InChISub/0312/minutes

IUPAC Division (VIII) of Chemical Nomenclature and Structure Representation
International Chemical Identifier (InChI) Subcommittee

Minutes of the meeting on 21-22 March 2012 at the Hilton Hotel, Gaithersburg, USA

Present: *Subcommittee members:*
Steve Heller (Chairman, National Institute of Standards and Technology, USA)
Steve Bachrach (Trinity University, San Antonio, USA)
Colin Batchelor (Royal Society of Chemistry, UK) (in part; via Skype)
Evan Bolton (National Center for Biotechnology Information, USA)
Jonathan Goodman (Unilever Centre for Molecular Informatics, Cambridge, UK) (in part; via Skype)
Günther Grethe (Alameda, USA)
Alan McNaught (Secretary, Cambridge, UK)
Marc Nicklaus (Frederick National Laboratory of Cancer Research, USA)
Igor Pletnev (ex-officio developer) (Moscow State University, Russia)
Hinnerk Rey (Elsevier, Frankfurt, Germany)
Keith Taylor (Accelrys, San Diego, USA)
Dmitrii Tchekhovskoi (National Institute of Standards and Technology, USA)
Tony Williams (RSC ChemSpider, UK)
Andrey Yerin (Advanced Chemistry Development, Moscow, Russia)

Observers:
John Barnard (Digital Chemistry Ltd) (in part; via Skype)
Yulia Borodina (Food and Drug Administration, USA)
Don Burgess (National Institute of Standards and Technology, USA) (in part)
Lawrence Callahan (Food and Drug Administration, USA) (in part)
Dusanka Janezic (National Institute of Chemistry, Ljubljana, Slovenia) (in part)
Peter Linstrom (National Institute of Standards and Technology, USA)
Daniel Lowe (NextMove Software, Cambridge, UK)
Dave Martinsen (American Chemical Society, USA) (in part)
Matej Penca (National Institute of Chemistry, Ljubljana, Slovenia) (in part)
Tyler Peryea (Food and Drug Administration, USA) (in part)
Roger Sayle (NextMove Software, Cambridge, UK)
Markus Sitzmann (Frederick National Laboratory of Cancer Research, USA)
Bill Wallace (National Institute of Standards and Technology, USA) (in part)

Apologies: *Subcommittee members:*
Nicko Goncharoff (SureChem, UK)
Steve Stein (National Institute of Standards and Technology, USA)
Chris Steinbeck (European Bioinformatics Institute, Hinxton, UK)
Ted Wilks (Hockessin, USA)

1.0 History of the InChI project 1999-

Alan McNaught summarised the history of the project (see Attachment A)

2.0 InChI Subcommittee history and status

Alan McNaught outlined the evolution of the subcommittee and its development as the scientific advisory board for the InChI project (see Attachment B). He noted the problems encountered in finding a suitable home within IUPAC for chemoinformatics projects: it was difficult for the structure of such a cumbersome organisation to adjust quickly to changes in the world of chemistry. The possibility of setting up a body (a new Division or perhaps a joint subcommittee of Division VIII and the Committee on Printed and Electronic Publications), with associated project funds, should be considered seriously.

3.0 InChI Trust history and status

Steve Heller described the background to the formation of the Trust, set up in 2009 to provide a stable environment for management and development of the InChI standard. The Trust had been responsible for providing two updates (1.03 and 1.04) to the InChI/InChIKey software, source code documentation, and a certification suite for validation of the software installation. Requirements defined by the working groups of the IUPAC InChI subcommittee were to be applied to the further development of the software. Programming work on polymers and mixtures was about to begin, and work on extension to Markush structures would proceed when funding was available.

3.1 Lack of sufficient funding had meant that developments had not proceeded as quickly as the Trust would have liked, and the possibility of charging non-members for access to InChI extensions was being considered. It was recommended that the Trust should develop a donation model for organisations not wishing to become members or associates.

3.2 The importance of raising the profile of InChI was emphasised. The Trust was looking for ways of doing this, and would like good examples of the utility of InChI for use in publicity.

