InChI Trust Project Director's Newsletter Report

September 2015 (09/6/15)

Summary:

Since the January 2015 report there continues to be good progress with InChI and the InChI Trust in a number of areas. I received the Patterson-Crane award for 2015 and will go to Columbus Ohio in September to accept the award and give an InChI presentation. The detailed technical paper on the InChI algorithm was finally published in the Journal of Chemoinformatics. Bio-Rad has offered to rewrite parts of the InChI code to allow for threading (essentially parallel processing) so the algorithm would run faster. The US Patent Office, with help from the US NTIS (National Technical Information Service) asked the Trust to participate in a workshop/seminar about InChI for Markush. A new InChI working group on chemical mixture composition, which is related to lab safety, has been initiated. Related to the matter of the working groups, the main issue that, as usual, needs to be improved on is having the working groups be more active in moving towards their goals and getting more organizations, databases, and publications to use the InChI algorithm.

Items covered in this report:

Membership/Support:

Summary: A number of organizations are still in the process of joining or talking about joining, but there still is little progress in the past 6 months. In most organizations, since InChI works and it is not high on their immediate priority lists, actual real progress is slow without a dedicated champion within an organization.

As of September 1, 2015

Existing Members and Associates: 14

Supporters: remains at 48

InChI RFP/Contracts

The contract for Markush structures remains on hold awaiting potential funding. There is hope that July 2015 US Patent Office workshop/seminar on InChI for Markush structures will enable this to move forward.

The contract for taking forward the RInChI work that Jonathan Goodman and Chad Allen did at Cambridge University with Dr. Gerd Blanke (Germany) is progressing well. A RInChI working party meeting will be held during the Fall Boston ACS meeting to assess the progress and plan for future needs.

InChI development work

Igor Pletnev continues to do a superb and a very responsive job as the InChI programmer. What follows is a summary of his recent work:

"The new API has recently been analyzed and tested by Paul Thiessen of the NCBI/NLM/NIH. His uncensored comments are:

I really like this API. It's clean and straightforward. In retrospect, simpler to learn and use than the "native" API. I wrote an interface that directly translates PubChem's native format (a "PC-Compound" class object) to/from the IXA "molecule" – rather than, say, going through MOL/SDF. (Mostly I could copy this from the same interface I built for the regular InChI 1.04 stereo0d api.) I have run InChI generation, and conversely, generation of structure from InChI, across all of PubChem Compound. I did "round trips" – generating structure from InChI, then (re)generating InChI from that structure, to compare to the original InChI (and structure). I did not have any stability problems; no crashes/segfaults or anything like that.

Had to do some tweaking to get it to compile with GCC (4.8.1) on linux. Mostly type casts that GCC is kinda picky with. Had to replace memicmp with strncasecmp. Had to do a few small tweaks to compile all this within a unique C++ namespace – which I do so that, for example, I can build an application that uses both the IXA and the standard InChI 1.04 libraries simultaneously. (Which I needed to do to compare the two APIs side-by-side across all of PubChem.) Note that I do not have a standalone unix makefile for this library; the way I use it, it is compiled as a library into our NCBI C++ toolkit, and hence the Makefile configuration is embedded in that system, not useful outside that context. I did not compile a standalone application like inchi-1.

Fixed a substantial memory leak, by adding "FreeINCHI(&output);" to the end of BUILDER_Update() in ixa_builder.c.

It would be convenient to have MOL_GetBondOtherAtom() exposed in the public API.

Where I had to spend the most time was in figuring out how IXA's "stereo descriptors" work. The documentation on this was basically non-existent (unless I missed it, always possible). In particular, the array of atoms used as "vertices" in IXA are not (in general) the same as the atoms used as "neighbors" in the original library's stereo0D api. I actually had to read IXA code to figure out how these vertices are ordered and which atoms around the stereocenter they represent. So, it would of course be good to include in the documentation a complete discussion of how to create and read these stereo descriptors, and what it expects for each vertex atom. Graphical illustrations of how these vertices are arranged around the stereo atom/bond would be extremely helpful, rather than just text description.

When scanning across the entire PubChem compound database, results were basically identical for IXA vs 1.04 stereo0D.

There were a tiny number of cases (<1000 out of all 68M compounds) where IXA actually did better than 1.04, in generating structure from InChI. Mostly seemed to be cases of double bonds in a large (>6 atom) ring, e.g. CID 10059912. However, I can't swear for sure that this is a bug in the InChI 1.04 library, as opposed to a bug in my stereo0D->compound conversion.

I did not detect a single case where IXA failed but 1.04 succeeded.

