What is the Internet doing to chemistry and our brains?

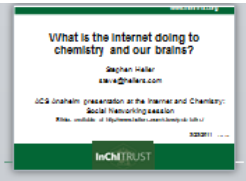
Stephen Heller

steve@hellers.com

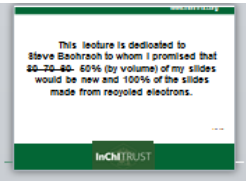
ACS Anaheim presentation at the Internet and Chemistry: Social Networking session

Slides available at http://www.hellers.com/steve/pub-talks/

This lecture is dedicated to Steve Bachrach to whom I promised that ~~80 70 60~~ 50% (by volume) of my slides would be new and 100% of the slides made from recycled electrons.



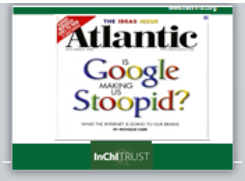
1



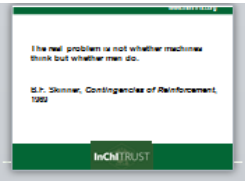
2

3

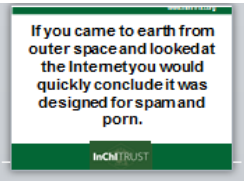
4



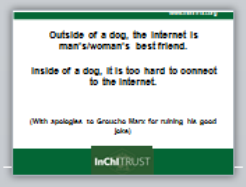
5



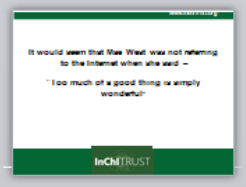
6



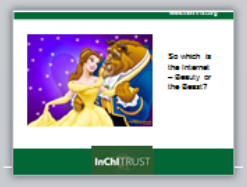
7



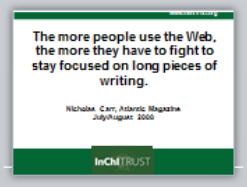
8



9



10



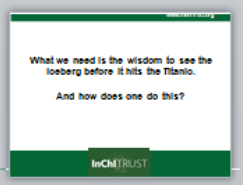
11

12

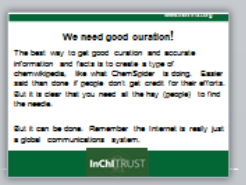
13

14

15



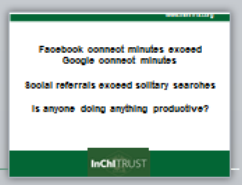
16



17

18

19



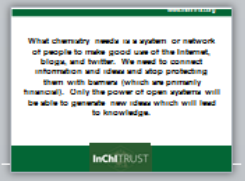
20

21

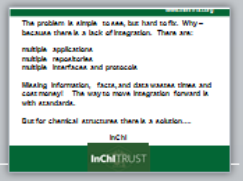
22

