



## IUPAC International Chemical Identifier (InChI) Subcommittee

### Minutes of the inaugural meeting, September 15th 2008, at the US National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

**Present:** *Subcommittee members:*

Steve Heller (Chairman) (NIST)  
Evan Bolton (US National Center for Biotechnology Information)  
Sandy Lawson (Elsevier, Frankfurt)  
Alan McNaught (InChI project coordinator, Cambridge, UK)  
Marc Nicklaus (US National Cancer Institute)  
Steve Stein (NIST)  
Dmitrii Tchekhovskoi (ex-officio developer) (NIST)  
Graeme Whitley (Wiley, New York)  
Jason Wilde (Nature, London)  
Tony Williams (ChemSpider)

*Observers:*

Igor Filippov (US National Cancer Institute)  
Peter Linstrom (NIST)  
Dave Martinsen (American Chemical Society)  
Marcus Sitzmann (US National Cancer Institute)  
Keith Taylor (Symyx Technologies, CA)

**Apologies:** *Subcommittee members:*

Colin Batchelor (Royal Society of Chemistry)  
Igor Pletnev (ex-officio developer) (Moscow State University)  
Chris Steinbeck (European Bioinformatics Institute)  
Andrey Yerin (Advanced Chemistry Development, Moscow)

Stefan Spiegel (Wiley-VCH, Weinheim) would attend for Graeme Whitley when meetings were in Europe.

Alan McNaught would act as secretary for the present meeting.

#### **1.0 Purpose and goal of the subcommittee**

The subcommittee had been set up under Division VIII (Chemical Nomenclature and Structure Representation) of the International Union of Pure and Applied Chemistry (IUPAC) in response to the perceived need for a formalised structure to respond to InChI user requirements in a businesslike and professional manner. It was intended that the subcommittee would oversee and control three principal aspects of the IUPAC InChI project: (1) maintenance, (2) direction for future InChI/InChIKey extension work, and (3) publicity/promotion to the scientific community. Thus the subcommittee will be concerned to set policy and priorities and oversee future InChI development. The membership consists of the InChI project team and representatives of organisations



actively using InChI and developing InChI applications in their products; the latter group will expand as more organisations adopt and use the InChI standard. In addition to publishers and data providers, it would be important to uncover InChI use in the pharmaceutical industry and include appropriate representation.

## 2.0 The standard InChI and standard InChIKey

2.1 Decisions were needed to enable launch of standard (i.e. fixed-option) InChI and InChIKey in response to user requests. Evan Bolton had tried to summarise requirements, pulling together pre-meeting e-mail discussions, and led a discussion of these. As a result, the following points were agreed:

[1] Standard InChI is for the purposes of interoperability/compatibility between large databases/web searching and information exchange.

[2] Standard InChI and non-standard InChI need to be distinguishable.

[3] Standard InChI is a stable identifier; however, periodic updates may be necessary and these should be reflected in the identifier version designation, which should be included in the InChI string. Furthermore, older versions of standard InChI should be archived and the ability to render InChI strings from all versions should be preserved.

[4] Standard InChI organometallic representation should not include bonds to metal for the time being (see minute 2.6).

[5] Standard InChI needs to distinguish between chemical substances at the level of same 'connectivity', same 'stereo', and 'isotopes', where:

connectivity means tautomer-invariant valence-bond connectivity (different tautomers give the same connectivity/hydrogen layer);

stereo means undefined/unknown-invariant stereo;

isotopes means same mass number (when specified)

[6] Standard InChIKey is computed only from a standard InChI.

[7] Standard InChIKey is for the principal purpose of a search-engine-style lookup of chemical information.

[8] Standard InChIKey and non-standard InChIKey need to be distinguishable at the level of the key itself.

[9] Standard InChIKey is a stable identifier; however, periodic updates may be necessary and these should be reflected in the version designation, which should be included in the InChIKey string.

[10] Any shortcomings in standard InChI/InChIKey may be addressed using non-standard InChI/InChIKey.

2.2 In the light of the agreed requirements listed in minute 2.1, the following options for standard InChI/InChIKey were agreed:

tautomerism:	FixedH	omit (i.e., turn mobile H perception on)
drawing style:	RecMet	omit (reconnection of metal)
	Newps	include (narrow end of wedge to stereocentre)



bug fixes:	Fb	include
	Fb2	include
	Fnud	include
stereo:	Suu	omit (designation of unknown/undefined stereo)
	SPXYZ	include (stereo at phosphorus)
	SAsXYZ	include (stereo at arsenic)
	Sabs/rel/rac	include abs but not rel/rac
new tautomerism:	Ket	omit
	15t	omit

**2.3** In the course of discussing ways of designating the standard forms the InChIKey format was revisited and the following points were agreed:

**2.3.1** The number of protons reinserted would not be encoded in the hash but would be indicated as a separate 2-character block at the end, where one character is a hyphen, as -N for neutral, -M for -1 hydrogen, -O for +1 hydrogen, etc.\*

**2.3.2** The checksum character was not very useful and would be removed.

**2.3.3** The version number would not be included in the flag character but would be indicated explicitly after it.

**2.3.4** It was suggested that the prefix 'InChIKey=' should be an integral part of communicating the presence of an InChIKey string in a document but not required for InChIKey lookup purposes. There was no decision on a precise location or syntax for 'InChIKey=' inclusion.