Note that PubChem does not currently support allenic/cumulenic stereocenters, so I did not test that explicitly.

Overall, my impression of IXA is very positive. I would definitely recommend it being included in the "official" InChI code distribution."

Support of simple polymers.

The programming has been delayed due to numerous revisions in the InChI manuscript just published (see below) and is now expected to be finished by the end of 2015.

IUPAC InChI subcommittee & working groups

IUPAC Committees

Chemical mixture composition

Leah McEwen at Cornell has initiated a working group for chemical mixture composition. Recent highly damaging events in chemical laboratories and classrooms [Sheharbano (Sheri) Sangji, a

23-year-old chemistry research assistant, died from injuries sustained in a chemical fire on December 29, 2008, in a laboratory at UCLA] have led to increasing focus on chemical information management in laboratory organizations. The diverse teaching and research environment in the academic sector particularly is raising awareness of the complexity of chemical safety information resources and formats available. A key concern in this regard is that documentation of chemicals with current identifiers is a persistent challenge for tracking and managing chemicals across the chemical enterprise, from process planning to manufacture to waste disposal and emergency response.

The objective of this project is to establish requirements and guidelines for the generation of a unique identifier for all forms of a chemical (liquid, gas, solid, powder, etc.). Currently, many chemical identifiers exist, but very few reflect these bulk properties of substances, which may commonly exist in many forms and mixtures. Furthermore, most existing identifiers present cross-referencing challenges between systems designed around different initial applications and editorial principles.

The intended outcome of this project is global adoption of the InChI notation in chemical inventories and information systems across commercial, industrial, government, academic and educational sectors to facilitate accurate documentation, handling and exchange of chemical information in support of safer management and use of chemicals.

This project is complementary to another user-focused project that is developing a QR code version of the InChI to facilitate labeling and other communication of chemical safety information. That project will be consulting with global stakeholders to determine deployment and use approaches. This project will focus the specificity and usefulness of the information being encoded in the InChI.

This working group is probably unique for the InChI project in that it is of clear scientific value, but even of more importance and value to all the chemistry labs around the world. Safety is something that makes the front page of newspapers and TV news programs

Positional Isomers

Considerable technical interest in positional isomers has developed in the past few months but at the same time Chris Steinbeck at EBI who had hoped to lead this effort has been promoted and does not have sufficient time to chair the working group. Chris is now looking for a new person to lead this working group.

The current members of this working group are:

Christoph Steinbeck
Egon Willighagen
John May
Steffen Neumann
Steve Stein
Roger Sayle
Evan Bolton
Oliver Fiehn

Resolver – The work is now being done under Markus Sitzmann, with assistance from Evan Bolton at NIH/NLM/NCBI/PubChem. Markus has a test web site up and running. It is hoped to have wider testing by year end.

Polymers – This work was finished by the working group under Andrey Yerin. Igor has started programming this standard. The work will be completed and tested in 2015.

Reactions –With the August 2014 signing of a contract with Gerd Blanke this project is again moving ahead nicely. The RInChI working group held a meeting at the Boston ACS meeting to discuss the progress Gerd has made and what next steps are to be taken. The results were that testing of the current code/work should be completed by the end of 2014.

009-043-2-800 Standard InChI-based Representation of Chemical Reactions
http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pi1[project_nr]=2009-043-2-800

Chairman: Grethe, Gunther

Members:
Colin Bachelor
Jonathan Goodman
Hans Kraut
Martin Schmidt
Keith Taylor

Markush – The US Patent Office, with help from the US NTIS (National Technical Information Service) has asked the Trust to participate in a workshop/seminar about InChI for Markush structures expected to be held the end of July 2015. Other than this there has been no progress as there is as yet no funding.

Electronic States – Don Burgess at NIST has developed plans for using InChI for Representations of Species at the Molecular Level. He has recently published the 3 papers on this subject about InChI-ER (Elementary Reactions). The last two came out in the June 2015 issue of IJCK. Being manuscripts from a US Government employee PDF copies are freely available from Don.

D. R. Burgess, Jr., J. A. Manion, and C. J. Hayes
International Journal of Chemical Kinetics **46**, 640-650 (2014)
Data Formats for Elementary Gas Phase Kinetics, Part 1: Unique Representations of Species at the Molecular Level
DOI: 10.1002/kin.20875

D. R. Burgess, Jr., J. A. Manion, and C. J. Hayes

International Journal of Chemical Kinetics **47**, 334-350 (2015)
Data Formats for Elementary Gas-Phase Kinetics: Part 2. Unique Representations of Reactions
DOI: 10.1002/kin.20912

D. R. Burgess, Jr., J. A. Manion, and C. J. Hayes
International Journal of Chemical Kinetics **47**, 361-378 (2015)
Data Formats for Elementary Gas-Phase Kinetics: Part 3: Reaction Classification
DOI: 10.1002/kin.20914

InChI for Materials – No news from the NIST staff about this.

