 3465 (1963); (c) F. Ramirez and N. B. Desai, *ibid.*, **85**, 3252 (1963); **82**, 2652 (1960).

(2) Acknowledgment is made to the Cancer Institute of the National Institutes of Health (CY-04769) and to the National Science Foundation (G19509) for support of this research.

TABLE I
CHEMICAL SHIFTS IN THE P^{31} N.M.R. SPECTRA OF OXYPHOSPHORANES^a



Cyclic Oxyphosphorane		δP^{31}
Pentaoxy 1,3-dioxaphospholane	III ^b	+54.83
Pentaoxy 1,3-dioxaphospholene	IV ^c	+51.27
Tetraoxyalkyl 1,4-Dioxaphospholane	V ^{d-f}	+48.92
Tetraoxyalkyl Δ^4 -Oxaphospholene	I ^e	+34.16
	VI ^{h-j}	+27.89

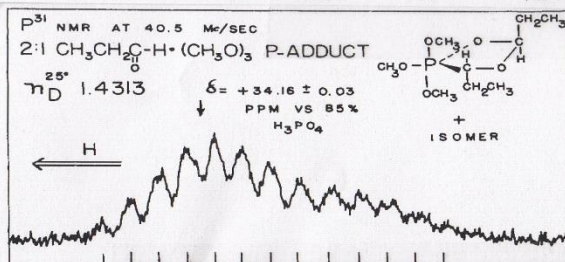


Figure 1

Journal of the
American Chemical Society
86 (3), 514-516, 1964

Acknowledgment.—We are grateful to Prof. P. C. Lauterbur of this department for advice on P^{31} n.m.r. spectroscopy and to Dr. E. M. Banas (American Oil Co.) and Prof. E. Eliel (University of Notre Dame) for some of the earlier H n.m.r. spectra.

(7) (a) V. A. Kukhtin and K. M. Kirillova, *J. Gen. Chem. USSR*, **31**, 2078 (1961); (b) *Zh. Obshch. Khim.*, **31**, 2226 (1961).

(8) A. Arbuzov and V. M. Zoroastrova, *Izv. Akad. Nauk SSSR Otd. Khim. Nauk*, 1030 (1960).

(9) A. C. Poshkus and J. E. Herweh, Abstracts, Division of Organic Chemistry 141st National Meeting of the American Chemical Society, Washington, D. C., March, 1962, p. 17-O.

(10) V. Mark, *J. Am. Chem. Soc.*, **85**, 1884 (1963).

DEPARTMENT OF CHEMISTRY
STATE UNIVERSITY OF NEW YORK
STONY BROOK, NEW YORK

FAUSTO RAMIREZ
A. V. PATWARDHAN
STEPHEN R. HELLER

RECEIVED SEPTEMBER 26, 1963

My first 4 peer reviewed papers were in the prestigious ACS journal – JACS. Why? I was in the right place at the right time.

This made it easy to be thought of as an expert when there is no one else around in the field at the time.

PS. I have never been able to get another paper into JACS.

The NIH/EPA/NIST mass spectrometry database – 1970's – date

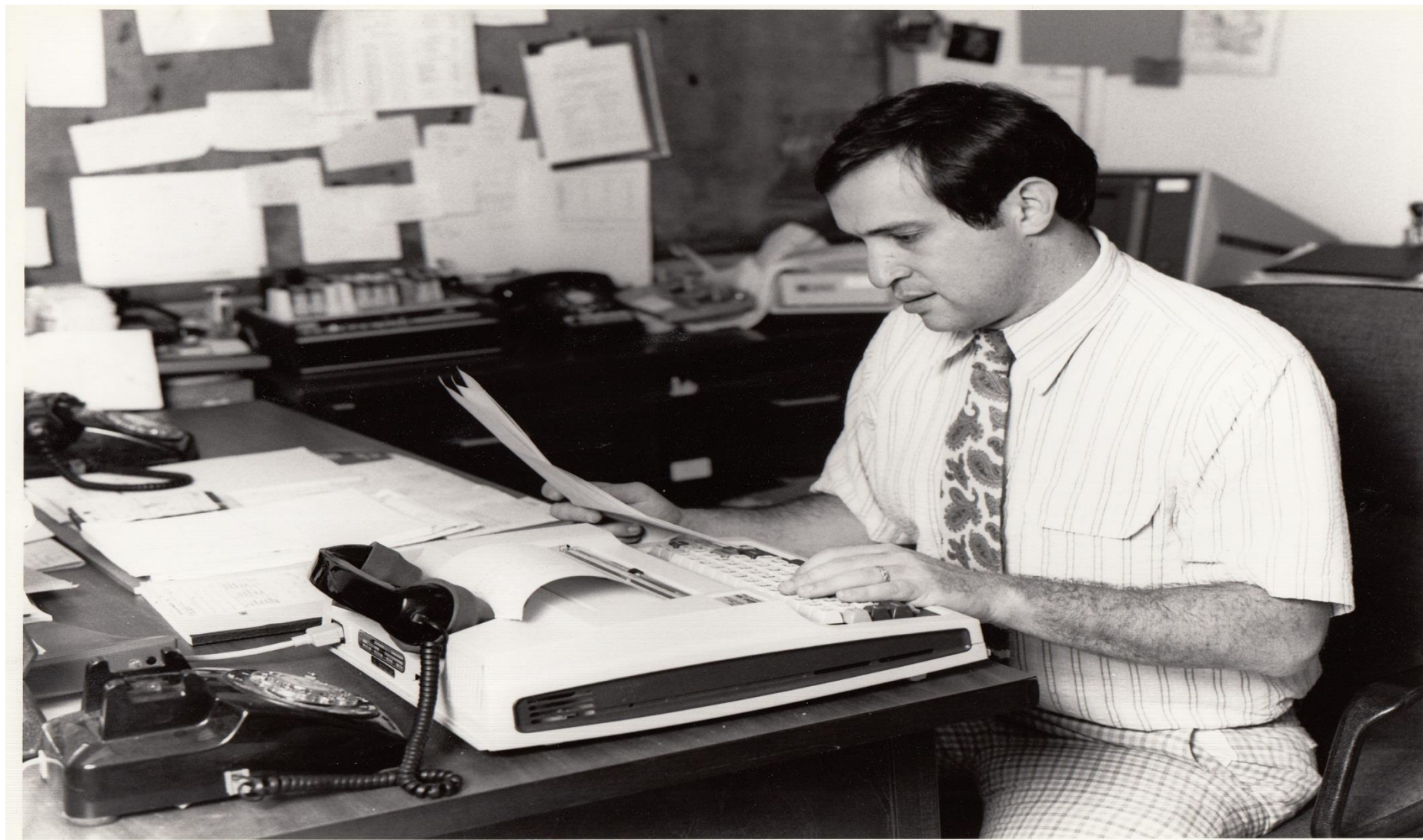
**Initial database came from MIT/Klaus Biemann, a colleague of Hank Fales
Modified search software came from Richard Feldmann at NIH/DCRT**

