# The IUPAC InChI Chemical Structure Standard – Today and the Future

Stephen Heller

NIH/NLM/NCBI

**The main web sites for the IUPAC InChI project are:**
**http://www.iupac.org/inchi**
**http://www.inchi-trust.org**

**FDA Cheminformatics Workshop October 18-20, 2023**

**Slides are available at http://www.hellers.com/steve/pub-talks/fda-2023.pdf**

# The opinions presented on these slides are those of the slides, not the presenter.

**Feel free to interrupt at any time – it will probably improve my talk immensely.**

# Technical success is uncoerced adoption

# Sustainable success is permanent sustained funding

# What about funding?

**This talk will cover two areas of the InChI project. The non-technical (management and infrastructure) and the technical (chemistry and programming). The former area being rather depressing while the latter is rather encouraging.**

# What is InChI by Chat GPT

The Chemical Identifier (InChI) is a textual notation used to represent chemical structures in a way that is both human-readable and machine-readable. It was developed by the International Union of Pure and Applied Chemistry (IUPAC) to provide a standardized and unique way of describing chemical compounds.

InChI is designed to represent the connectivity of atoms in a molecule, as well as stereochemistry and isotopic information. It consists of a series of alphanumeric characters and layers of information. The basic format of an InChI consists of:

1. The version identifier (InChI=1/): This indicates the version of the InChI standard being used.
2. The main layer: This layer contains the chemical formula and connectivity information, representing how atoms are bonded together in the molecule.
3. Sublayers: These sublayers provide additional information, such as stereochemistry, isotopic information, and charges.

InChI strings are unique for each chemical compound, meaning that no two different compounds will have the same InChI. This makes InChI a valuable tool for database searches, chemical structure comparisons, and the exchange of chemical information between researchers and databases.

In summary, the InChI (Chemical Identifier) is a standardized and machine-readable way to represent the structure of chemical compounds, providing a unique and consistent means of describing molecules in the field of chemistry.

sustainability

In the broadest sense, sustainability refers to **the ability to maintain or support a process continuously over time.**

**A quick history of InChI starting from the Big Bang as to how I was able to get sustainability of the mass spectrometry database which was started at NIH. The NIH/EPA/NIST mass spectral database began when I was at NIH in1970. NIH management, at that time, did not like anything by R01 research grants so the project was not high on the list of projects at the Heart Institute. I moved to EPA which needed a way to analyze polluted air and water and used mass spec primarily to do it. To be able to see if new spectra were duplicates/better quality, etc. in the mid 1970's I chose to contract with CAS for Registry Numbers. Funding and support issues at EPA led me to convince OSRD at NIST in the early 1980's to take over the project and actual copyright (unique in the US Government) the database to keep control and be able to fund it. Hence the database because quite sustainable – to the about of $6-7 million income per year.**

**Skipping forward a decade or so to about 1999, CAS decided it no longer wanted to keep the original contract of a couple decades supplying NIST with CAS Registry Numbers to deal with duplicates. Between this and my need to do something in retirement, the InChI project was initiated. The first email from me to Steve Stein in the NIST mass spec lab documenting the idea of an IUPAC Chemical Registry System was 11/15/1999.**

**At the same time IUPAC was realizing the need to move into the 21st century and they organized a meeting in March 2000 on Representations of Molecular Structure: Nomenclature and its Alternatives. It was a success and IUPAC (who had no resources to do this) asked NIST (who would not do it without being formally asked by IUPAC) to undertake the project. The initial InChI software was finished in 2005.**

**Rather than continue to maintain and develop this standard (as many had thought made sense since NIST is the scientific standards agency of the US Government) which has become very widely accepted, the lab decided it wanted nothing to do with any ongoing efforts with InChI as its mass spec database needs were satisfied by the original brilliant work of Dmitrii Tchekhovskoi.**

I was able to convince a few publishers as to how they could benefit from having an InChI, especially when they could not have a CAS Registry Number attached to the chemical in their publications. The bottom line of all this history is InChI needed a home to support ongoing work of maintenance and expansion. Thus, the InChI Trust was established to fund the management and further funding for the development of InChI.

