chem-bla-ics

New InChI software beta: license issues resolved and InChIKey

Egon Willighagen

Published September 7, 2007

Citation

Willighagen, E. (2007). New InChI software beta: license issues resolved and InChIKey. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/cqv2x-hph36

Keywords

Inchi, Openscience

Abstract

The IUPAC/NIST team made a beta release of the next InChI software release:

Copyright

The IUPAC/NIST team made a beta release of the next InChI software release:

> *The principal new features of this release are:*
>
> 1. *A fixed-length (25-character) condensed digital representation of the Identifier to be known as InChIKey. In particular, this will:*
>   - *facilitate web searching, previously complicated by unpredictable breaking of InChI character*
>   - *strings by search engines*
>   - *allow development of a web-based InChI lookup service*
>   - *permit an InChI representation to be stored in fixed length fields*
>   - *make chemical structure database indexing easier*
>   - *allow verification of InChI strings after network transmission.*
> 2. *Restructured InChI-generating software that separates key steps in its creation from an input chemical structure file. Among other uses, this allows checking of intermediate results to enable easier testing and development of InChI-based applications.*
> 3. *Bug fixes designed to withstand malicious attempts to attack a Web server by providing a specially designed InChI string input to InChI binaries.*
>
> *We would welcome reports of your experiences with this new release and, of course, any problems.*

InChIKey

A had heard about the InChIKey extension earlier, and it solves the issue some people have with the InChI: it is too long. Well, molecules can have many atoms indeed. It is important to realize the InChIKey is not a replacement: it simply is not unique. The collision probability is calculated to be rather small, though. But clashes may occur, and sees from the above statistics quite likely for the number of molecules estimated to be drug-like, which is estimated at ~10^{60}. Moreover, these are theoretical probabilities which may not apply to the subset of molecules we actually tend to look at.

Anyway, the InChIKey is not a unique identifier, and never use it as such; that's what you need to remember.

An interesting feature is that addition of a check character, which enables some verification of typos. Nothing said about collision clashes there, which exist too. And the fixed length has its virtues too. That said, it certainly helps as sort of prefiltering. Google does a quite decent lookup of InChIs nowadays, and there is a growing amount of semantic markup of InChIs like use of microformats , as RDF/RDFa, stored in HTML @alt attributes, embedded in PNG images to address the issues of the InChI length.

Two final comments, and I hope Alan, Steve, Igor, Steve and Dmitrii will pick this up:

1. the InChIKey lost the version layer, which will cause trouble when the InChI moves to a next version (as in InChI=2/.... I would really like to see InChIKey=1/RYYVLZVUVIJVGH-UHFFFAOYAW as key instead.
2. an online service to validate the key using the check character would be most welcome

LGPL license

Not reported in the above announcement is the fact that this release also addresses a issue brought forward by the opensource community. License ambiguity has been addressed, and it is reported that the release now clearly states the LGPL license in the distribution as well as source code headers. This will make packaging for, for example, Linux distributions possible.

Modularization

One of the reasons why there has not been a Java port developed was the lack of modularization in the InChI software. This apparently has now been added, and I am very interested in reading about the effective modules available now. In particular, the canonicalization is interesting. The resulting atom ordering find its use in chemoinformatics algorithms, and a standard for that is most welcome.

Maybe now is the time to develop a Java version of the software.