Without the NIH PDP-10 timesharing computer connected to a dial up phone line (the precursor to the cloud and the internet) no one outside NIH would have been able to see or use the database.

At NIH, the NLM (pre NCBI) along with the ACS objected to any non-NLM databases being developed and disseminated outside the NLM

Moving to EPA who needed a way to identify pollutants, was looking back, the perfect place to work. But EPA is run by lawyers, so a long term solution needed to be found.

Hence, the database was moved to NIST/SRD in 1980's while I went off to USDA/ARS.



The NIH/EPA/NIST mass spectrometry database – 1970's – date

NIST was the ideal home, except some at NIST thought the quality of data was not up to NIST standards. But since NIST had the unique authority in the US Government to copyright data it was literally a gold mine for NIST, which now collects a few million dollars a year in mass spec database royalties.

Summary:

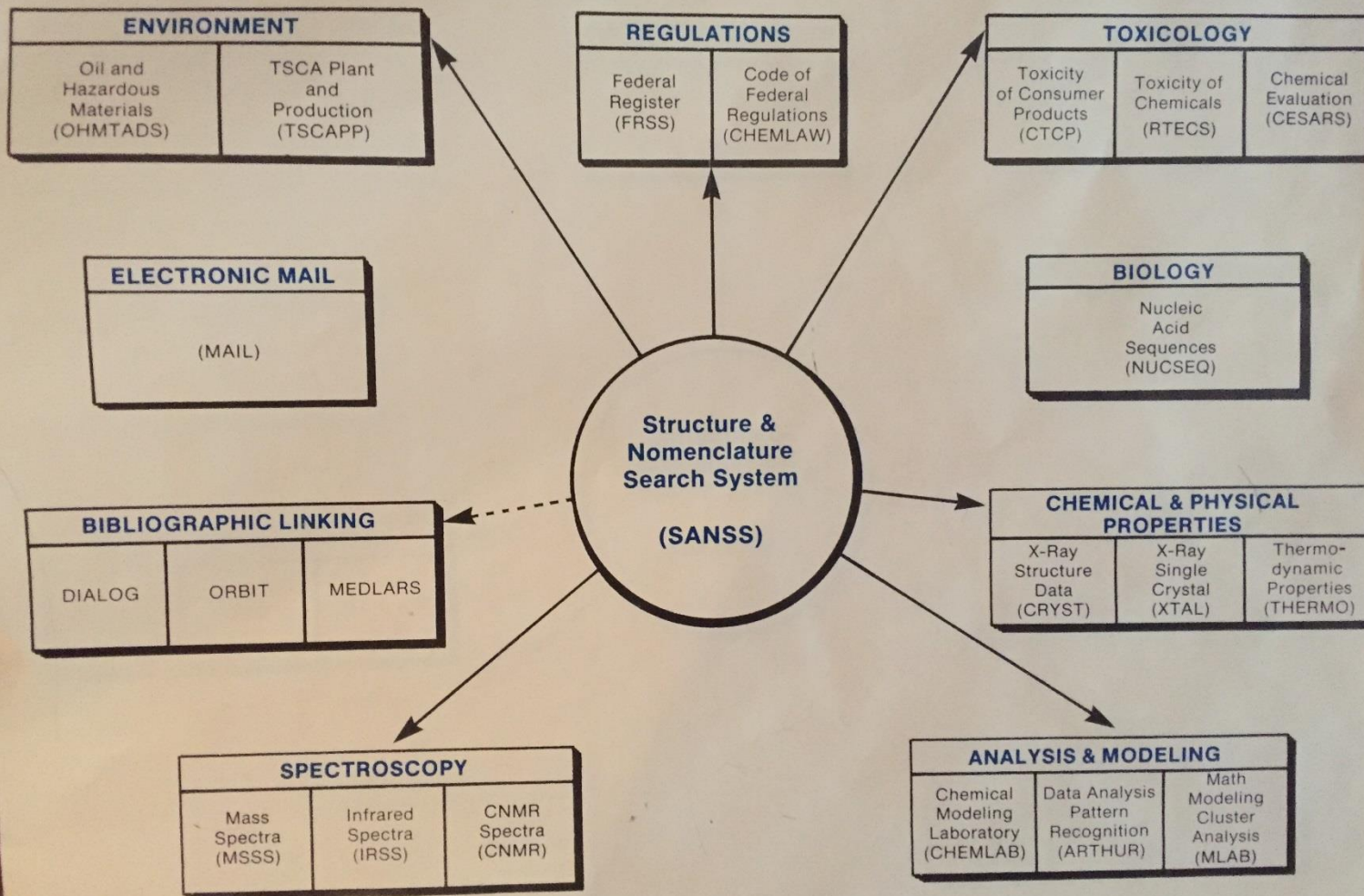
Find a home at NIST for the mass spec database was being at the right place at the right time, work with the supportive people and was a plan for my exit as to whom would take it to the next stage.

