IUPAC International Chemical Identifier (InChI) Subcommittee

Minutes of the meeting on March 23rd 2009 at the Sheraton City Centre, Salt Lake City, UT, USA

Present: *Subcommittee members:*

Steve Heller (Chairman)
Colin Batchelor (Royal Society of Chemistry)
Evan Bolton (US National Center for Biotechnology Information)
Alan McNaught (InChI project coordinator, Cambridge, UK)
Marc Nicklaus (US National Cancer Institute)
Steve Stein (NIST)
Chris Steinbeck (European Bioinformatics Institute)
Dmitrii Tchekhovskoi (ex-officio developer) (NIST)
Graeme Whitley (Wiley, New York)
Jason Wilde (Nature, London)
Tony Williams (ChemSpider)
Andrey Yerin (Advanced Chemistry Development, Moscow)

Observers:

Steve Bachrach (Trinity University, San Antonio, Texas)
Richard Kidd (Royal Society of Chemistry)
Peter Linstrom (NIST)
Dave Martinsen (American Chemical Society)
Marcus Sitzmann (US National Cancer Institute)
Keith Taylor (Symyx Technologies, CA)

Apologies: *Subcommittee members:*

Sandy Lawson (Elsevier, Frankfurt)
Igor Pletnev (ex-officio developer) (Moscow State University)

1.0 Minutes of the previous meeting etc.

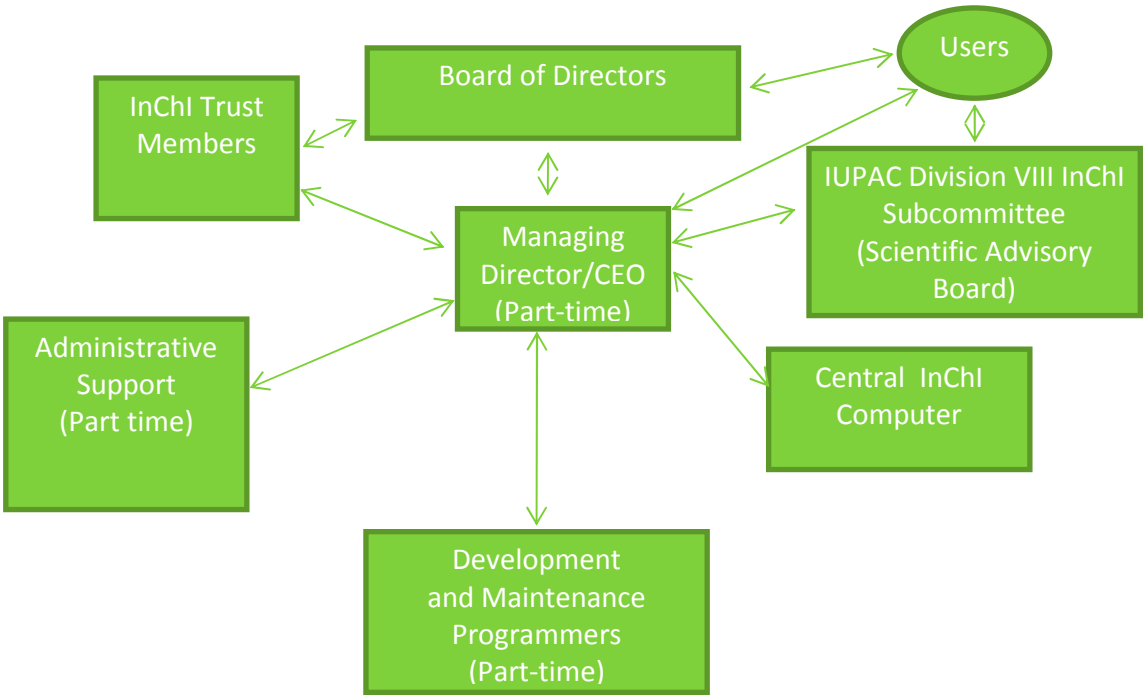
The minutes of the meeting at NIST on September 15th 2008 had been modified in the light of comments received, and re-circulated. Steve Heller circulated copies of correspondence with IUPAC Officers and a summary from Igor Pletnev of issues raised in feedback from the release of InChI1.02final.

2.0 InChI Trust status and revised role of InChI subcommittee

Steve Heller reported progress with set-up of the InChI Trust, which would in future manage the development and maintenance of the InChI standard. When this was established, the role of the InChI Subcommittee would be modified: it would become a Scientific Advisory Board for the InChI Trust, while remaining a subcommittee of IUPAC Division VIII. It was noted that the IUPAC website did not yet include details

of the InChI Subcommittee; the IUPAC Secretariat should be asked to deal with this as soon as possible.