4.0 InChI subcommittee working groups

4.1 InChIKey resolver

A report from Tony Williams and Markus Sitzmann is included as Attachment C. Markus had taken over leadership of the group. He would consider introducing coverage of some non-Standard InChIKey options, with inclusion of the Standard InChIKey in the response. About 120 million InChIKeys were now available for lookup from the NCI resolver. It was intended to develop a federated resolver portal listing all resolvers using the IUPAC InChI protocol. The Standard resolver would return both a Standard InChI and its source.

4.2 InChI for polymers and mixtures

A report from Andrey Yerin is included as Attachment D. The working group had developed recommendations for both source- and structure-based InChIs. It was thought that the source-based model would be the more valuable. The input for both would be molfile initially. The need for specification of conditions of polymerisation and proportions of reactants was discussed: it was thought that the addition of such data would be left to the user. The possibility of including information relating to multiple attachment points would be considered after the development of a simple source-based InChI. John Barnard pointed out the need for polymer and Markush extensions to employ a mutually consistent approach.

It might be necessary/helpful to give the output from the polymer software a specific designation such as PInChI.

4.3 RInChI (InChI for reactions)

A report from Jonathan Goodman is attached (Attachment E); a full write up of the work in Cambridge would be available soon. Günther Grethe would ask the working group to evaluate and analyse the results obtained so far. Progress with routines for reaction analysis was good. Progress had been made also with studies on both long and short forms of RInChIKeys: the shorter version may be preferable for most purposes, but RInChIKeys would not be publicised until their usefulness had been investigated further. Then the proposed RInChI/RInChIKey utilities would be publicised to database owners and InChI Trust members and associates. Günther intended to write a descriptive paper for publication.

4.4 InChI for Markush structures

The proposal from Digital Chemistry for implementation of the working group proposals is included as Attachment F. This would proceed when funding was available. The first step would be the decoupling of internal and external structure representations in the API, and it was agreed that this would be beneficial not only for the Markush extension but also for maintenance and other future extensions of the InChI software. The certification suite would enable a check that no changes to version 1 InChIs had been introduced. It was recommended that the Trust proceed with the API modification without waiting for funding for the rest of the project .

4.5 InChI for electronic states

Don Burgess outlined some desirable extensions to InChI, covering sp^3 conformers, resonance structures, and elementary reactions. It was agreed that he should discuss these matters further with Dmitrii Tchekhovskoi and then put together a draft IUPAC project proposal (or proposals) for preliminary consideration by the subcommittee officers.

4.6 InChI for organometallics

Colin Batchelor reported that the working group was waiting for the release by Accelrys of the new structure representations corresponding to molfile V3000. Keith Taylor said that these would be available in the new version of the Accelrys drawing package, expected in about one month. The working group would then consider a range of options for defining InChI for organometallics. If it was decided to proceed on the basis of V3000 this would need to be incorporated in the InChI software.

Dmitrii Tchekhovskoi pointed out that it would be possible to proceed on the basis of V2000, without atom coordinates, using atoms and connections only. V3000 capability could be added when generally available through several drawing packages. However, it would be necessary to define an InChI-compliant drawing protocol. Also it would be essential for the InChI software to continue to interpret legacy representations as well as the new V3000 type.

It seemed inevitable that introduction of new organometallic capabilities would require specification of output as InChI version 2. It was important to ensure that the proposed API changes (minute 4.4) could handle organometallics as well.

The subcommittee wished to remind the organometallic working group that they should concentrate on requirements rather than methods of implemention.

4.7 InChI for inorganics

It was agreed that Hinnerk Rey would take over leadership of this group from Nigel Wheatley. There was a draft proposal to IUPAC that should be developed further.

4.8 Biopolymers etc

Development of a project was probably premature. To deal with these systems might require a product independent of the established InChI software.

4.9 Interlocking structures (e.g. rotaxanes)

Andrey Yerin said that he would keep this area in mind while proceeding with work on rotaxane chemistry for Division VIII.