23

24



25

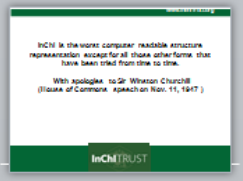


26

27

28

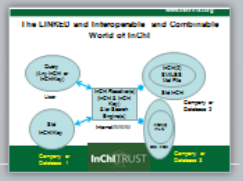
29



30

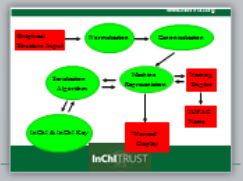
31

32

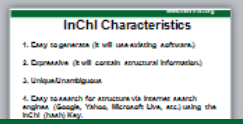


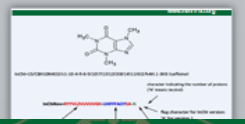
33

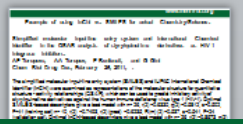
34



35







InChI TRUST

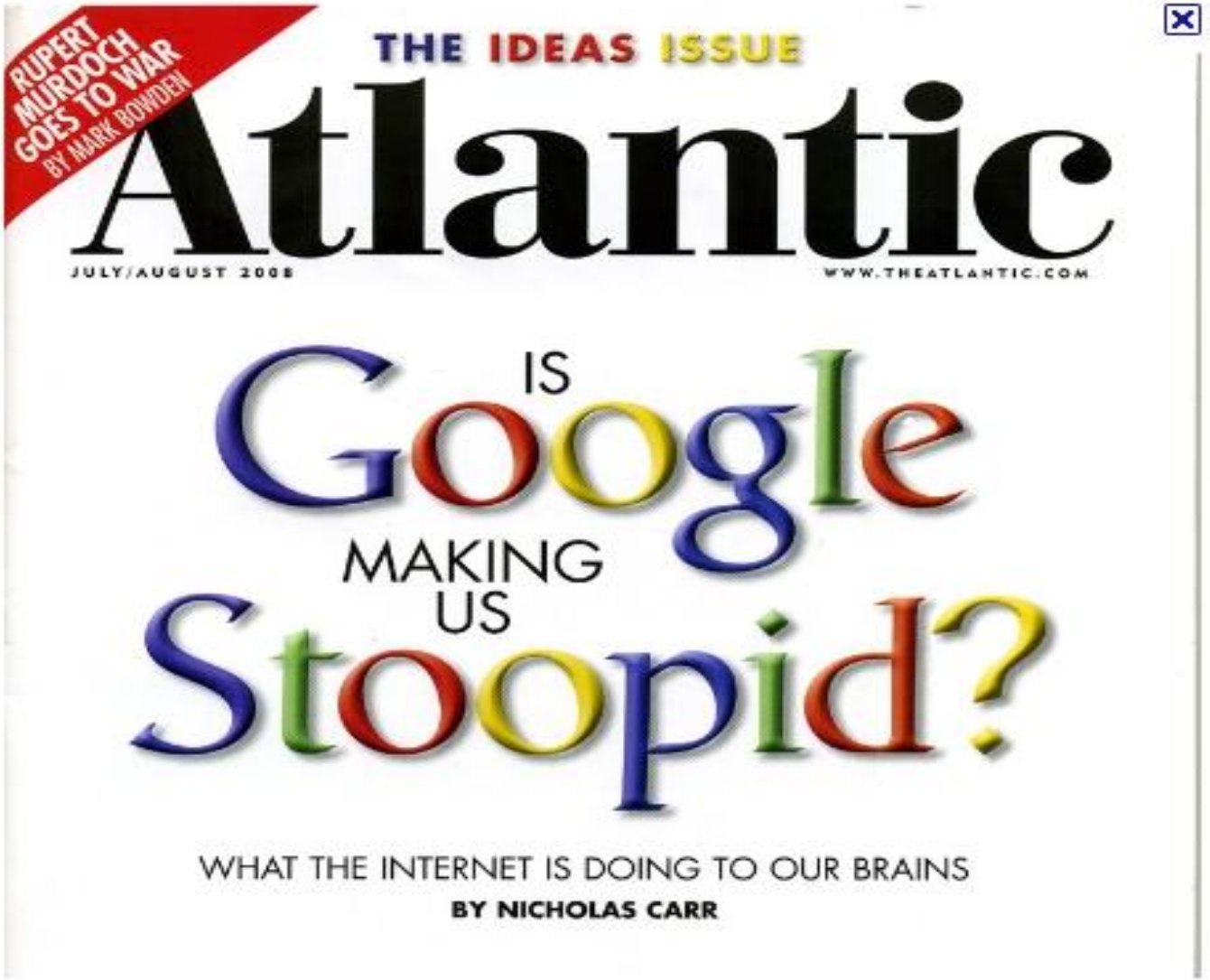
Outline

Hardware
Internet
Data Quality
Social Networks
InChI

Hardware developments over the past few decades:

Mainframe
Timesharing
Mini-computer (DEC, Wang)
Desktop
Laptop
Netbook
Smartphone
iPod
Very Smartphone
Cloud (and reduced software costs)
Genius smartphone
?? Is next

Let me assure everyone that it is a completely false rumor that the US State Department still uses Wang computers for their word processing. They have all migrated to PDP-8's.


RUPERT MURDOCH GOES TO WAR
BY MARK BOWDEN
THE IDEAS ISSUE
Atlantic
JULY/AUGUST 2008
WWW.THEATLANTIC.COM
IS
Google
MAKING
US
Stoopid?
WHAT THE INTERNET IS DOING TO OUR BRAINS
BY NICHOLAS CARR

The real problem is not whether machines think but whether men do.