The NIH/EPA Chemical Information System – CIS

A collection of chemical structures with links to various databases supporting environmental and scientific needs. It also had a number of analysis and prediction programs All databases had CAS Registry Numbers as their link. The CIS worked for a number for years but never had the political backing at NIH or EPA and was not supported or encouraged by the “establishment” of the ACS and NLM/SIS. It died in the mid 1980’s. It was a bit ahead of its time.

CIS COMPONENTS

SPRING, 1983



the united states environmental protection agency

Awards to

Dr. Stephen R. Heller

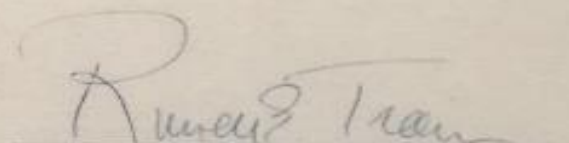
the

Gold Medal for Exceptional Service

For outstanding leadership in the design and implementation of the Chemical Information System (CIS) which coordinates and links all EPA chemical files with the vast chemical literature and numerous other government chemical files.



December 13, 1976


RUSSELL E. TRAIN, ADMINISTRATOR

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.**

...

InChI Project Goal

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

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 supported 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>

The InChI Team **(The right people)** (alphabetical order)

Stephen R. Heller

Alan McNaught

Igor Pletnev

Stephen E. Stein

Dmitrii Tchekhovskoi

Four Requirements for a Computer Representation Standard

Need
Definition/Specification
Timing/Infrastructure
Acceptance/Use

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,
Associate Director for Data Science, NIH**

Definition/Specification

A computer algorithm to insure consistency and reproducibility.

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*³)
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 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, without the lack of support from the ACS and CAS, 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 of 5-7 people at this ACS meeting talk.

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(C)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(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)c(=O)c2c1ncn2C
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(n(cn2)C)c(=O)n(C)c1=O)C
n1(C)cn2c1c(=O)n(C)c(=O)n2C
O=c1c2c(n(C)c(=O)n1C)ncn2C
n1(c2c(nc1)n(c(=O)n(c2=O)C)C)C
n1(C)c(=O)c2c(n(c1=O)C)ncn2C
n1(c2c(c(=O)n(C)c1=O)ncn2C)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)ncn2C
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)n(c2c1n(cn2)C)C)C
n1(c2c(c(=O)n(c1=O)C)ncn2C)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
n1(cnc2c1c(=O)n(C)c(=O)n2C)C
n1(C)cnc2c1c(=O)n(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
Cn1c2c(n(c(=O)n(c1=O)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(cn2)C)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)ncn2)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(c2c(n(C)c(=O)n(C)c2=O)nc1)C
O=c1n(c2c(n(C)cn2)c(=O)n1C)C
c1(=O)n(C)c(=O)n(C)c2c1n(C)cn2
Cn1c(=O)n(c2c(c1=O)n(cn2)C)C
n1(c2c(n(c(=O)n(C)c2=O)C)nc1)C
Cn1c2c(c(=O)n(c1=O)C)n(C)cn2
c1(=O)n(C)c2c(c(=O)n1C)n(cn2)C
O=c1n(C)c2c(c(=O)n1C)n(cn2)C
c1(=O)n(C)c(=O)n(c2c1n(C)cn2)C
Cn1c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(nc1)n(c(=O)n(C)c2=O)C)C
O=c1n(c(=O)n(c2c1n(cn2)C)C)C
c1(=O)n(C)c2c(c(=O)n1C)n(C)cn2
O=c1n(C)c(=O)n(C)c2c1n(cn2)C
c1(=O)n(C)c2c(c(=O)n1C)n(C)cn2
n1(C)c(=O)c2c(ncn2C)n(C)c1=O
Cn1c(=O)n(c2c(ncn2C)n(C)c1=O)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(cn2)C)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
Cn1(c(=O)n(C)c2c(c1=O)n(cn2)C)C
O=c1n(C)c(=O)n(c2c1n(cn2)C)C
O=c1n(c(=O)c2c(ncn2C)n(c1=O)C)C
c1(=O)c2c(ncn2C)n(C)c(=O)n1C
Cn1c2c(n(C)c(=O)n(c2=O)C)nc1
n1(C)c(=O)c2c(n(C)c1=O)ncn2C
n1(C)c(=O)n(C)c2c(c1=O)n(C)cn2
Cn1c2c(c(=O)n(C)c1=O)n(cn2)C
n1(C)c(=O)n(C)c2c(n(C)cn2)c1=O
n1(c2c(n(C)cn2)c(=O)n(C)c1=O)C
n1(C)c(=O)n(c(=O)c2c1ncn2C)C
c1(=O)n(c(=O)n(c2c1n(cn2)C)C)C
c1(=O)n(c(=O)n(c2c1n(C)cn2)C)C
n1(C)c2c(nc1)n(c(=O)n(C)c2=O)C
Cn1c(=O)n(C)c2c(c1=O)n(C)cn2
O=c1n(c2c(c(=O)n1C)C)ncn2C
n1(C)c2c(n(c(=O)n(c2=O)C)C)nc1
n1(C)c(=O)n(C)c2c(c1=O)n(cn2)C
n1(C)c2c(nc1)n(C)c(=O)n(C)c2=O
n1(C)c2c(n(cn2)C)c(=O)n(c1=O)C
n1(C)c(=O)n(c2c(c1=O)n(cn2)C)C
n1(C)c2c(c(=O)n(C)c1=O)n(cn2)C
n1(c(=O)n(c2c(c1=O)n(C)C)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(c1=O)C)n(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



