IUPAC International Chemical Identifier (InChI)

Current status and future development in relation to IUPAC activities

The IUPAC International Chemical Identifier (InChI) is a non-proprietary machine-readable chemical structure representation format enabling electronic searching, and interlinking and combining of chemical information from different sources. It was developed from 2001 onwards at the US National Institute of Standards and Technology (NIST) under the auspices of IUPAC's Chemical Identifier project. Since 2009, the InChI Trust, a consortium of (mostly) publishers and software developers, has taken over responsibility for funding and oversight of InChI maintenance and development. Responsibility for scientific aspects of InChI development remains with the IUPAC Division VIII InChI Subcommittee.

The many potential uses of InChI may not be immediately apparent to many chemists, since it was developed to be handled by computer tools, not by human beings. Its role as a unique chemical identifier is analogous to the role of a barcode in general commerce – while not intended for reading or generation by humans, the identifier provides a unified electronic encoding for chemicals which can be built upon by various computer services to collect, search and exchange chemical information. Its chemical 'intelligence', single source and vendor neutrality allow us to consider InChI as a chemical barcode giving to every chemist instant and reliable access to the electronic world of chemistry. However the InChI is not supposed to be a replacement for established means of identification, like names or registration numbers, but a very powerful addition to the chemist's arsenal of weapons for dealing with chemical informatics.

The sections below deal with various aspects and needs of the InChI project and their relationship with IUPAC activities.

1. InChI as the most significant IUPAC initiative for e-chemistry

Current chemistry is now practically "paper free" and critically dependent on computer tools, from the planning of an experiment to the publication of the results, not to mention chemical calculations and modeling, which have significant practical importance but are performed exclusively in the virtual space of computer memory.

However IUPAC's activities in the area of e-chemistry have been extremely limited hitherto. In fact there is little else apart from the JCAMP-DX vendor-independent file format from the Committee on Printed and Electronic Publications (CPEP) Subcommittee on Spectroscopic Data Standards, developed to exchange spectroscopy and chromatography data, and some recommendations on spectroscopic data standards.

InChI is probably the most important and relevant IUPAC activity in relation to e-chemistry needs; its facilities are readily available via the desktops of virtually every chemist dealing with chemical structures. Hardly any other IUPAC activity is available for electronic use by the whole of the chemistry community.

IUPAC must not neglect computerized activities for all areas of chemistry; indeed proactive involvement in more e-chemistry projects is vital for the IUPAC activities to be seen as relevant to current needs. However, the current subject-based Divisional structure is not conducive to the development of such projects, and there is no budget assigned specifically to e-chemistry. Perhaps the best way forward would be to reorganize CPEP into an e-chemistry committee or even a division and provide it with a suitable project budget. Since IUPAC converted to self-publishing, CPEP's involvement with IUPAC publishing activities has been much reduced and the Committee has sometimes struggled to define an appropriate framework for its activities.

2. InChI as chemical informatics tool

Searching of existing publications and data is now impractical without the use of electronic databases. An efficient way of finding specific chemical information is vital both for science and industry. However, although ways to search publications by bibliographic and other textual data are straightforward and well implemented, searching data by chemical structure is so far not standardized and still significantly vendor-dependent.

InChI and especially its fixed length hash representation InChIKey provide unique tools for indexing and searching structure-related information suitable both for huge chemical databases and for specific scientific papers in electronic form. However, their introduction into the chemical literature is slow, perhaps inevitably. Both InChI and InChIKey can be used now to search over the internet but they still give fewer hits than searches by chemical name.

Nevertheless the take-up of InChI continues to grow and chemists are becoming increasingly aware of its advantages in facilitating dissemination, collection, indexing and retrieval of chemical information. In this way, InChI serves the whole chemical community and the InChI Project thus fits perfectly the aims of IUPAC.

3. InChI and structure representation

The InChI text string is a unique encoding of a chemical structure but being structure-based it is a more powerful representation than the structure itself. The reason is that the there are often several ways of graphically representing the same chemical substance. This often prevents recognition of the equivalence of different tautomers, mesomers and alternative protonation variants. The recognition of tautomerism and mesomerism is built into InChI algorithms allowing InChI to be a substance identifier rather than a structure identifier. Chemistry deals with substances; thus InChI is a more useful tool for identification of chemicals than chemical structure.