**2.4** It was then agreed that the standards should be designated as follows:

Standard InChI:

InChI=1S/..... (i.e. including 'S' after the version no.)

(standard InChI version numbers should always be whole numbers).

Standard InChIKey:

InChIKey=[14-block]-[8-block][flag][version]-[proton number]

- proton numbers will be conveyed using a case-insensitive alphabetic character where 'N' designates neutral, 'M' designates -1, 'O' designates +1, 'L' designates -2, 'P' designates +2, etc.\*

- version numbers start with case-insensitive 'A' for version 1, 'B' for version 2, etc.\*

---

*Secretary's note*

\* In subsequent e-mail discussion it was noted that the developers should feel free to consider alternative character mappings of the proton number, and of the version number, to achieve the same results.



- standard InChI will be distinguished from non-standard InChI by using a case insensitive flag character 'S'

**2.5** The foregoing decisions should enable Igor Pletnev to prepare InChI version 1.02(final), which should be circulated as a pre-release to the subcommittee. The new release should retain the Ket and 15t tautomerism options for evaluation by the community. It was suggested that the release version should be designated by the year label, i.e. version 1.08 for 2008, specifically to designate InChIs generated with such unsupported options. Another suggestion for denoting such strings was to replace "InChI=" with "XInChI=".

**2.6** With regard to item [4] of minute 2.1, Steve Heller would point out to IUPAC Division VIII that the recently published recommendations for graphical representation of chemical structures did not deal adequately with organometallic systems, and that the lack of standard representation principles in this area was a serious hindrance to the development of InChI to cover organometallic structures. Division VIII should address this as a matter of urgency.

### **3.0 Future InChI/InChIKey requirements**

The following were identified as areas requiring further InChI development:

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

Statements of preference by those present indicated that polymers, organometallics and extended stereo concepts were of roughly equal importance and should be given priority.

### **4.0 Funding**

The problems associated with funding further development were discussed. IUPAC was accustomed to provide travel and subsistence funds for people attending meetings but it was difficult to extract funds for maintenance and development. Steve Heller and Alan McNaught would continue to explore ways of doing this. NIST had made a major investment in InChI but could not be expected to continue funding at a similar level. However a proportion of Igor Pletnev's present NIST funding would be applied to InChI work, and application would be made to IUPAC for further supplementary funding when the current arrangement expires at the end of 2008. Steve Heller circulated a list of possible alternative funding sources that would also be explored. Meanwhile subcommittee members were asked to suggest other approaches to the problem. Steve Heller's hitherto very successful marketing efforts would continue with IUPAC funding if available.



**4.1** Sandy Lawson outlined a suggestion by a colleague that IUPAC should offer fee-based InChI certification to user organisations, in the form of a tag denoting InChI validation by IUPAC that certified users could employ for marketing purposes. Jason Wilde pointed out that this idea might be developed analogously to the CrossRef model as a lookup database whereby a charge could be made for deposition of InChIKeys. A sliding scale of charges could be developed in relation to ability to pay. It was agreed that Jason Wilde would consult with Sandy Lawson, Graeme Wiley and Tony Williams to produce a paper elaborating these ideas.

## **5.0 InChI/InChIKey lookup service**

See minute 4.1.

## **6.0 InChI/InChIKey-based reaction schema**

Steve Heller described progress with a project to develop a IUPAC computer-readable chemical reaction database standard based on InChI/InChIKey. A meeting of interested parties had been held in Berlin in May 2008; he would circulate the minutes to the present subcommittee. A first stage of this work, funded by the Royal Society of Chemistry and carried out at the Unilever Centre for Molecular Science Informatics in Cambridge, UK, would establish formats for data items, and a second stage, still to be defined and approved by IUPAC, would develop an XML-based schema.

## **7.0 InChI information on the web**

The 'unofficial' InChIFAQ prepared by Nick Day, though very valuable, was seriously out of date and ways of remedying this would be explored. Beda Kosata's inchi.info was less elaborate but somewhat more current. Tony Williams reported plans for ChemSpider to provide a Wikipedia services page.

## **8.0 Future meetings**

An InChI symposium was to be held at the ACS Salt Lake City meeting, on the morning of Sunday March 22 2009. The theme was to be InChI applications. Contributions from Evan Bolton, Graeme Wiley, Keith Taylor, Tony Williams and a Pfizer representative had been agreed; more were expected. It was agreed that the InChI subcommittee should meet in Salt Lake City on March 23rd (pm). Jason Wilde would investigate whether a CrossRef representative could attend. The following meeting would take place in Glasgow on July 30 2009, at the IUPAC General Assembly.

In view of the valuable e-mail discussions preceding the present meeting, it was felt that interim decisions might well be achieved by e-mail.

Alan McNaught

22 September 2008