4.10 Extended stereo concepts

It was agreed to form a working group (Evan Bolton, Andrey Yerin, Marc Nicklaus, Markus Sitzmann) to consider what improvements to stereochemical representation could be made easily. In particular, the group should try to identify stereochemical problems that can be solved by business rules to be included in the technical manual.

5.0 Teaching and training requirements

5.1 The development by Bill Armstrong at Louisiana State University of written and video materials aimed at faculty and students was in progress.

5.2 Contributors to the forthcoming InChI Symposium in San Diego were being invited to submit their presentations to a special issue of *J. Chem. Informatics*.

5.3 A revised and extended InChI FAQ prepared by Igor Pletnev was nearly ready for release.

5.4 Subcommittee members and observers were asked to send details of any interesting uses of InChI to Steve Heller for use in InChI publicity.

6.0 InChI version 1.04

The recent maintenance release of version 1.04 was noted. This release was accompanied by a more permissive software licence. Support for Chemical Markup Language had been removed (as part of an exercise to ensure InChI Trust ownership of all aspects of the InChI code), support for elements to 112 had been introduced, and the software had been modified to accept multiple input files.

7.0 The future

Dmitrii Tchekhovskoi presented an account of various possible improvements to InChI that could be introduced at the same time as an upgrade to version 2 (see Attachment G). These could not be introduced earlier since they would result in many changes to the version 1 InChI strings. It was agreed that a working group led by Marc Nicklaus (Dmitrii Tchekhovskoi, Markus Sitzmann, Igor Pletnev, Evan Bolton, Andrey Yerin, Tyler Peryea, Hinnerk Rey, Marc Nicklaus) should be asked to recommend desirable changes coincident with establishment of version 2. In particular they should discuss alternative possible approaches to tautomerism.

It was agreed that in the event of a version 2 being established, version 1 should be retained alongside it.

It seemed likely that upgrade to version 2 would be required when the organometallic extension was developed. The Markush extension would not require version change, but might carry a designation such as 1.04M.

The possibility of introducing ‘error’ InChI (InChI=1//error) should be considered.

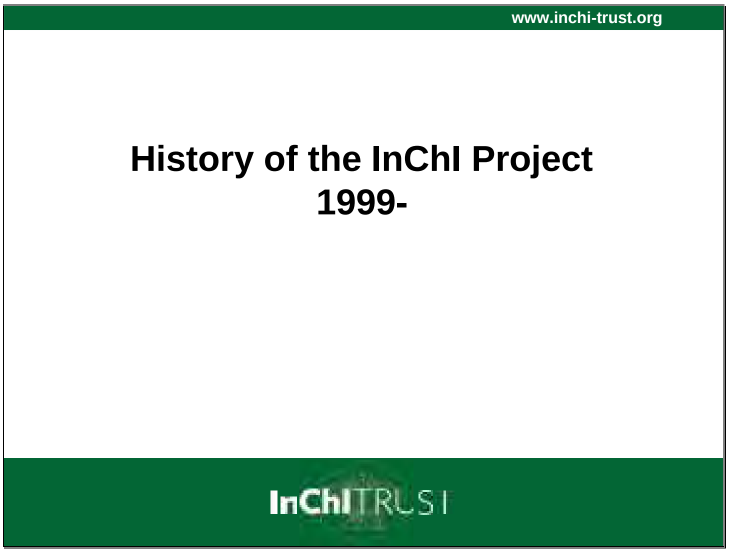
It was agreed that there was insufficient justification for change to the InChIKey format,

8.0 Next meeting

A further meeting would probably be arranged in 2014-2015.