Correct operation of the InChI software is dependent on correct electronic structure input. While common organic compounds have well agreed representation conventions, there is still a lack of standards for representing many other classes of chemicals, especially for electronic representation. There are many available chemical drawing tools, ranging from desktop programs to web-based applets, but the tools and conventions are still vendor-dependent especially for classes of chemicals with poorly defined representation standards.

Further development and wider adoption of InChI as a vendor-neutral structure identifier will force chemical drawing software producers to conform to InChI procedures and unify their representation and encoding of chemical structures.

Thus, InChI needs and allows involvement of chemical software vendors in development of unified structure representation standards via the combined resources of Division VIII and InChI Trust. Any IUPAC structure representation recommendations need to take into account electronic representation. Special projects must be initiated to assure chemically intelligent representation conventions for all classes of chemicals, to allow correct and uniform coverage of all areas of chemistry in electronic and printed media. This is probably the most important task both for development of InChI and to make IUPAC chemistry computer-friendly.

4. InChI in relation to chemical nomenclature

Traditional chemical nomenclature is aimed at development of conventions for naming chemical substances in a human-friendly common language. Current chemistry often deals with molecules that are very complex and conventional nomenclature developments lead to long scarcely pronounceable names derived from large sets of rules. This complexity prevents most chemists from assigning chemical names manually and software tools are now very useful for

chemical name generation. Thus, even the development of classical nomenclature needs to take into account electronic representation and computer naming tools.

A consequence of this complexity is that such systematic names are no more human-friendly than the corresponding InChI text string generated from the chemical structure. From this point of view, InChI may be considered as a special kind of nomenclature allowing explicit definition of a chemical compound. A very important aspect is that traditional chemical nomenclature is largely structure-dependent and inherits most of the limitations of structure discrepancies mentioned above. InChI still has some areas, for example complex tautomers, that need further development but being significantly substance-aimed, InChI by design does not have such limitations. The involvement of Division VIII is highly desirable for development of principles to deal with multiple structure representations. At the same time InChI concepts can be useful for development of nomenclature for chemical structures represented in delocalized form.

InChI text strings can hardly be treated as a replacement for conventional chemical names but they are far more suitable for identification of chemicals in various media. Any chemist with any level of nomenclature knowledge and any chemical software available will be able to generate the same unique InChI text string for the same chemical structure. InChI can be treated as an additional type of nomenclature and needs no less attention from IUPAC than traditional nomenclature.

5. Current state and further development of InChI

The InChI project is now quite mature and provides a well developed set of tools adopted by many open access and commercial chemical media. Most prominent examples include the Royal Society of Chemistry (ChemSpider), Chemical Abstracts Service (SciFinder) and the US National Institutes of Health (PubChem). Existing services allow retrieval of data for chemical compounds including many experimental and predicted properties, environmental and biomedical data, spectra, and links to patents, articles and other databases. Facilities for InChI generation are already present in practically all chemical drawing programs allowing chemists to use various InChI-based tools.

Although well developed and tested for organic structures, InChI still needs further development to cover inorganic, biochemical and polymer structures, and advanced stereochemical and tautomeric features. Each area needs first of all the development of representation conventions with essential involvement of IUPAC resources to ensure the choice of chemically intelligent and widely acceptable conventions. This will define requirements for further development of InChI tools.

The degree of InChI acceptance critically depends on rapid extension of coverage to all classes of chemicals and integration of InChI tools into a wide variety of commercial and open-access chemistry-oriented services.

6. InChI support by industry and public organizations

Most chemical drawing programs already include InChI procedures thus making InChI tools available virtually to every chemist. However industrial chemistry still awaits better support from InChI since industry often deals with impure substances and mixtures that lack agreed representation and encoding conventions.