Relationships amongst the various bodies concerned are summarised in the following diagram:



Staff at the Royal Society of Chemistry were making arrangements for the Trust to be established in the UK; when this was done it would be possible to proceed with start-up administrative tasks and project work. FIZ-Chemie (Berlin) had offered to provide administrative and computer facilities for the Trust.

It was noted that IUPAC had recently approved funding towards continued development, maintenance and publicity for InChI up to the end of 2010. A suggestion from IUPAC Officers of an alternative scenario for the future, with IUPAC in charge of the Trust, was noted; the Subcommittee felt that this would be unrealistic when IUPAC was no longer providing funding for InChI.

The level of annual contributions to be required for membership of the Trust was still under discussion but was expected to be in the range $5000-25000 according to ability to pay. Exceptional zero contributions would be permitted if appropriate.

The InChI administrative office would probably be expected to deal with InChI correspondence, routing items to appropriate individuals, to take responsibility for or assist with maintenance of InChI information sources (e.g. InChIfaq, inchi.info); this should include an up-to-date list of projects involving InChI. The domain name inchi.org had been acquired.

InChI applications (such as the InChI Resolver) would be outside the scope of the Trust.

It was considered very unlikely that any external body would be in a position to patent the InChI algorithm.

Current commitments to the InChI Trust are as follows:

Royal Society of Chemistry	$25000
Nature	$25000
Symyx Technologies	$5000
FIZ-Chemie	admin and computer facilities

Portions of the contributions from RSC and Nature might be regarded as loans for set-up purposes; this was still under consideration.

The American Chemical Society/Chemical Abstracts had been kept fully informed of progress, and their participation would be welcomed.

Subcommittee members were asked to send suggestions of other potential participants to Steve Heller. He would review his mailing list for other possibilities.*

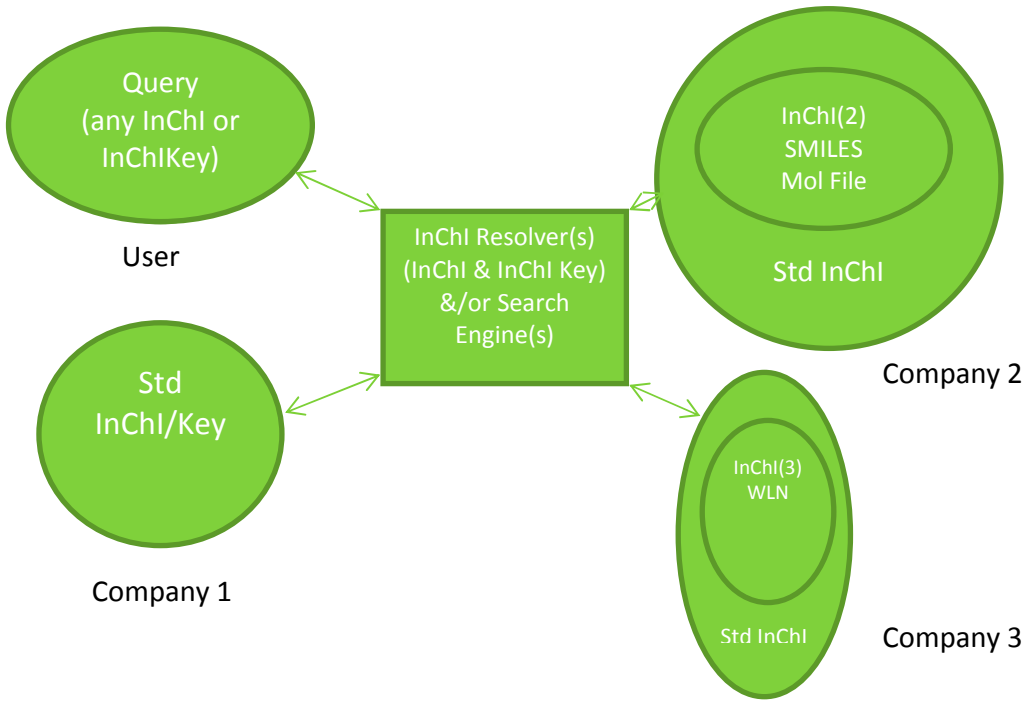
---

* *Secretary's note:* Other contacts both before and after the meeting include the following (if no comment, no reply received yet):

