InChI: part of the solution

Richard Kidd
@rkiddr
@InChI_Trust

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 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)

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, David Evans, Jeremy Frey, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Steve Heller, Jaroslav Kahovec , Richard Kidd, Hans Kraut, Alexander Lawson , Peter Linstrom, Gary Mallard, 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.


ACD/Labs


ALDRICH
Chemistry


BIO-RAD


ChemAxon


CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY


IUPAC


FDA


NIH
National Institutes
of Health


NIST
National Institute of
Standards and Technology
U.S. Department of Commerce


OpenEye


RELX Group


ROYAL SOCIETY
OF CHEMISTRY


SPRINGER NATURE


Taylor & Francis Group
an informa business


WILEY
Publishers Since 1807

What has succeeded?

disambiguation
linking

at a price of

arbitrary but agreed compromise
normalisation

What can we do?

What should we do?

How far?

Extension *vs* application?

How fast?

1.04 in 2011

1.05 and RInChI in 2017

Who is ready for InChI v2?

Extension

Tautomerism
Organometallics
Large molecules
Positional isomers
Inorganics
Markush

Tautomerism

Created a database of experimental tautomerism results

Approx. 1,700 cases; mostly pairs, some triples and higher
Will evaluate this DB for possibly additional types of tautomerism

Some advantageous rule changes identified

For how many of the counted cases are current InChIs different?
Final list of proposed transforms (with counts)
Vetting and commenting by Working Group and InChI community
Putting up for public comments
Final decision on Working Group's recommendations by August 2017

Tautomerism

Proposed transforms (“rules”) based on:

–Existing 20 tautomerism transforms in CACTVS
–Ca. 10 new rules not (yet) covered by CACTVS

No recreation of CACTVS tautomerism engine proposed!
Checked for occurrence rates in >290M cpds
–AMS, PubChem, CSDB, ZINC

Even rarest rules found applicable to >100 cpds

Organometallics

Higher stereochemistry
Delocalized bonding
Advanced normalization and business rules for drawings
(related) InChI validations and warnings

Determine tasks for stereo package for developer.
Identify supported V2000 extensions, idioms and bugs.
Review examples of advanced normalization and suggest drawing rules.

Large molecules

Discussed the “Who, What, When, Why, How” necessary to cross integrate HELM and InChI

Scoping
What is amenable to InChI in the short-term
What small molecule content is HELM-able
Planning
What can we achieve in the next year?
Community engagement / involvement

Java vs C/C++
MOL V3000 implementations
Extension of InChI to handle ambiguity

Positional isomers
Inorganics
Markush

Application?

Reactions
Mixtures
InChI Resolver
QR codes for InChI
InChI teaching/educational materials
Publishing

Reactions - RInChI

Future:

a) Error correction and Problem resolution

b) Collecting more data: bigger databases; cross-searching data sources

c) New features and applications

i. Stereochemistry issues –especially mixtures and mysteries

ii. Adaption to new InChIreleases; introduce a standard set of flags

iii. Include more data (in AuxInfo): must not create barriers to use

iv. Tools for analysis / regularisation(no multiple molecules)

v. Details of 'no structure' for metabolism, natural materials, etc

vi. Benefits of keys? definitely identifiers, unless there is a decoder

vii. RInChI *vs* RAuxInfo *vs* ProcAuxInfo

viii. Generic transformations: SMIRKS-alike -use Markush; ??; ??

Reactions - RInChI

How best to draw reactions to avoid ambiguity?
Give authors advice on how to represent reaction schemes
Building on stereochemistry and general papers for structures
Should be recommended to students, examiners, *etc*
Need similar advisory for reactions; will be simpler

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

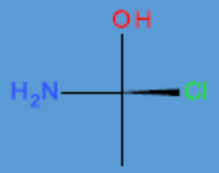
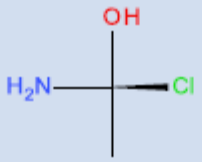
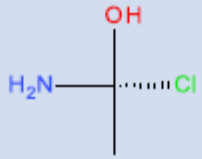
Use /n layer to notate repeated InChI units (e.g., tautomers) no structures, AND/OR logic within isomer groups

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

Workflow for file input and conversion

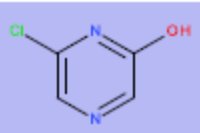
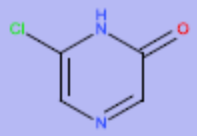
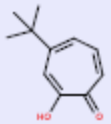
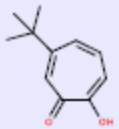
Mixtures - MInChI

Racemic mixtures

Example	Meaning	InChI	MInChI
AND Enantiomer 	**Mixture of**  and 	Standard InChI=1S/C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/m0/s1 (Note: absolute stereochemistry!) Parameter /SUCF or /SREL InChI=1/C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/s2 Parameter /SRAC InChI=1/C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/s3	MInChI=0.0.1S/C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/m0/s1 & C2H6ClNO/c1-2(3,4)5/h5H,4H2,1H3/t2-/m1/s1 /n{1&2}

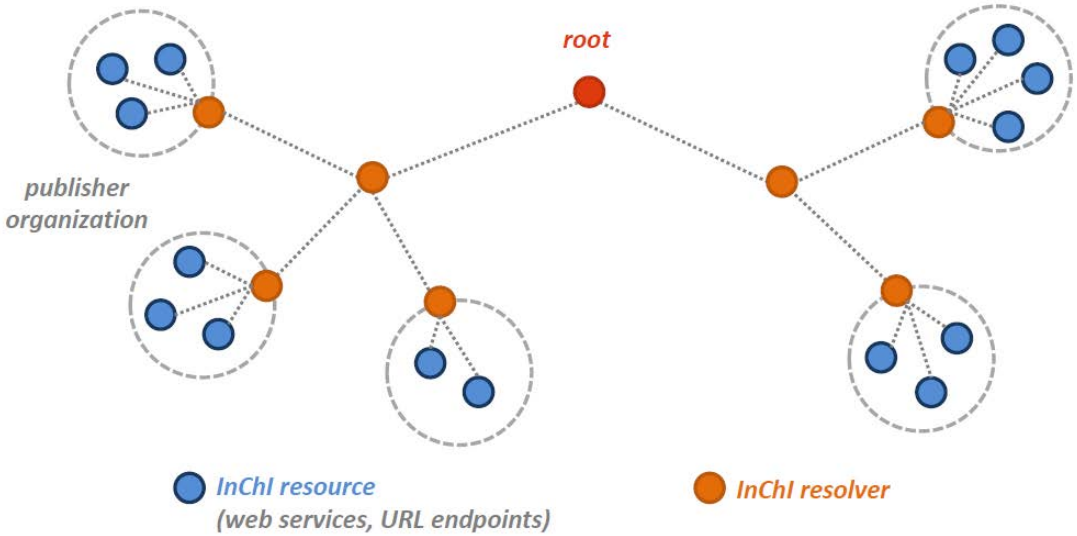
Mixtures - MInChI

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)

Education and support Publishing

How to support a standard?

No written standard
Software is the implementation
Centralised expert resource

How should we encourage the best open source practices?
Community input and contributions while maintaining a standard

Governance and funding
Participation through time and money

Open file formats

Define open standards for an important subset of widely used formats used for interchanging chemical information: SMILES, SMARTS, CTAB/Mol

For each format:

Standardized documentation of format, including extensions
What is not allowed
Examples and validation set
Open reference implementation
Stewardship

State and Future of the IUPAC InChI

16-18 August 2017

NIH Bethesda

Signup: www.inchi-trust.org

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