**This all worked well for about 15 years with InChI becoming a real and well accepted standard. Sufficient funds from membership dues in the Trust were available to support the rather low level of enhancements to and maintenance of the original work with only myself and Igor Pletnev being funded for any InChI promotion and programming work.**

**But now are now in 2023 we have a funding problem. As you all know the InChI algorithm has been and continues to be freely available. The only source of income are membership dues to the Trust. Between dues of $5,000 to $25,000 a year or $0 (zero dollars if you choose to be a free loader) it is probably not surprising there is a problem. That zero dollar membership fee has been winning amongst organizations around the world. In fact, there have been more organizations discontinuing their support that the number of organizations continuing their financial support.**

# InChI Trust Membership History/Status

Paying Members: (10)

Elsevier
ACS/CAS
RSC
Springer-Nature
John Wiley
NIH/NLM/NCBI
Chinese Chemical Society
ACD Labs
OpenEye
ChemAxon

Previous Members (14)

OntoChem
FDA
BioRad
Taylor & Francis
IBM
Sigma Millipore/Aldrich
Perkin Elmer/Cambridgesoft
Mcule
Zinc project
Microsoft
Thompson Reuters
U of California
CCD
Proquest

Contract signing problems(2) : Google & University of Luxemburg

**The last new member to join the Trust was in 2020: (3 years ago – Chinese Chemical Society)**

**If you can't keep your current supporters, how do you convince new ones to join? Is there a real foundation/sustainability for InChI – or is it something a few people are championing? It is very nice to have champions, advocates, and supporters when you start a new project but eventually they retire or die. You need a broader list of funders.**

**For those who may have seen a recent InChI poster presentation there was a list proposing ways to financially support InChI:**

**1. Become a member of the Trust**
**2. Make a charitable donation to the Trust**
**3. Make an in-kind contribution (i.e., a programmer)**
**4. Work through Trust certified consulting companies**

**Only number 3 has happened**

**This is not good. All suggestions on charging for InChI downloading have been rejected by the Trust. The only choice left to the Trust was to reduce expenses, which Is what is planned for 2024 and onwards. Some of the technical work (programming) has been solved in the short run by some in-kind programming support from Germany – Aachen University and the Beilstein Institute. But this is not likely to be permanent or likely sustainable, and the required in-kind support needs to agree on what specific improvements (organometallics but not, for example large molecules or extended tautomers) to InChI that they would support. What the project obviously needs is a clear and permanent supporter to be sustainable. No Government agency (e.g., NIST )-seems to want to or be able to do that for whatever reasons. The Trust will soon be running out of dues income and reserves. The project needs a champion or an advocate, or a dedicated supporter or a hero to become sustainable.**

**The best definition of insanity is doing the same thing over and over and expecting a different outcome.**

**InChI needs a change, a revamping, a renovation**

# Now that I have depressed most everyone I do have some potentially good news.

**There has been interest from an outside party to make the project sustainable, but it is still very early in the process, and I am not at liberty to divulge the organization.**

# End of part 1 (sustainability) and start of part 2 (technical)

# Now comes the status of the science and programming part of this talk – all the chemical issues technical issues of InChI and the InChI Key

# 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 layered structure design

**The current InChI layers are:**

**1. Formula**
**2. Connectivity (no formal bond orders)**
  **a. disconnected metals**
  **b. connected metals**
**3. Isotopes**
**4. Stereochemistry**
  **a. double bond (*Z/E*)**
  **b. tetrahedral (sp3)**
**5. Tautomers (on or off)**

**Charges are added to end of the string**

**The InChI Algorithm normalizes chemical representation and includes a "standardized" InChI, and the 'hashed' form called the InChIKey**

# 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.**

# Large Databases with InChIs/InChIKeys

**EBI UniChem –157 million**
**NIH/NCI – 120 million**
**NIH/PubChem -115 million**
**RSC/ChemSpider – 128 million**
**Elsevier/Reaxys – 179 million**
**IUPAC – 0 million**

**(Numbers come from Google searches and may not be accurate except for the IUPAC number)**

# InChI is FAIR

**The most critical and valuable thing InChI does is support the FAIR (Findable, Accessible, Interoperable, and Reusable) activities that make chemistry FAIR.**

**InChI/InChIKey is an identifier. InChI and the InChIKey are machine readable. It is Non-proprietary. It is reusable by all, as anyone can freely download and use it. It has become a standard for structure representation and is now used universally used by chemistry, biochemistry, and related resources. InChI and the InChIKey are critical for data exchange/interoperability. And lastly InChI and the InChI Key makes things findable by many search engines.**

# Maintenance of InChI Code

As we all know too well the InChI code requires maintenance. And the code requires security updates. And changing computer languages and computer architectures with time (operating systems, hardware, compilers, etc.). cases. And expansion of applicable use cases (e.g., new entity types). And finding and fixing bugs/errors (for which we thank Google staff for their ongoing help in this matter.). And lastly, newly encountered new chemistry. The IUPAC Division VIII Chemical Nomenclature and Structure Representation Division has assured me they will never run out of new chemical structures which will need to be named and *for* which there will be a need to generate an InChI.