Taylor & Francis	$5000 committed
ChemSpider	request to join with fee waiver
Pistoia Alliance	probable
PNAS/Natl Academy of Sciences	
Google Open Source group	
BASF	
Microsoft	may consider in next fiscal year
Wiley	awaiting approval
American Chemical Society Publications	
Thomson	under consideration
Elsevier	under consideration
Thieme	probable
US National Cancer Institute	probable
PubChem	probable later in 2009
Merck Index	
Bio-Rad	
CambridgeSoft	under consideration
IBM (Steve Boyer)	
Fujitsu	
ChemIndustry	
IUPAC	
InfoChem (Munich)	possible in-kind donation for InChI/reactions
ACS-CINF Division	to be brought up at next board meeting
Bentham Science Publishers	
CSIRO Publishing	
Accelrys	do not wish to join at this time
NIST	do not wish to join at this time
JAICI	do not wish to join at this time

3.0 Standard InChI/InChIKey

The final version of InChI1.02 had been launched in January 2009, as an implementation for Standard InChI/InChIKey. Steve Heller outlined a scheme for appropriate use of Standard InChI/InChIKey by organisations with various requirements, as shown in the following diagram:



Members of the Subcommittee were concerned that version 1.02final had been implemented only for Standard InChI/InChIKey. They had expected a release of 1.02 including options for generating non-standard InChI and also the new format InChIKey, making the previous format obsolete. The current arrangement involved a bifurcation of the InChI source code. It was explained that the restriction to Standard InChI/InChIKey had been made to expedite the release. It was agreed that the preparation of a full release including non-standard options and the new InChIKey format for non-standard InChI should now be a top priority.

It was agreed that in future more control of development work would be advantageous, and steering groups would be established to control various aspect of future work (see minute 4.0); also it could be helpful, when feasible, to introduce an online programmers' development space. In any case, a full copy of all working developments should be stored at FIZ-Chemie, as well as an authoritative source for the current InChI release. In the interim, Steve Heller would ask Igor Pletnev to upload a weekly backup to NIST. As soon as possible, a second part-time programmer should be acquired.

It was noted that some facilities contained in the InChI source code are currently unactivated, and could be introduced if thought desirable. These are:

keto-enol tautomerism
1,5- tautomerism
ring-chain tautomerism in sugars

de-derivatisation

4.0 Future requirements

It was agreed that the top priority was now the full release of version 1.02 with both standard and non-standard options, as described in minute 3.0.

Second priority was the publication of full documentation of InChI, InChIKey and their Standard implementations. This would be most helpfully provided as two papers, one an authoritative presentation aimed at the InChI user, and the other a full technical specification. Igor Pletnev was currently preparing a paper for publication, and would be asked to report progress. It seemed likely that this paper would be the full technical account; in that case other people should be asked to prepare the user-oriented publication.

The requirements for further InChI development remained as specified in September 2008, as follows:

polymers
organometallics
extended stereo concepts
Markush structures
3-D structures
excited states
unattached groups
undefined substituents
interlocking structures (e.g. rotaxanes)

Also an InChI checker should be built into the software to enable a user to establish whether any particular InChI generator is legitimate.

Further suggestions were:

application of InChI to correct wrongly designated stereochemistry in the input structure

development of accurate round-tripping, perhaps based on auxinfo

5.0 InChI/InChIKey Resolver

The launch of the initial implementation of the RSC/ChemSpider InChI/InChIKey Resolver was welcomed. It was necessary to establish the IUPAC standard protocol for such resolvers to communicate with each other, and a sub-group would be set up to deal with this.

6.0 InChI/InChIKey-based reaction schema

A summer student at the Unilever Centre for Molecular Science Informatics in Cambridge, UK, was to be appointed to work on the first stage of this project.

7.0 Subcommittee membership

It was agreed that Keith Taylor (Symyx) should be added to the membership.

8.0 Subcommittee subgroups

The following sub-groups were established, to report back as soon as possible and in any event at the next meeting (July 30th):

Requirements for organometallic structures:

Colin Batchelor (coordinator)
Andrey Yerin
Marcus Sitzmann
Keith Taylor

InChI/InChIKey resolver protocol

Tony Williams (coordinator)
Evan Bolton
Marc Nicklaus
Marcus Sitzmann
Steve Bachrach

Structure input control (business rules)

Keith Taylor (coordinator)
Tony Williams
Andrey Yerin
Dmitrii Tchekhovskoi
Colin Batchelor
Wolf-Dietrich Ihlenfeldt (if willing)
A Nature representative (possibly)

A starting point for extension to polymers should be the notes of the InChI subgroup meeting in Prague on June 13th 2005 (attached).