InChI
(Trump's Lamborghini)

172 SMILES representations



E Pluribus Unum
Out of many, One

Whatever the controversies and different opinions, InChI has now been more widely adopted than SMILES. In addition, three US Government agencies - FDA, NIH, NIST - now have become paying members of the InChI Trust which would seem to indicate more official and institutional support leading to further widespread usage. And leading to a long term, personality independent, standard.

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 meet the everyday 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 – 137 million

NIH/NCI – 128 million

NIH/PubChem - 91 million (68 million online)

RSC/ChemSpider – 34 million

Elsevier/Reaxys – 30 million

IUPAC – 0 million

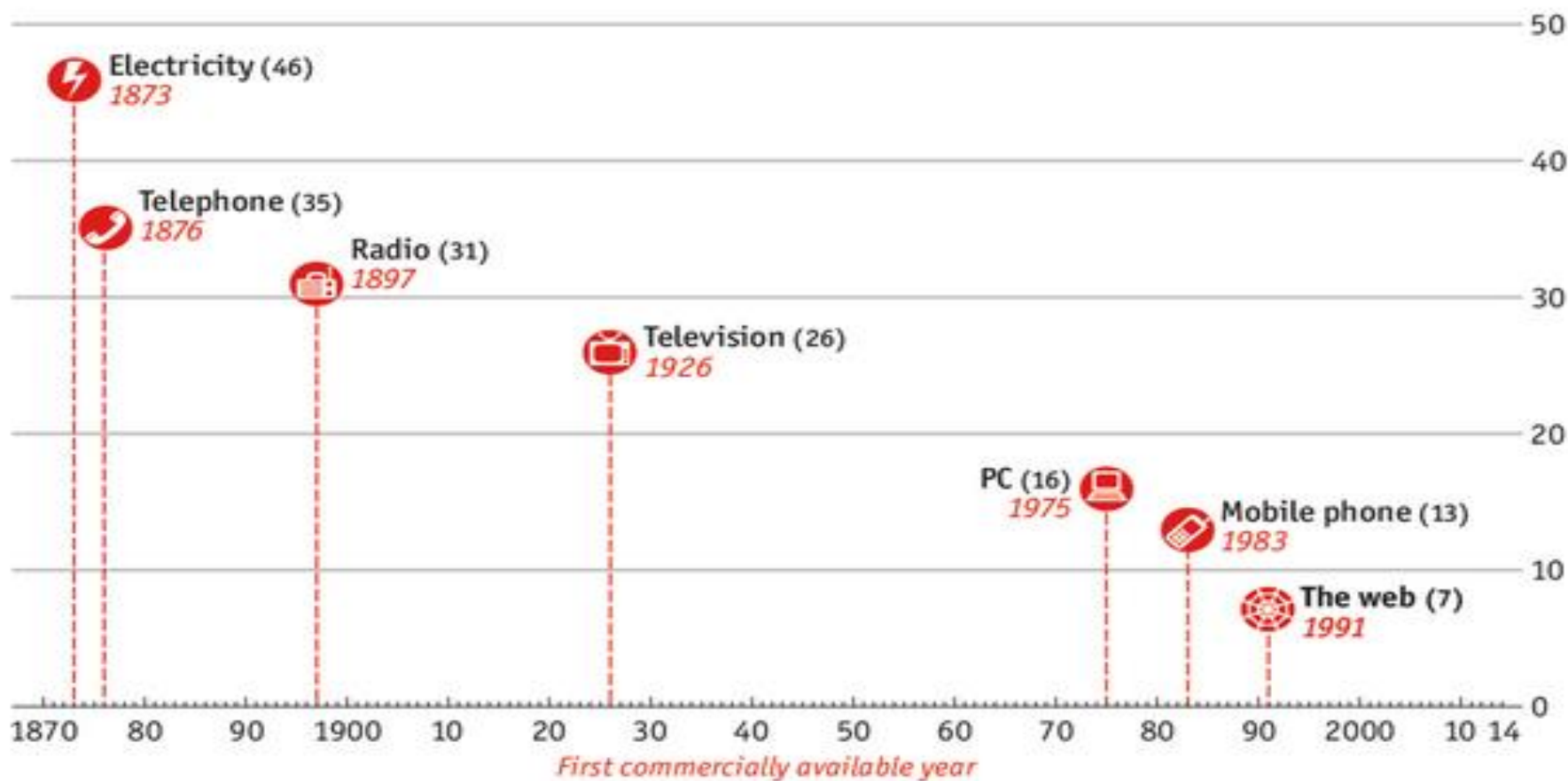
Why is InChI a Success

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.**

Technology adoption

Years until used by one-quarter of American population



Source: Singularity.com

Economist.com/graphicdetail

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 all scientists is in **FINDING** and **LINKING** information**

Internal

Your representation (e.g. WLN, SMILES)

Your format(s)

External

Same representation (Standard InChI/InChIKey)

Same one format

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.