B.F. Skinner, *Contingencies of Reinforcement*, 1969

“ A Trinity College Dublin survey showing that a third of Brits under 30 can’t remember their own home land-line number. 'Our gadgets have eliminated the need to remember such things anymore.‘ “

Maureen Dowd, NY Times, 3/9/11

If you came to earth from outer space and looked at the Internet you would quickly conclude it was designed for spam and porn.

Outside of a dog, the Internet is man's/woman's best friend.

Inside of a dog, it is too hard to connect to the Internet.

(With apologies to Groucho Marx)

It would seem that Mae West was not referring to the Internet when she said –

“Too much of a good thing is simply wonderful”

So which is the Internet – Beauty or the Beast?

The more people use the Web, the more they have to fight to stay focused on long pieces of writing.

Nicholas Carr, Atlantic Magazine
July/August 2008

Japanese poetry

I am busy now;
The Internet has stolen
So much precious time

Martin Marks
The New Yorker
10/25/2010, page 49

It is clear that users are not reading online in the traditional sense; indeed there are signs that new forms of “reading” are emerging as users “power browse” horizontally through titles, contents pages and abstracts going for quick wins. It almost seems that they go online to avoid reading in the traditional sense.

Nicholas Carr, Atlantic Magazine
July/August 2008

There is a tsunami of supply of information – published and unpublished – but what is authoritative and how do you separate quality from quantity?

Peer review – flawed but a currently accepted social standard (fact checking is down; after fact checking is up)

Wikipedia – flawed but a good social experiment

For the most part, what has the Internet brought us far?

1. Massive amounts of information, but chemistry is not a mass market. We have more information, but it seems we have less time to use it.

2. Print to digital (i.e., pdf's)
(The Internet Journal of Chemistry was a unique exception to this, but was too far ahead of its time to survive. It is dead, but 10 years later it is still ahead of its time.)

3. Social networks – Facebook (500+ million users), LinkedIn, Twitter (200+ million users). 2+ billion tweets a month; 1+ billion tests a month. Number of intellectual conversations per month – 42.

4. Mobile applications

5. Direct Selling – i.e., No middleman

But very, very little new and usable content.

What we need is the wisdom to see the iceberg before it hits the Titanic.

And how does one do this?

We need good curation!

The best way to get good curation and accurate information and facts is to create a type of chemwikipedia, like what ChemSpider is doing on a small scale. Easier said than done if people don't get credit for their efforts. But it is clear that you need all the hay (people) to find the needle.

For a start I would propose that editors and reviewers require all data to be published as supplemental material with all manuscripts. Even this will take a major change in attitude and procedures.

But it can be done. Remember the Internet is really just a global communications system.

Texting & Tweeting- The new heroin's

Do chemists go out anymore for a beer to discuss a lecture or a manuscript – or is that not possible with only video conferencing, texting, and tweeting. Do chemists interact face-to-face with one another very much anymore?

The trouble with social networks is that it is getting harder and harder for others to complain about you behind your back.

“The Bed of Procrustes”
Nassim Taleb
(author of The Black Swan)
Random House 2010

Facebook connect minutes exceed Google connect minutes

Social referrals exceed solitary searches

Is anyone doing anything productive?

The next two slides are examples of hard working, driven chemists drawing structures and doing searches on their phones.


ups
1W42

Trial by Twitter & Blogs

Blogs and tweets are ripping papers apart within days of publication.

NPR Radio report:

“Scientists discover keys to long life. Who will live to be 100? Genetic tests might tell.

Blog:

“We expect that most of the results of this study will not have the same longevity as its participants”

http://www.nature.com/news/2011/110119/full/469286a.html?s=news_rss

Sex predators target children using social media
By Byron Acohido, USA TODAY (3/1/2011)

Sexual predators, pornographers and prostitution rings are capitalizing on the rising popularity of mobile devices and social media to victimize children, say police and child safety experts.

The popular practice of making declarations about oneself on social websites puts valuable intelligence into predators' hands, say police and child safety experts.

CyberTipline, the nation's hotline for reporting sexual exploitation of children, received 223,374 reports in 2010, nearly double the 2009 number.