# Keeping up with scientific advances

**As Evan Bolton has pointed out to me - science, including chemistry, is not static. Constant evolution of science and related fields make accurate (machine) identification critical/crucial. Evolving use cases and needs by the chemistry community. Advancing science requires constant input from IUPAC InChI sub-committee members to improve and expand the capabilities of the InChI algorithm and to innovate and find new ways of doing things.**

# Update on InChI technical issues

**With difficulty and with time, the efforts of Igor Pletnev who worked on the InChI software for over a decade and died suddenly of COVID in 2021 have been taken over by Gerd and others.**

**With the recent in-kind support from Aachen University and the Beilstein Institute the project has 2+ full time people working on expanding the capabilities of the InChI algorithm – primarily molecular inorganics (which includes organometallics). Some support from the to-be-hired cheminformatics person at the Beilstein Institute will also be available shortly.**

# Technical topics agenda and overview

Release timetable/status

Code review

Testing

Organometallics

New Architecture

# InChI Timetable

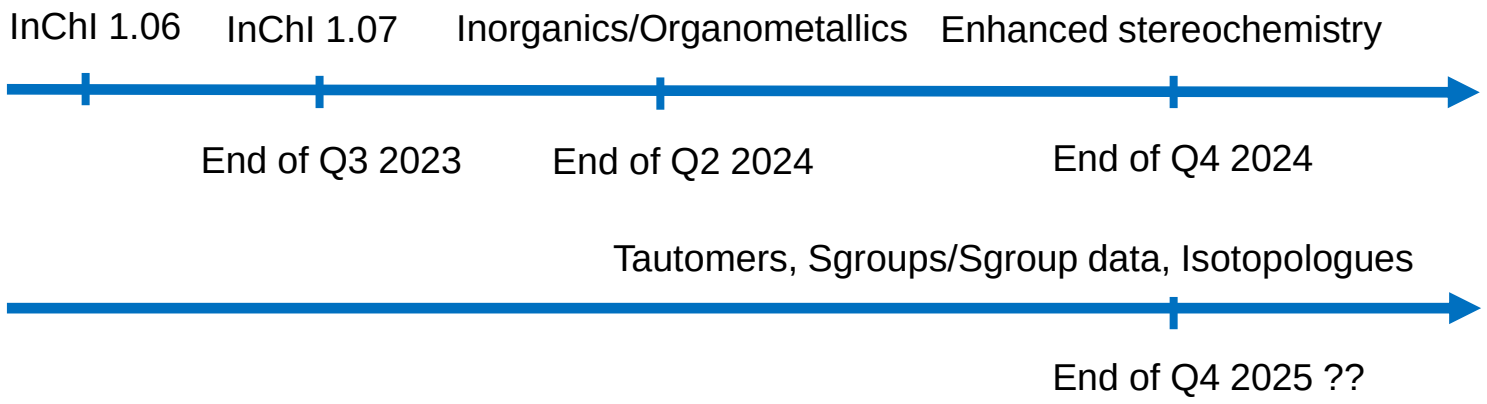
## InChI

- Released (1.06, December 2020)
  - Polymers (in beta status)
- Code clean-up (1.07)
  - Polymer in standard InChIs
  - Tautomer transformations
- Under development
  - Requirements (nearly) ready
    - Molecular inorganics
    - Stereochemistry
    - Tautomers
    - Sgroups and Sgroup data
    - Isotopologues
- Longer evaluations
  - Variable molecules / Markush InChIs
    - Prototypes being tested
    - https://github.com/topics/inchi
  - Large molecules

## InChI Application Framework

- Released
  - RInChI 1.0
- Published
  - QR code
- Test phase
  - InChI Web Demo
    - https://iupac-inchi.github.io/InChI-Web-Demo/
  - Resolver
    - https://github.com/inchiresolver/inchiresolver#readme
  - RInChI (1.1)
    - https://github.com/IUPAC-InChI/RInChI
- Awaiting coding
  - MInChI
    - Prototype released
      - https://github.com/cdd/mixtures
  - RInChI (1.2)
- Longer evaluations
  - NanoInChI (NInChI)

# InChI Timetable




Test: InChI Web Demo, Resolver, RInChI 1.1
Coding: MInChI, RInChI 1.2

Evaluations:
Large molecules
Variable molecules / Markush InChIs
NanoInChI (NInChI)

All technical developments depend on further funding.