Critical words/phrases for InChI

Link

Addition; not replacement

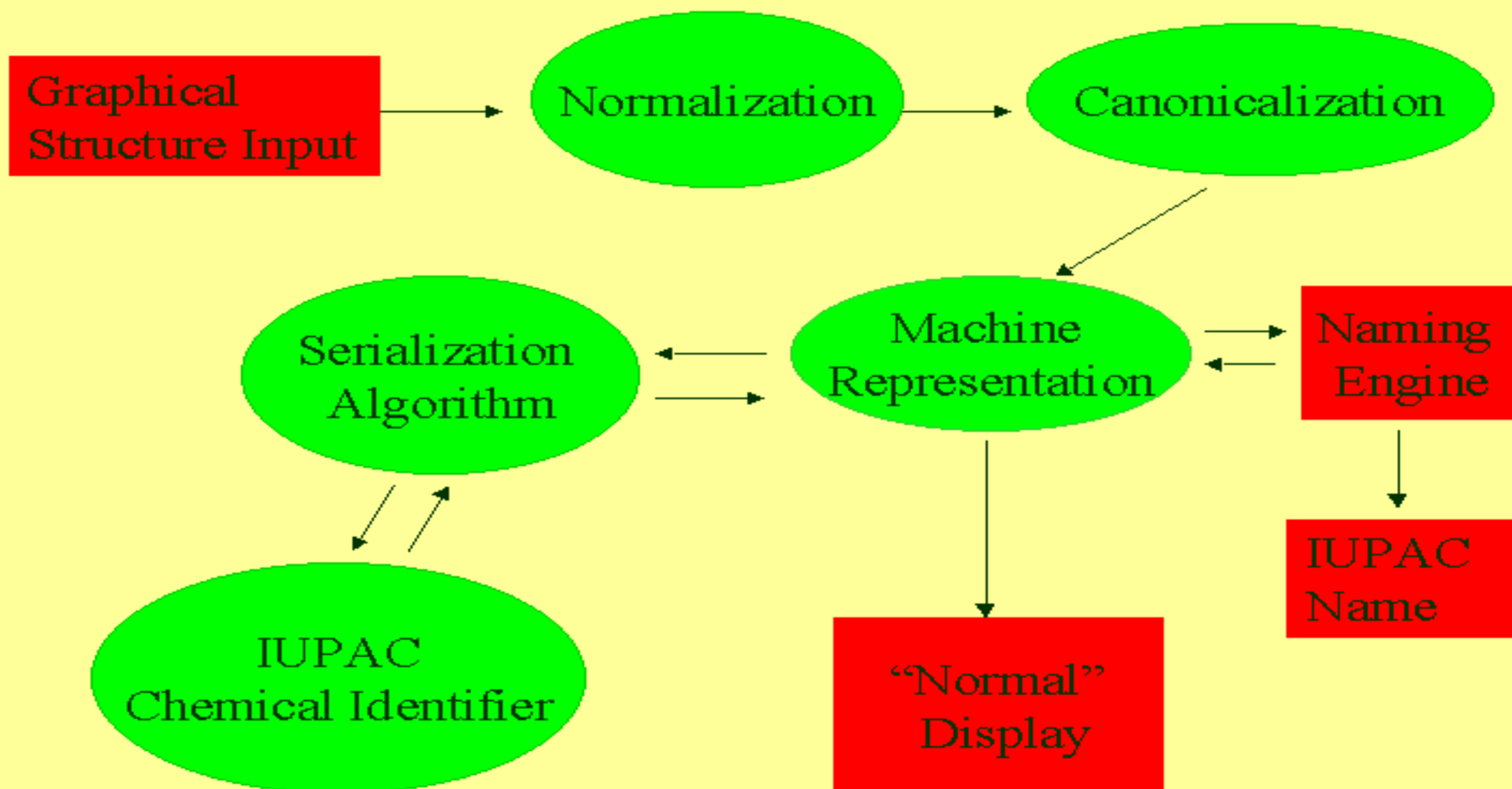
Algorithm

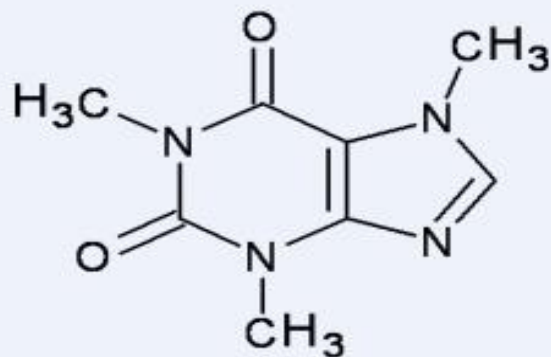
Synonym

No bureaucracy/Almost no staff

Decentralized

A Bottoms Up Project





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)