Alan McNaught
22 April 2012

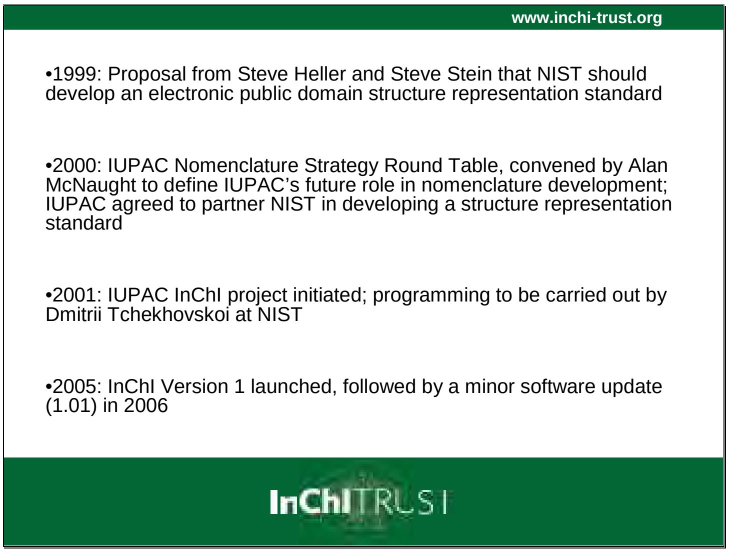
Slide 1



www.inchi-trust.org

History of the InChI Project 1999-

Slide 2



www.inchi-trust.org

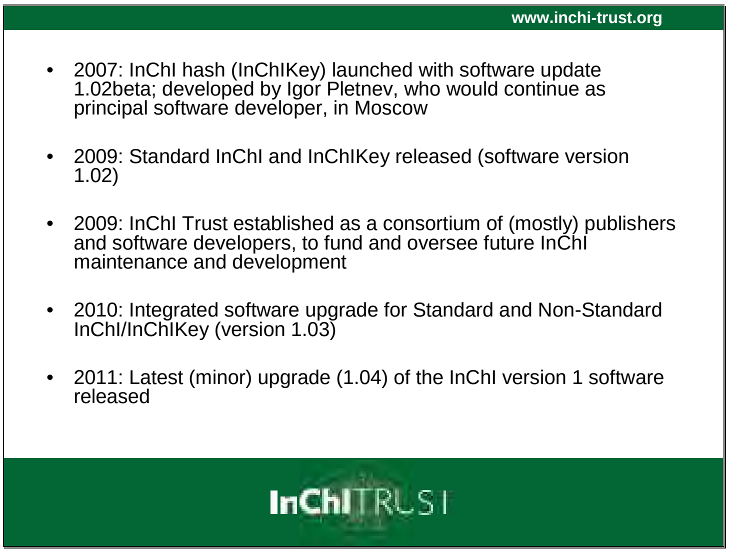
•1999: Proposal from Steve Heller and Steve Stein that NIST should develop an electronic public domain structure representation standard

•2000: IUPAC Nomenclature Strategy Round Table, convened by Alan McNaught to define IUPAC’s future role in nomenclature development; IUPAC agreed to partner NIST in developing a structure representation standard

•2001: IUPAC InChI project initiated; programming to be carried out by Dmitrii Tchekhovskoi at NIST

•2005: InChI Version 1 launched, followed by a minor software update (1.01) in 2006