Alan McNaught
15 April 2009

The IUPAC International Chemical Identifier (InChI): Promotion and Extension (project 2004-039-1-800)

Extension to polymeric systems

Notes of a meeting at the Institute of Macromolecular Chemistry, Heyrovského námestí 2 CZ-162 06 Praha 6, Czech Republic, on June 3rd 2005 at 9.00 am

Present: Dr Stephen Heller, Dr Jaroslav Kahovec, Dr Alan McNaught, Dr Andrey Yerin

1.0 **Objective.** The meeting was convened to consider what methods for two-dimensional depiction of macromolecular structures could be advantageously incorporated into the InChI specification. In discussion, reference was made to the 1994 IUPAC Recommendations on Graphic Representations of Macromolecules:http://www.iupac.org/publications/pac/1994/pdf/6612x2469.pdf

2.0 **Questions.** Dr Stephen Stein (absent Task Group member) had suggested the following questions for discussion:

2.1 *What, if any, features of polymer structure can be usefully 'canonicalized'?*

This was discussed in relation to representations recommended in the 1994 IUPAC Recommendations (see minute 1.0); conclusions are given in minute 3.0.

2.2 *Is the Chemical Abstracts representation of polymers helpful/adequate?*

Task Group members considered that the largely text-based representations used by CAS were not directly relevant to the extension of InChI.

2.*3 Is there any need to reflect source-based in addition to structure-based depiction?*

For a variety of reasons it was considered that InChI should continue to reflect actual structure without regard for origin. In particular, it was noted that there was no generally accepted source-based 2-D representation, and that the same macromolecular structure could arise from several sources. However, source information might usefully be included in the Auxiliary Information field.

2.4 *To what extent should we deal with the variability inherent in polymeric structure (crosslinking, variable end-groups, etc). Is there or should there be a polymer markup language (PolymerML?) or equivalent to separate the variables*

Task Group members were not aware of any moves to establish a markup language for categorising the parameters defining a macromolecular structure.

3.0 **Recommendations.** The following recommendations arose from consideration of the 1994 IUPAC Recommendations referred to above (minute 1.0) in the light of the questions listed in minute 2.0. Reference numbers (e.g. 2-E8) quoted for structures are those used in the 1994 paper.

3.1 Work to include polymers should be categorised as follows:

(a) Regular single-strand organic polymers (essential)

(b) Copolymers (desirable)

(c) Irregular polymers (low priority; not for the next version)

3.2 Acceptable 2-D representations should show the constitutional repeating unit (CRU) with 'free valences' as bonds, enclosed (or crossed) by parentheses (), brackets [ ] or braces { }, followed by a letter subscript, e.g. $[-CH_2-CH=CH-CH_2-]_n$. Steric configuration should be included if known, but there is no requirement to recognise Fischer representations (e.g. 2-E2).

3.3 The InChI should be the same regardless of the direction of reading of the backbone chain, and regardless of where the CRU is considered to begin. A CRU beginning within a ring should not be allowed, i.e. the number of CRU free valences should be minimised. There is no requirement to handle 'ladder' polymers, e.g. 2-E-14 (where the CRU can only start within a ring).

For example $[-O-CH_2-CH_2-CH_2-C_6H_4-CH_2-CH_2-]_n$ can have 14 different representations, all of which should have the same InChI.

3.4 The molecular formula field could contain the formula of the CRU plus a polymer flag.

3.5 End groups should be included as a separate 'layer' if known; however the algorithm will need to deal with the fact that the end groups vary according to where the CRU starts. It would be necessary to relate end group citation to a specified CRU starting point.

3.6 There is no requirement to deal with inorganic polymers, e.g. 2-E17, 2-E18.

3.7 Alternating and periodic copolymers (4.1-E1 to E3) can be represented and treated like regular single-strand polymers.

3.8 Statistical (4.2-E1 to 7) and block (4.3-E1 to E6) copolymers should both be included: $[-A-]_x-[-B-]_y$ (statistical and random where $x + y = 1$; block where x and y are both > 1); perhaps in the former case they can be treated like mixtures?

3.9 Graft and star copolymers are not to be covered at this stage.

Alan McNaught
13 June 2005