The soaring use of social networks, online games, smartphones and webcams has translated into "more opportunities for potential offenders to engage with children," says Ernie Allen, CEO of the National Center for Missing & Exploited Children.

GOD TEXTS THE TEN COMMANDMENTS

1. no1 b4 me. srsly
2. dnt wrshp idols/pix/jpegs/gifs
3. no omg's
4. no wrk on w/end (sat 4 now; sun l8r)
5. pos ok – ur m&d r cool
6. dnt kill ppl
7. :-X only w/ m8
8. dnt steal
9. dnt lie re: bf
10. dnt ogle ur bf's m8. or ox. or dnkey. myob.

(Thus, next time, God will eliminate the middleman – Moses – and just deal directly with the whole world.)

Exaptation

Exaptation means a shift in the function of a trait as it evolves. While usually associated with biological evolution, it seems well suited for the Internet and how it is used in chemistry and other areas. Use began as PC's were glorified typewriters to submit manuscripts to journals. Then chemists used computer readable conversion of simple text coupled with text searching. This was followed by electronic versions databases, then manuscripts, and various types of computational activities (modeling, lab notebooks, and so on). But these are still all just facts, not knowledge. The future of chemistry and the Internet must be to move and evolve past this and create knowledge.

What chemistry needs is a system or network of people to make good use of the Internet, blogs, and twitter. We need to connect information and ideas and stop protecting them with barriers (which are primarily financial). Only the power of open systems will be able to generate new ideas which will lead to knowledge.

The problem is simple to see, but hard to fix. Why – because there is a lack of integration. There are:

multiple applications
multiple repositories
multiple interfaces and protocols

Missing information, facts, and data wastes times and cost money! The way to move integration forward is with standards.

But for chemical structures there is a solution….

InChI

“No, no, not another structure standard!!!”

Why InChI? - Too Many Identifiers

Structure diagrams
- various conventions
- contain ‘too much’ information

Connection Tables
- MolFiles, SMILES, ROSDAL, …

Pronounceable names
- IUPAC, CAS, trivial

Index Numbers
- EINECS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAF

Why Use InChI

For publishers and database providers using InChI gives one a competitive advantage being able to LINK content from multiple sources. It offers users the ability to help in new discoveries from existing information and data by easily being able to integrate, remix, and retell. InChI is a small, but vital, part of new business models and technologies involving chemicals that will lead to new discoveries. Combinability increases the value of information and data.

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)