Slide 3



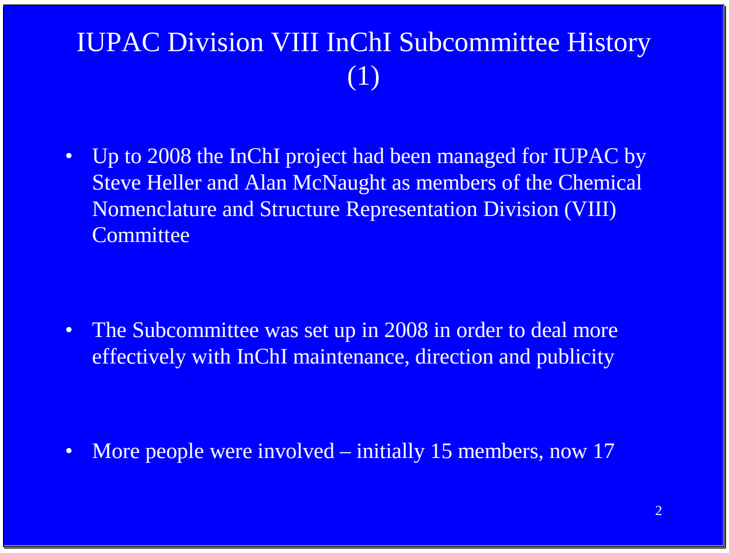
www.inchi-trust.org

- 2007: InChI hash (InChIKey) launched with software update 1.02beta; developed by Igor Pletnev, who would continue as principal software developer, in Moscow
- 2009: Standard InChI and InChIKey released (software version 1.02)
- 2009: InChI Trust established as a consortium of (mostly) publishers and software developers, to fund and oversee future InChI maintenance and development
- 2010: Integrated software upgrade for Standard and Non-Standard InChI/InChIKey (version 1.03)
- 2011: Latest (minor) upgrade (1.04) of the InChI version 1 software released

Slide 1

IUPAC
Division VIII
InChI Subcommittee

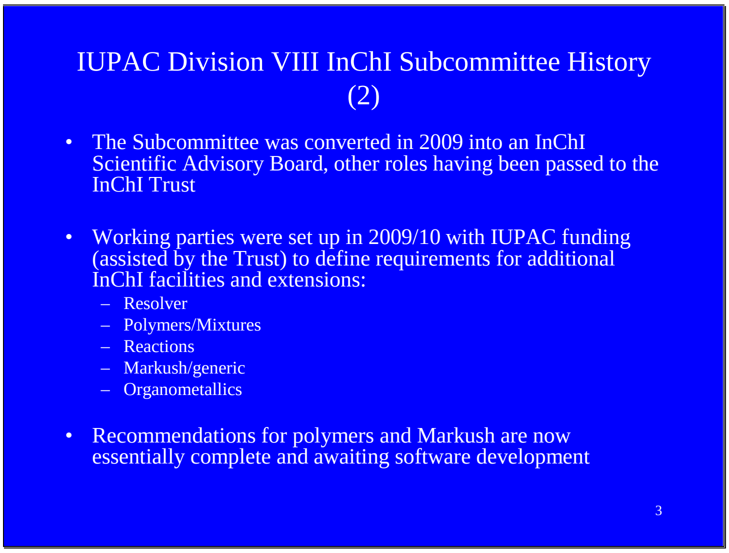
Slide 2



IUPAC Division VIII InChI Subcommittee History (1)

- Up to 2008 the InChI project had been managed for IUPAC by Steve Heller and Alan McNaught as members of the Chemical Nomenclature and Structure Representation Division (VIII) Committee
- The Subcommittee was set up in 2008 in order to deal more effectively with InChI maintenance, direction and publicity
- More people were involved – initially 15 members, now 17

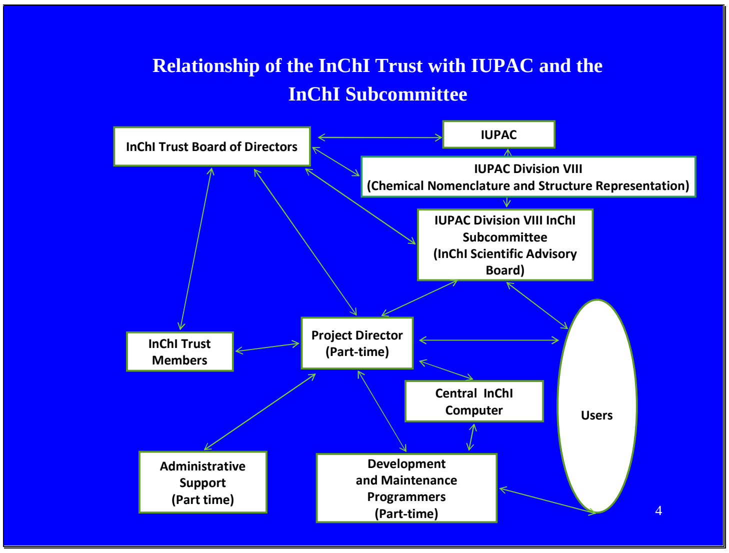
Slide 3



IUPAC Division VIII InChI Subcommittee History (2)