Second block (8 letters)

Encodes stereochemistry and isotopes

First block (14 letters)

Encodes molecular skeleton
(connectivity)

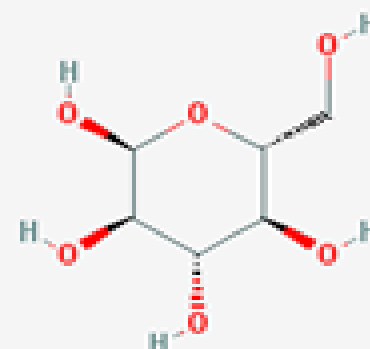
InChI TRUST



InChI is a string

InChI=**1**S/**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

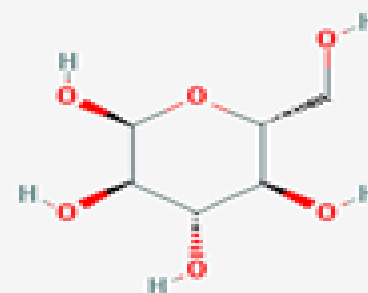
InChIKey is a “hashed” InChI

- 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

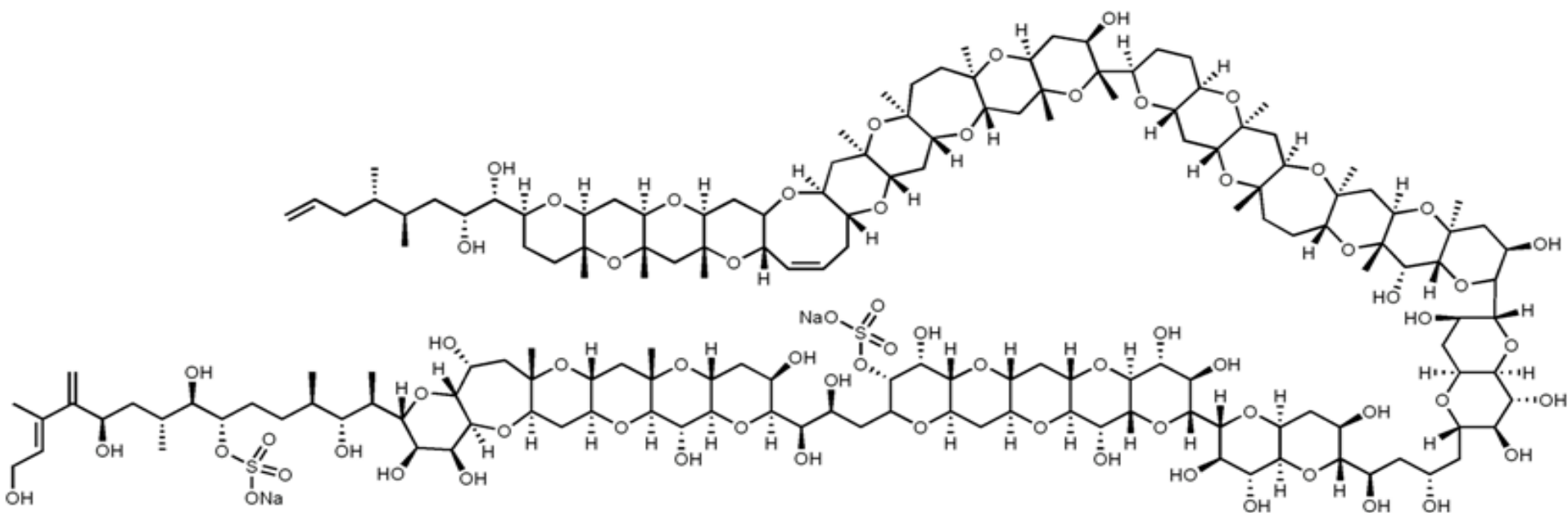
“layered” line notation



alpha-D-Glucose

InChI for Maitotoxin (from Nextmove Software, UK)

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;;/h24-25,28,35,65,67-69,71-149,165-192H,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