# Code review - Google® Oss-Fuzz

- Switch to Google development environment based on **Clang**
  - Stricter interpretation of C syntax
  - An additional set of more than 4,000 warnings have been detected and cleaned
    - A few of them are serious buffer issues
  - Side effect: InChI becomes about 10% faster

# Testing

- Test sets
  - MCULE: test sets with 2,000, 20,000, and 200,000 structures
  - Test set of RInChI 1.06
    - Extended by new findings and other known “trouble makers”
  - PubChem

# Testing

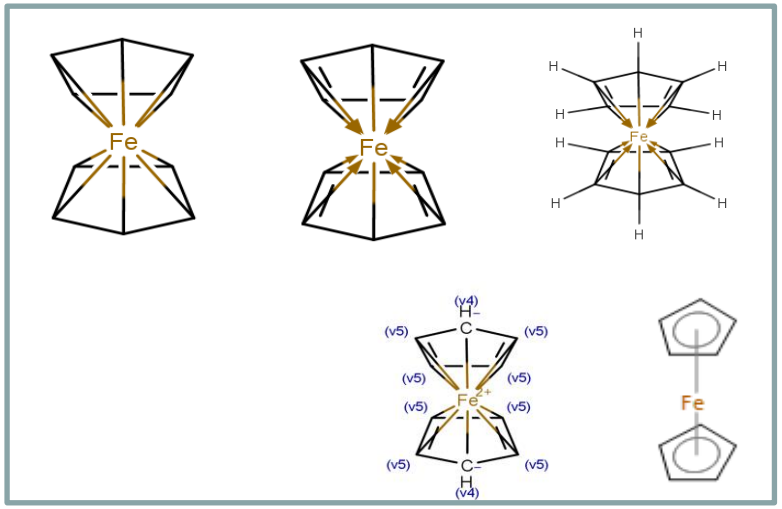
## PubChem Tests

- Standard InChI comparison1.06 / 1.07 (without polymers)
  - Simple regression tests (no invariance etc.)
    - About 7 hours on a Linux box with 16 cores
    - 4 deviations within more than 100 million structures!
      - » 2 worked in 1.06 but not in 1.07 and vice versa
  - Further tests will include additional options like the invariance

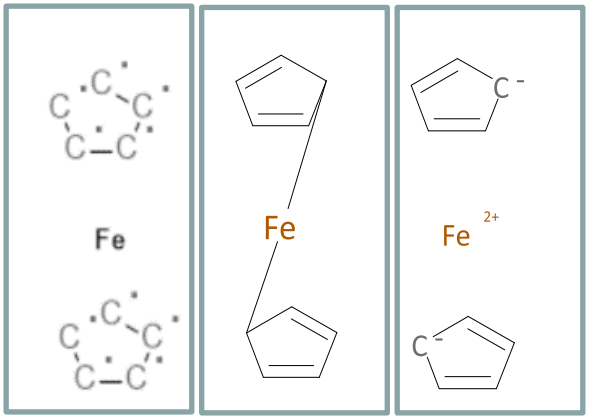
# Organometallics

Identity rules for organometallics, inorganics and coordination compounds defined

– Example Ferrocene




Equivalent representations




Non identical representation

(taken from Hinnerk Rey's Cambridge presentation, June 2022)

# Molecular Inorganics, Organometallics and Coordination compounds

- Definitions must be fully elaborated
  - Full set of examples and test sets
- Stereochemistry of organometallics
  - Per-Ola Norrby's proposal of opposite pairs
  - nextMove's Cambridge representation via affiliation boards
  - Ulrich Schatzschneider's stereographs

# Proposed new architecture


– Preprocessor for structure normalization
  - Tautomer transformation
  - Organometallics normalizations
– Main processor: canonicalization and connectivity
– Postprocessing for additional layers

# Working Group Efforts

Working groups of the InChI subcommittee are currently working on extended stereochemistry and extended tautomers, as well as the OER (Open Education Resource) and inorganics which are part of the overall organometallics efforts – now called molecular inorganics.. Efforts are being made to find someone to take the lead in the InChI extension to large molecules. Positional Isomers/Variability/Markush are underway as are additional capabilities for some areas of polymers. Lastly work continues on adding to and improving RInChI. Programmers are looking into full V2000/V3000 Molfile support.

# Summary

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

# Acknowledgements

**(Current and past members for the IUPAC InChI subcommittee and associated InChI working groups – 10/23)**

**Steve Bachrach, Djordje Baljozovic, Colin Batchelor, John Barnard, Bob Belford, Evan Bolton, Ray Boucher, Steve Boyer, Jan Brammer, Ian Bruno, Steve Bryant, Alex Clark, Szabolcs Csepregi, Rene Deplanque, Josef Eiblmaier, Vincent Scalfani, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Sonja Herres-Pawlis, Nauman Kahn, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Frank Lange, Alexander Lawson , Peter Linstrom, Gary Mallard, Leah McEwen, Bill Milne, Hunter Moseley, Moss, Peter Murray-Rust, Heike Nau, Marc Nicklaus, Carmen Nitsche, Matthias Nolte, Wendy Patterson, Steffen Pauly, 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.**