- The Subcommittee was converted in 2009 into an InChI Scientific Advisory Board, other roles having been passed to the InChI Trust
- Working parties were set up in 2009/10 with IUPAC funding (assisted by the Trust) to define requirements for additional InChI facilities and extensions:
  - Resolver
  - Polymers/Mixtures
  - Reactions
  - Markush/generic
  - Organometallics
- Recommendations for polymers and Markush are now essentially complete and awaiting software development

Slide 4



Relationship of the InChI Trust with IUPAC and the InChI Subcommittee

InChI Trust Board of Directors

IUPAC

IUPAC Division VIII
(Chemical Nomenclature and Structure Representation)

IUPAC Division VIII InChI Subcommittee
(InChI Scientific Advisory Board)

InChI Trust Members

Project Director
(Part-time)

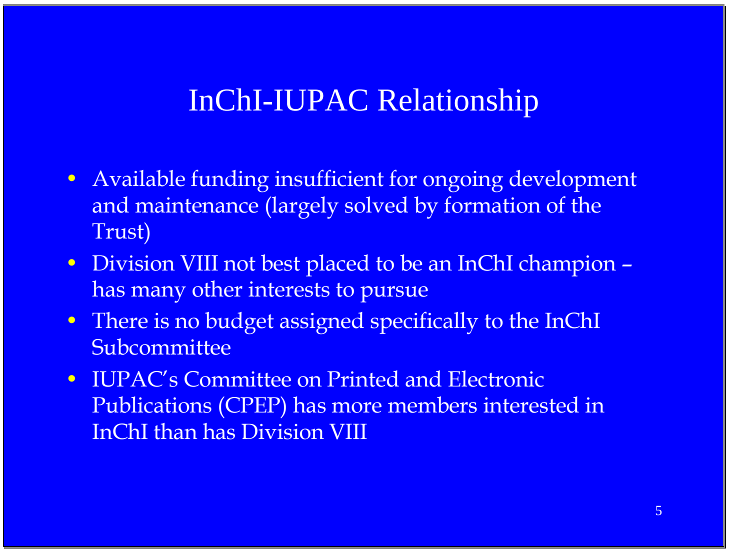
Central InChI Computer

Users

Administrative Support
(Part time)

Development and Maintenance Programmers
(Part-time)

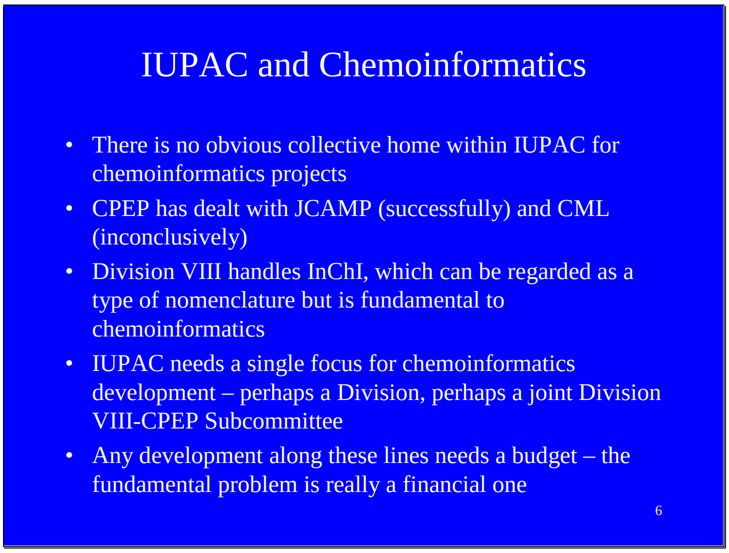
Slide 5



InChI-IUPAC Relationship

- Available funding insufficient for ongoing development and maintenance (largely solved by formation of the Trust)
- Division VIII not best placed to be an InChI champion – has many other interests to pursue
- There is no budget assigned specifically to the InChI Subcommittee
- IUPAC's Committee on Printed and Electronic Publications (CPEP) has more members interested in InChI than has Division VIII