Why InChI is becoming a success

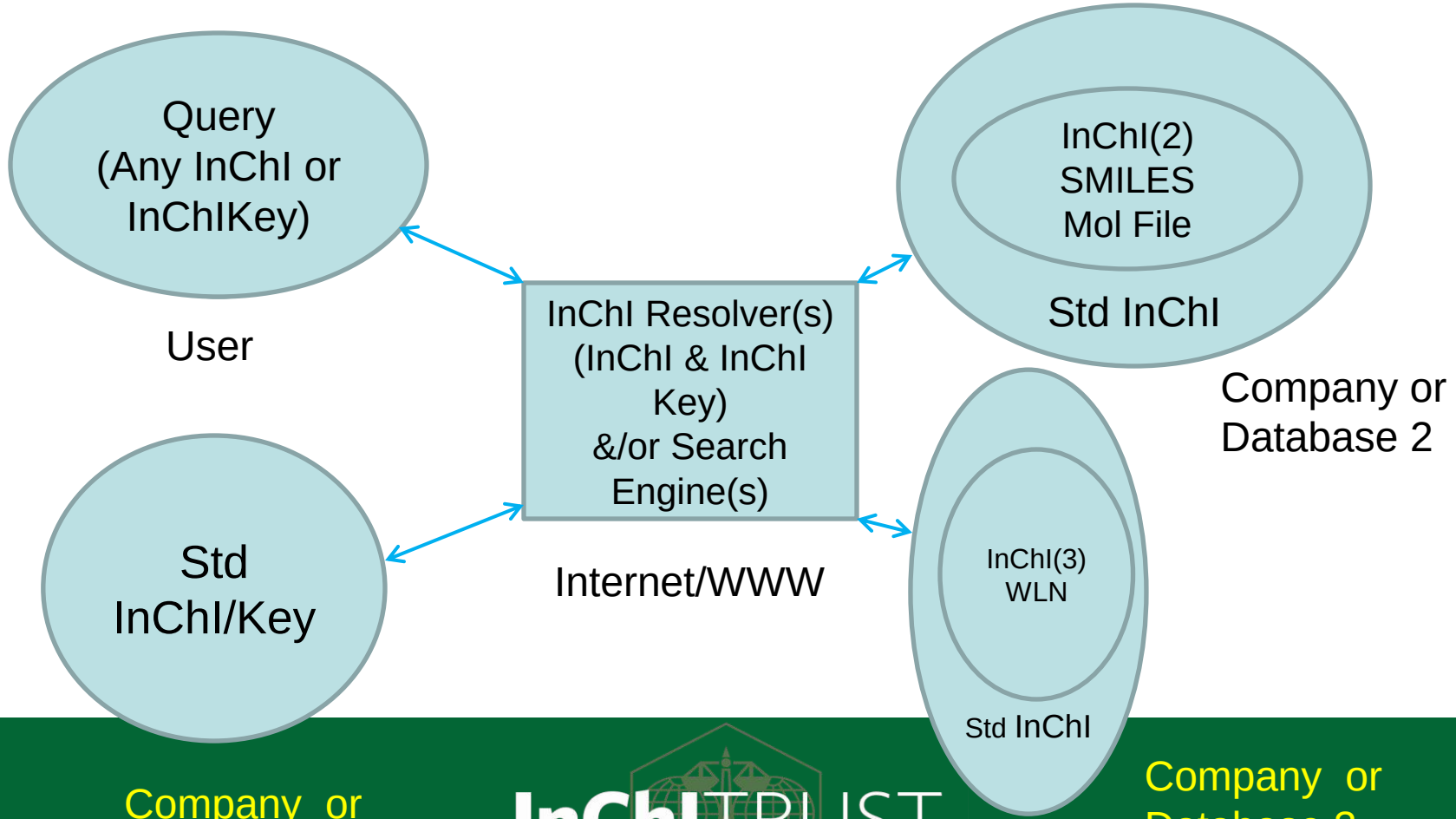
1. Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be LINKED to and combined with other content on the Internet.

2. InChI is a public domain algorithm that anyone, anywhere can freely use. By giving away the algorithm the project is building trust with the community.

How do we know the InChI project is beneficial?