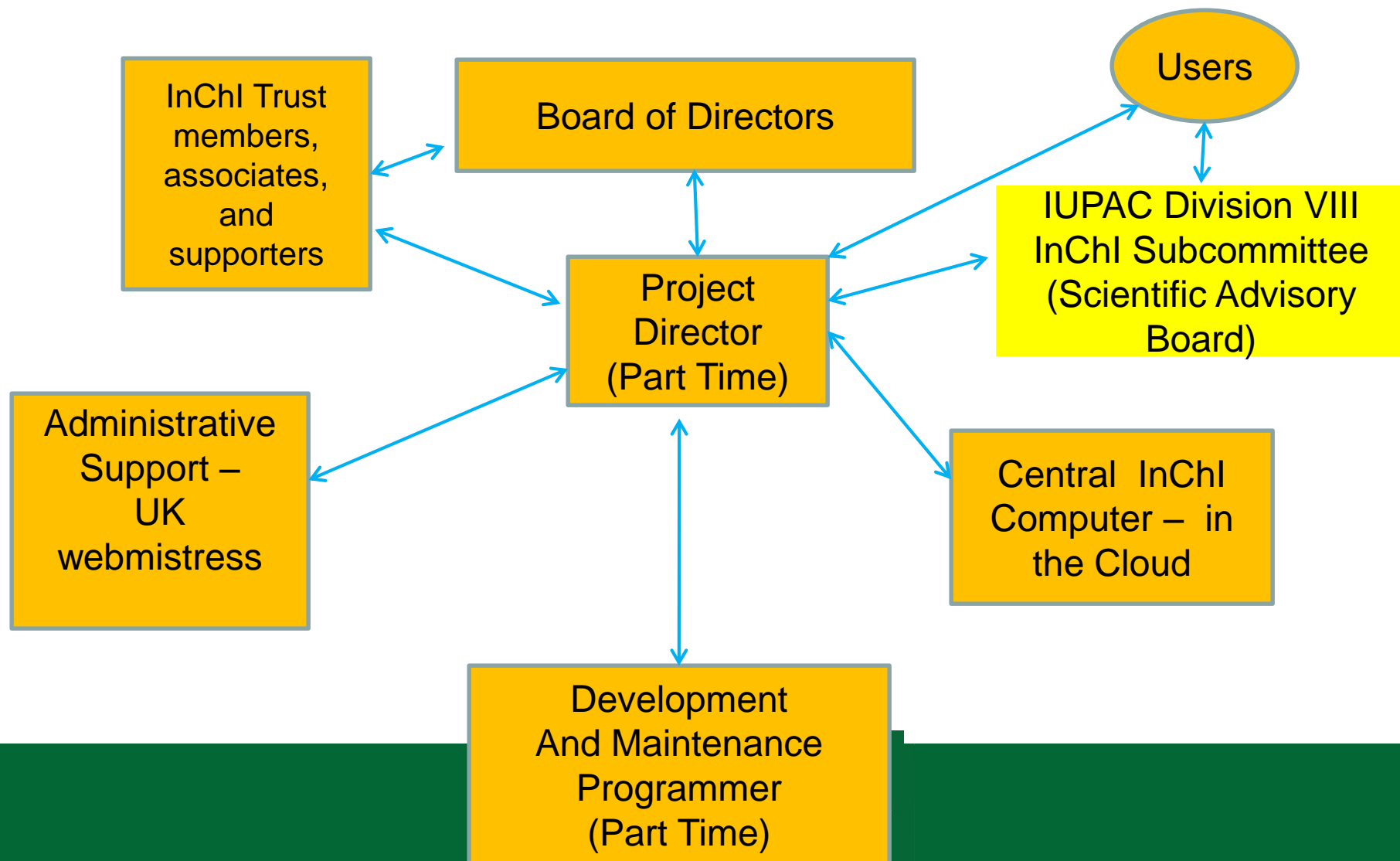
What about funding ?

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





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 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