Slide 6



IUPAC and Chemoinformatics

- There is no obvious collective home within IUPAC for chemoinformatics projects
- CPEP has dealt with JCAMP (successfully) and CML (inconclusively)
- Division VIII handles InChI, which can be regarded as a type of nomenclature but is fundamental to chemoinformatics
- IUPAC needs a single focus for chemoinformatics development – perhaps a Division, perhaps a joint Division VIII-CPEP Subcommittee
- Any development along these lines needs a budget – the fundamental problem is really a financial one

InChI Resolver Protocol

Task group progress report

March 2012

Participants

Antony Williams, Royal Society of Chemistry, Cheminformatics Group
Valery Tkachenko, Royal Society of Chemistry, Cheminformatics Group
Markus Sitzmann, National Cancer Institute
Marc Nicklaus, National Cancer Institute

Summary

The Federated InChI Resolver group was charged with the role of identifying a path by which the hosts of databases or repositories containing InChIKeys could be queried, using a standard protocol adapted by all systems, to allow for federated look-up for presence of a compound.

Members of the Royal Society of Chemistry and the National Cancer Institute have met to discuss approaches to producing a standard protocol for review, testing and adoption by the community. Both NCI and RSC presently have forms of "resolvers" which include InChIKey lookup and retrieval but, due to resource limitations, no progress has been made on defining a standard protocol. RSC has chosen to pass leadership of the project over to NCI, specifically under the guidance of Markus Sitzmann, but will remain engaged as a participant, supporter and adopter of a federated resolver approach and will dedicate resources, as appropriate, to demonstrate federation between our databases.

The NCI/CADD group is working on a substantial update of their local InChIKey Resolver (Chemical Identifier Resolver) database which currently allows the structure lookup for approx. 110 million InChIKeys.

March 5th, 2012

Antony Williams,	Markus Sitzmann
ChemSpider, RSC	National Cancer Institute (CADD Group)
Williamsa@rsc.org	sitzmann@helix.nih.gov

InChI Requirements for Representation of Polymers and Mixtures

IUPAC project 2009-042-1-800

Task group report

Task group members

Andrey Yerin, ACD/Labs, Russia, project chairman;
Ted Wilks, DuPont contractor, USA;
Jaroslav Kahovec, Institute of Macromolecular Chemistry, Czech Republic;
Roger Schenck, Chemical Abstracts Service, USA;
Dmitrii Tchekhovskoi, National Institute of Standards and Technology, USA

Experts involved in discussions:

Igor Pletnev, Lomonosov Moscow State University, Russia, InChI developer;
Keith Taylor, Accelrys, Inc., USA
Jonathan Brecher, PerkinElmer, Inc., USA;
Yulia Borodina, FDA, USA

Current state of the project:

The project can be considered as complete. The final report is submitted to InChI Trust on January 10th 2012.

Short summary of results

Structure-based representation and encoding of polymers

1. The electronic structure-based representation of homopolymers is well established and can be created in various computer drawing programs. The structure-based representation of polymers is based on structure of constitutional repeating units (CRUs) enclosed in polymer brackets with possible indication of end-groups.
2. InChI encoding of structure-based representation needs development of internal representation for such structures and the corresponding canonicalization procedures.
3. InChI encoding of structure-based polymers must be based on canonical CRUs. It would be desirable but not obligatory if the expected InChI canonicalization of CRU will follow at least basic IUPAC criteria to choose the preferred CRU.
4. The canonicalization of CRU by InChI procedures needs the corresponding accommodation of end-groups to maintain a correct constitution of a polymer.
5. Copolymers are defined by a set of CRUs enclosed in polymer brackets with specification of a type of copolymer and roles of the specific components. There is no general agreement about representation of copolymers.
6. InChI encoding of copolymers needs some way to store the information about the type of polymer and roles of components and can be achieved by introduction of a special “modification” layer.