An important industrial application concerns patents that deal with the special structure representation commonly referred to as Markush. The general principles for InChI support of Markush structures are already developed within the corresponding IUPAC project and await funding for implementation. It should be noted that there are currently no IUPAC recommendations on Markush representation. The support of Markush representation by InChI tools will significantly improve the applicability of InChI in patents.

It is clear that InChI is ideally suited for application in substance registration and regulations. Currently most official registration systems include support of chemical structures in addition to textual data such as registration numbers and chemical names. Being free from the shortcomings of chemical structure representations and aimed at substances InChI will allow all such systems to be made more chemically intelligent. Several government organizations already use InChI tools internally and some plan to make InChI and InChIKey a necessary part of their registration systems.

While the development of InChI tools now involves only IUPAC Division VIII resources, the acceptance of InChI by industry and government authorities needs involvement of other IUPAC bodies, primarily COCI and CPEP (see also section 1.0).

7. Extended IUPAC involvement in the InChI project

The development of InChI projects is covered by the InChI Trust and IUPAC Division VIII funds assigned to projects having general chemical importance. Wide acceptance of InChI needs better involvement of IUPAC resources including support by other IUPAC bodies and international organizations.

The table below lists the most important recent projects and several areas that need to be considered.

Area of chemistry or activity	IUPAC and other organizations	Status of the project
Recent and formulated projects		
Extended stereochemistry – chirality axis, plane, and helicity, high coordination stereo	Division VIII	Expected
InChI Requirements for Representation of Organometallic and Coordination structures	Division VIII	Expected
InChI Requirements for Representation and Encoding of Electronic States	Division VIII	Expected
InChI Requirements for Representation of Materials	Division VIII	Expected
Extended treatment of tautomerism by InChI tools	Division VIII	Submitted
InChI Requirements for Representation of Organometallic and Coordination structures	Division VIII	In progress
Standard InChI-based Representation of Chemical Reactions	Division VIII	In progress
InChI Requirements for Representation of Polymers	Division VIII	Completed Awaits implementation
InChI Requirements for Representation of Markush Structures	Division VIII	Completed Awaits implementation
Possible further projects and activities to extend InChI acceptance		
Investigation of possible extension to biochemicals	Division VIII, IUBMB	To consider
Representation conventions and encoding of industrial chemicals	Division VIII, COCI	To consider
Promotion of InChI for regulations and substance registration	COCI, CPEP	To consider

8. Proposed actions

It is clear that InChI and InChIKey have already become an important part of chemical space and serve well the needs of the chemical community. Any doubts about their usefulness were resolved long ago. However, the current implementation needs extension to other classes of chemicals and more active promotion of InChI to assure its position as the universal identifier for all chemicals dealt with by all types of scientific and industrial enterprise.

To extend IUPAC activities to fulfill the needs of e-chemistry and to ensure further successful development of InChI the following actions should be considered by the appropriate IUPAC bodies:

1. Recognize and support the InChI project as the most significant IUPAC activity in the area of e-chemistry. (All IUPAC bodies)
2. Take into account computer representation and treatment of chemical data in the development of all IUPAC recommendations. (All Divisions, Committees and especially ICTNS)
3. Reorganize CPEP into a Committee or Division aimed at fulfilling the needs of electronic chemistry and chemical informatics. (CPEP initially, then higher authorities)
4. Ensure involvement of the Divisions and the InChI Subcommittee in extension of InChI to cover additional classes of chemicals. (Divisions II, III, IV, and VIII)
5. Investigate the possibility to extend InChI to cover biochemicals, including peptide and nucleic acid sequences. (Divisions VIII and III, JCBN and IUBMB)
6. Promotion of InChI and InChIKey to industry and regulatory authorities as identifiers for regulation and registration of chemical compounds (CPEP, COCI)

Andrey Yerin
Titular Member, IUPAC Division VIII
Member, IUPAC Division VIII InChI Subcommittee

Alan McNaught
Secretary, IUPAC Division VIII InChI Subcommittee
Secretary, InChI Trust

Stephen Heller
Chairman, IUPAC Division VIII InChI Subcommittee
Project Director, InChI Trust