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**

ACS Acknowledgements



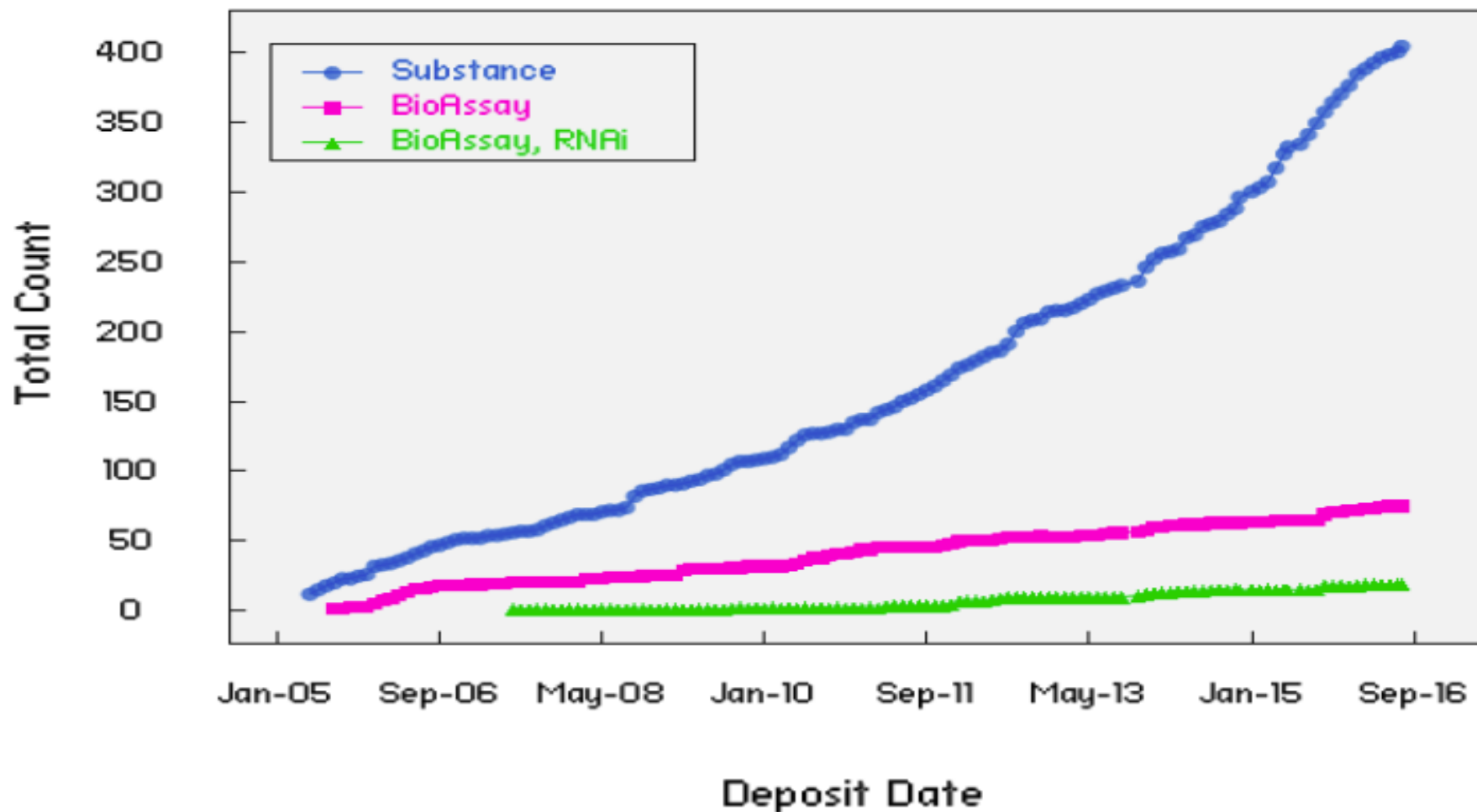
Acknowledgements

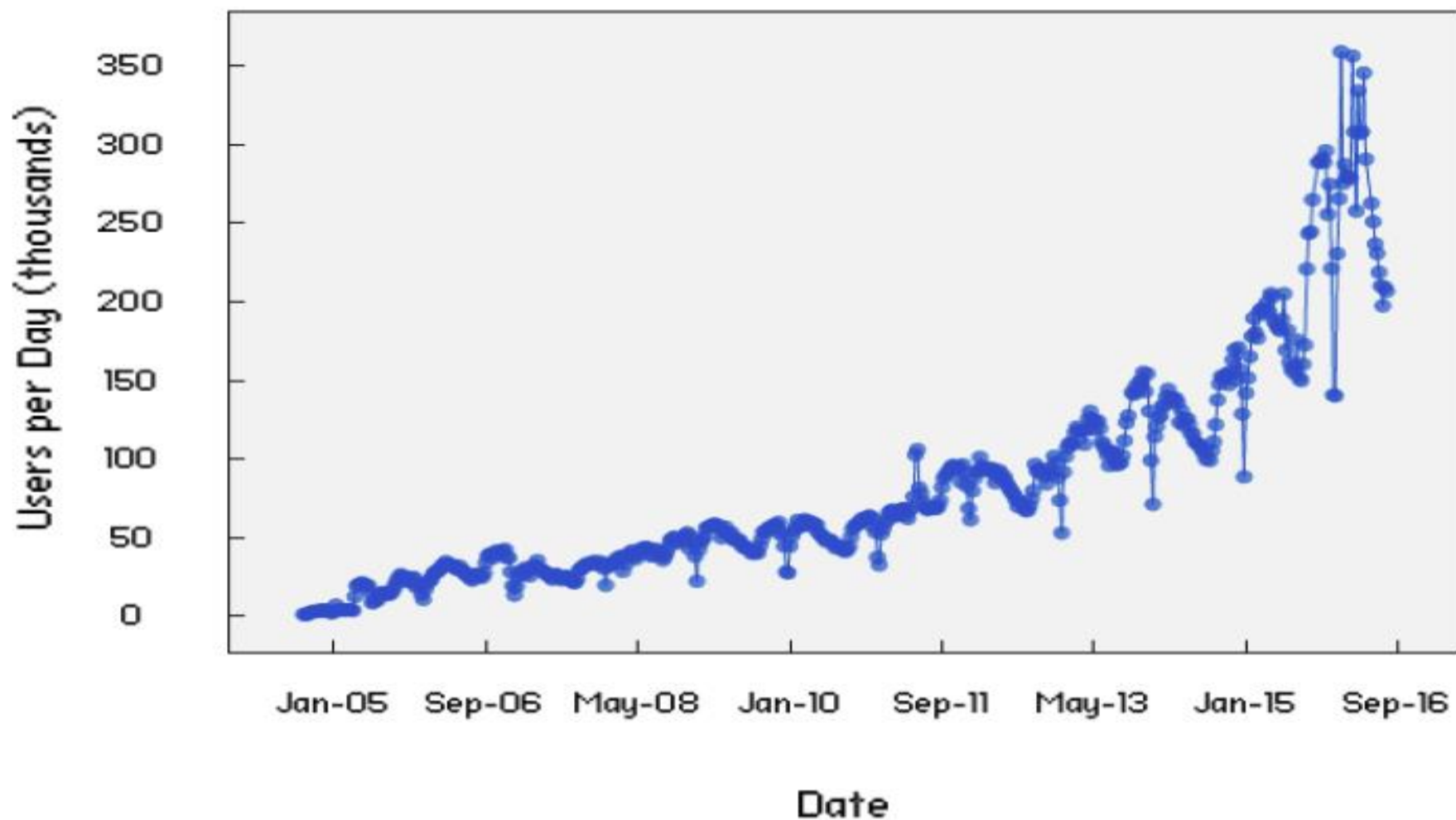
(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Colin Batchelor, John Barnard , Evan Bolton, Steve Boyer, Ian Bruno, Steve Bryant, Dominic Clark, 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.

Growth In PubChem Contributing Organizations





PubChem Usage

100,000+ searches a day
1.6 million unique users/month

PubChem, like InChI –

Success is uncoerced adoption

Have any questions?

If you think of a question later, email me:

steve@inchi-trust.org