Source-based representation and encoding of polymers

1. We propose to introduce and support source-based representation of polymers based on chemical structure of starting material with a special indication that the structure represents a polymer. Having general importance and strictly corresponding to source-based nomenclature of polymers this representation is especially useful for the polymers that are difficult or impossible to represent in a structure-based way.
2. Non-specific polymer nature is indicated with traditional enclosing marks that can be any with the preference for most often used brackets and one letter index.

3. The polymer components having a special role in a polymer (branching, cross-linking, modification) can be included in polymer brackets with an indication of their special function.
4. Source-based representation of copolymers - block, random, alternating and graft – is defined with the corresponding indication used instead of or in addition to one letter index used for unspecified polymer.
5. InChI encoding of source-based representation of polymers can be based on general InChI encoding with introduction of a special “modification” layer used to specify polymer nature, type of a polymer and the role and order of the components where needed.
6. The proposed introduction of modification layer can be used for specification of other specific modifications or nature for the set of chemical structures.

Relation between source- and structure-based InChI encoding

1. Source- and structure-based representations and their InChI encodings are independent and no procedures are implied for algorithmic conversion and relation from one type to the other.
2. The relation between source- and structure-based InChIs and InChIKeys can be established via the procedures close to InChI resolver protocol developed for discrete structures. For polymers this task is simple due to significantly smaller number of polymers compared to the number of discrete structures.

Encoding of mixtures

1. Chemical mixtures are uncommon objects both for printed and electronic media and miss agreed representation conventions. The representation of mixtures is possible in some drawing programs in a way close to source-based representation of polymers with a special mixture designation.
2. The encoding of mixtures can be based on general InChI encoding with an indication of mixture attributes inside the modification layer proposed above. The alternative encoding of mixtures can be based on a set of specific InChIs for each mixture component.
3. The task group has not come to general agreement about the encoding and computer representation of mixtures and we propose to make the decision after consulting with other InChI task groups, especially those that foresee InChI encoding of specific types of chemical objects via specification of a set of separate InChI codes for the components.

We think that the developed principles will help to unify ways of electronic representation and allow implementation of InChI encoding for polymers. The task group members can be contacted via the development of the corresponding InChI procedures for detailed specifications and examples of polymers.

March 16th 2012

Andrey Yerin,
Task group chairman
ACD/Labs, Russia
erin@acdlabs.ru

Reaction InChIs

`http://www-rinchi.ch.cam.ac.uk/`
March 2012
Jonathan M Goodman

Chad Allen, a final year undergraduate student at the University of Cambridge, has been working on Reaction InchI development as his research project. Work on the project continued until Friday March 16th, 2012, and the project write-up is currently being completed.

The work that he has done is now available on the website:
`http://www-rinchi.ch.cam.ac.uk/`

Improved Information Extraction

The main development has been the extraction of more information from RDFiles. In August 2011, the analysis of the database showed that a large number of reactions were duplicates. The reason for this was that the reactions had been recorded with different solvents or other conditions. As a result, the reactants and products were the same but the detailed procedure was different. This information was recorded in the RDFile, but was not extracted into the RInChI.

Chad Allen has improved the information extraction program, so that much more of this information is now transferred to the RInChI, and the small differences between similar reactions are reflected in the RInChI. There are now a hundred more unique RInChIs from the database, as a result of this increased differentiation. The code, which is written in Python, also runs much faster than before.

Reaction Analysis

Chad has also written routines that look for changes in stereochemistry and rings. This appears to be working effectively, and some examples of their use should be available soon.

RInChI-key

A trial process has been developed to generate RInChI-keys, and examples are available on the website:
`http://www-rinchi.ch.cam.ac.uk/database.html`
Two versions of the key have been tried. The long key encodes all of the molecules in the reaction as separate InChI-keys. The length of the key, therefore, increases with the number of molecules in the reaction. The short key encodes the starting materials, products and catalysts in separate blocks, with connectivity and stereochemistry held separately. The key is the same length for all reactions.

Conclusions

These three areas develop the future work that was discussed in August. The programs that Chad has developed appear to work effectively, and now they need to be tested using the existing database, and any additional data that may become available.

Jonathan Goodman
March 2012