Success is uncoerced adoption

The LINKED and Interoperable and Combinable World of InChI



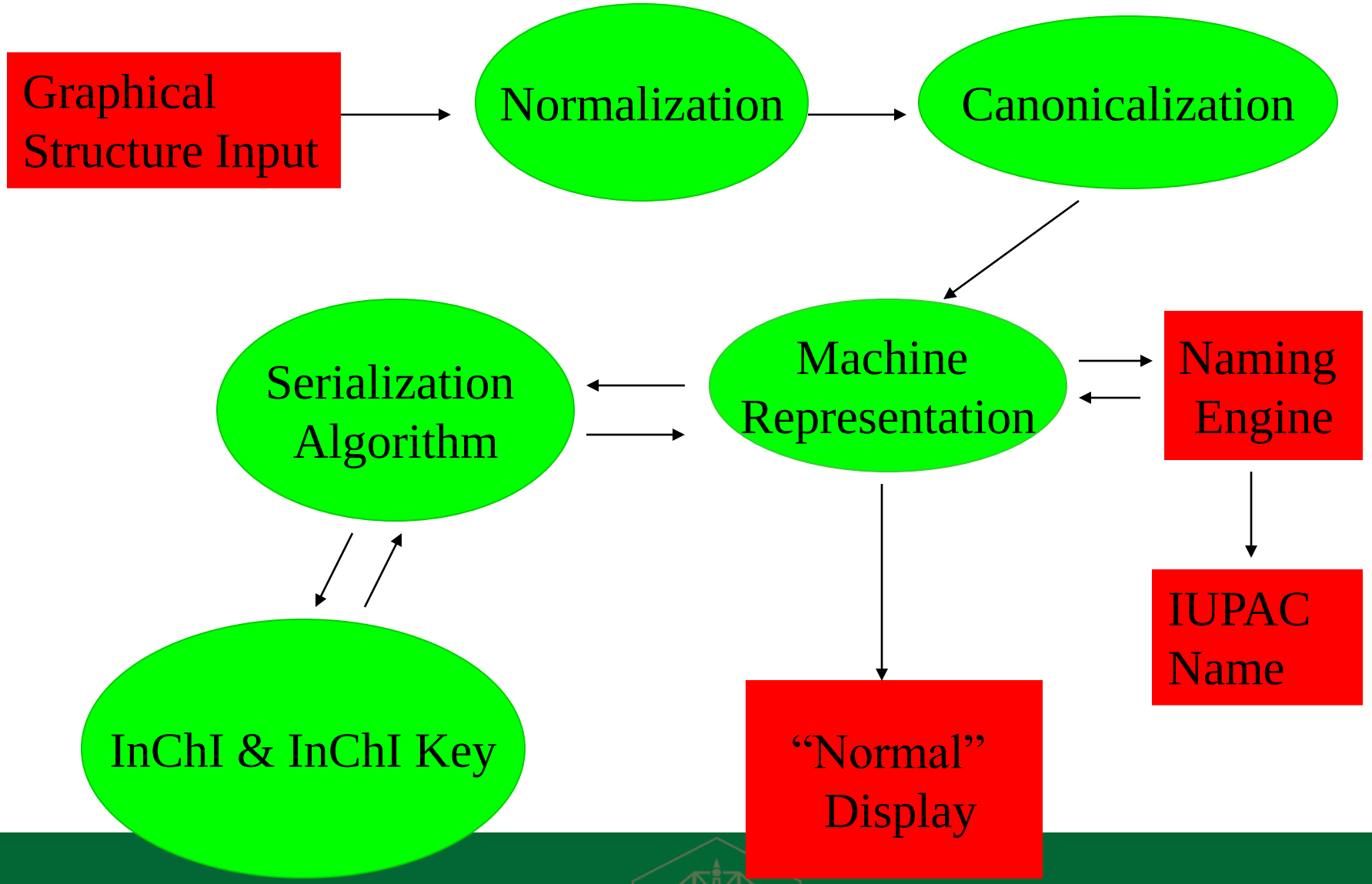
InChI Policy & Culture

Do not go outside our circle of competence.

No mission creep.

Staff is not territorial.

InChI Trust is doing well because it really doesn't require a lot of resources.


Graphical Structure Input
Normalization
Canonicalization
Machine Representation
Serialization Algorithm
Naming Engine
IUPAC Name
InChI & InChI Key
"Normal" Display

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

InChI Characteristics

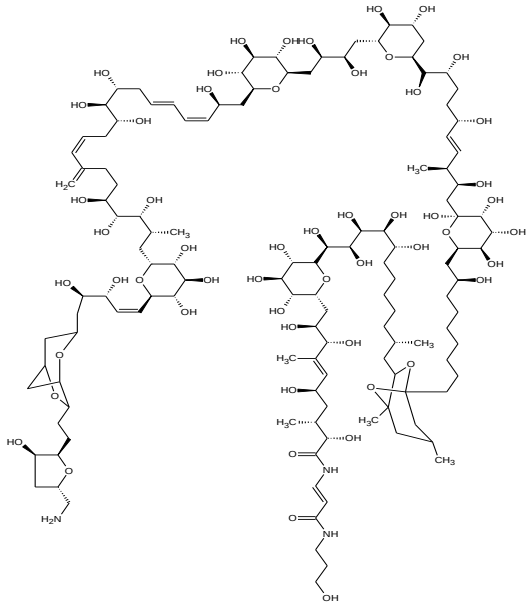
1. Easy to generate (It will use existing software.)

2. Expressive (It will contain structural information.)

3. Unique/Unambiguous

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

Really long InChI (Palytoxin)



Palytoxin

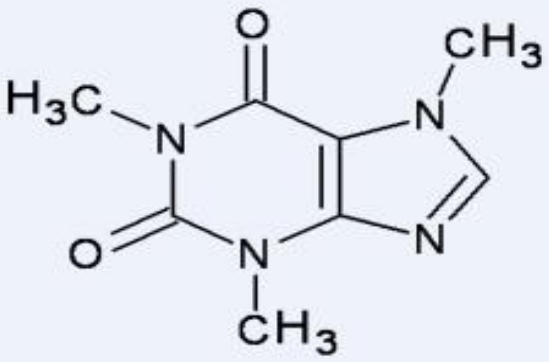
Isolated from Hawaiian soft coral

One of the most toxic non-peptide substances

Contains >70 stereochemical elements

InChI=1S/C129H223N3O54/c1-62(29-33-81(143)108(158)103(153)68(7)47-93-111(161)117(167)110(160)91(180-93)36-35-76(138)82(144)51-73-50-74-53-92(178-73)90(177-74)38-37-89-85(147)52-75(61-130)179-89)23-20-28-78(140)105(155)77(139)26-18-13-16-25-70(135)48-94-112(162)118(168)113(163)97(181-94)55-84(146)83(145)54-95-107(157)87(149)57-96(182-95)106(156)80(142)34-32-69(134)31-30-65(4)88(150)60-129(176)125(174)123(173)115(165)99(184-129)49-71(136)24-15-10-9-11-19-40-128-59-64(3)58-127(8,186-128)100(185-128)44-63(2)22-14-12-17-27-79(141)109(159)116(166)120(170)122(172)124-121(171)119(169)114(164)98(183-124)56-86(148)102(152)66(5)45-72(137)46-67(6)104(154)126(175)132-42-39-101(151)131-41-21-43-133/h13,16,18,20,23,25,30-31,35-36,39,42,45,63-65,67-100,102-125,133-150,152-174,176H,1,9-12,14-15,17,19,21-22,24,26-29,32-34,37-38,40-41,43-44,46-61,130H2,2-8H3,(H,131,151)(H,132,175)/b18-13+,23-20-,25-16-,31-30+,36-35-,42-39+,66-45+/t63-,64?,65-,67+,68+,69+,70+,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?,102+,103+,104-,105-,106?,107-,108+,109-,110+,111-,112-,113+,114-,115-,116-,117-,118+,119+,120+,121-,122-,123+,124?,125+,127?,128?,129-/m0/s1

InChIKey=CWODDUGJZSCNGB-DCBUCRFRSA-N



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 Certification Suite

The InChI certification suite is a software package developed and 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).

Once the programs are run and the results sent back to the Trust, an "InChI certified" logo will be sent to person/organization. The InChI Trust certification logo can then be put on the pages of the web site for all users to see.

The certification suite software is provided at no cost to all Trust members to use the logo in their business activities. For Trust supporters the suite and logo is provided at no cost, but may only be used for non-commercial activity.

The cost of certification suite to non-members is US $5000 per year.

The certification suite was developed for the Trust by GGA Software Services LLC.

Example of using InChI vs. SMILES for actual Chemistry/Science:

Simplified molecular input-line entry system and International Chemical Identifier in the QSAR analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors.

AP Toropova, AA Toropov, E Benfenati, and G Gini

Chem Biol Drug Des, February 26, 2011

The simplified molecular input-line entry system (SMILES) and IUPAC International Chemical Identifier (InChI) were examined as representations of the molecular structure for quantitative structure - activity relationships (QSAR), which can be used to predict inhibitory activity of styrylquinoline derivatives against the human immune deficiency virus type 1 (HIV-1). Optimal SMILES-based descriptors give a best model with n= 26, r(2) =0.6330, q(2) =0.5812, s=0.502, F=41 (training set) n= 10, r(2) =0.7493, r(2) (pred) =0.6235, R(m) (2) =0.537, s=0.541, F=24 (validation set). Optimal InChI-based descriptors give a best model with n= 26, r(2) =0.8673, q(2) =0.8456, s=0.302, F=157 (training set); n= 10, r(2) =0.8562, r(2) (pred) =0.7715, R(m) (2) =0.819, s=0.329, F=48 (validation set). **Thus, the InChI-based model is preferable**. The described SMILES-based and InChI-based approaches have been checked with five random splits into the training and test sets.

While there has considerable progress in the take-up and use of InChI, we still need to deal with the fact that chemists are very conservative and change their habits slowly, but

even glaciers are moving a lot faster these days due to climate change.

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.

Most people with visions of the future should go and see my brother-in-law, the optometrist.

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, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Randy Marcinko, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau , Marc Nicklaus, Carmen Nitsche, Matthias Nolte , Igor Pletnev, Josep Prous, 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.