Organometallics- Colin Batchelor and his working group have delivered a report in August 2015 which is being studied.

Inorganics - The lack of a Chair for this working group has been an ongoing frustration. The current chair, Hinnerk Rey from Elsevier/Frankfurt, who replaced Nigel Wheatley, has had to resign owing to insufficient time available. We hope to have a new chair shortly.

2012-046-2-800: Handling of Inorganic compounds for InChI V2
http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pi1[project_nr]=2012-046-2-800

Chairman: TBD
Members:
Damhus, Ture
Druckenbrodt, Christian
Hartshorn, Richard
Schenk, Roger
Sitzmann, Markus

Large molecules, biopolymers/Proteins/biological polymers/macromolecules/biomolecules etc. –

There was a half-day session at the Boston CAS meeting chaired by Keith Taylor on this work.

2013-010-1-800: Implementation of InChI for chemically modified large biomolecules
http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pi1[project_nr]=2013-010-1-800

Chairman: Taylor, Keith

Members:
Blanke, Gerd
Bolton, Evan
Chalon, Didier

Drijver, Alex
Jensen, Jan
Yerin, Andrey
Berman, Helen

Tautomers. – Under the leadership of Marc Nicklaus, NIH/NCI, InChI project #2012-023-2-800, "Redesign of Handling of Tautomerism for InChI V2" was approved for funding by IUPAC. Marc held a working group meeting on this at the Boston ACS meeting which resulted in a plan to implement a revised standard by 2017

2012-023-2-800: Redesign of Handling of Tautomerism for InChI V2
http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pi1[project_nr]=2012-023-2-800

Chairman: Marc Nicklaus

Members:
Bolton, Evan
Ihlenfeldt, Wolf-Dietrich
Peryea, Tyler
Pletnev, Igor
Rey, Hinnerk
Sitzmann, Markus
Tchekhovskoi, Dmitrii

.

QR Codes

The InChI QR code consultation workshop IUPAC project was approved in June 2015. This is the announcement for this project:

"The InChI Trust (http://www.inchi-trust.org/) is examining development of a QR code (2D bar code) version of the InChI. We wish to consult with industry/regulatory/academic sector users to identify and prioritise additional information that could/should be included in the QR code to enhance the value and commercial utility of the QR InChI. Possibilities to be evaluated and elaborated upon include: health/safety information (hazard code and/or safety data URL); catalog code; batch number; inventory information; sample composition/purity. This project is complementary to another user-focused project that is developing InChI for states and mixtures."

Workshop Locations/Dates/Times

Workshop 1: Busan, South Korea, Monday 10 August, 17:00-19:00, coinciding with the IUPAC Congress and General Assembly

Workshop 2: Boston, USA, Wednesday 19 August, 17:00-19:00, coinciding with the Boston ACS meeting.

Workshop 3: Brussels, Belgium, details yet to be determined

January 2015 – July 2015 activities

Manuscripts

The long overdue InChI technical manuscript was finally published in May 2015. The article "InChI, the IUPAC International Chemical Identifier" has now been published in **Journal of Cheminformatics, Volume 7, Issue 1** and is distributed with Open Access under the Creative Commons Attribution Noncommercial License.

The article is available electronically on SpringerLink:
http://www.springer.com/-/3/AU5bMkS9Khf9IB3kOfKX

InChI Trust web site

The Trust web site has left the IUPAC server and is now up on the InChI Trust cloud server. Aletia Rey who was hired to maintain and add content to the web site is doing a nice job.

InChI Usage

For lack of a better a better term, I use InChI Usage to refer to publications and blogs about InChI. Alan and I have been passing these on to Aletia and she has added these to the web site. There have been quite a number of publications using InChI. The numbers continue to grow. Searches on Google (and other search engines) continue to have more hits for InChI strings and InChIKey strings.

InChI Trust Videos - Access numbers:

InChI & the Islands – 728 (7/15); 629 views (12/14); 526 views (7/14)

The Googlable InChIKey – 915 views (7/15); 751 views (12/14); 597 views (7/14)

The Birth of the InChI - 984 views (7/15); 835 views (12/14); 687 views (7/14)

What on earth is InChI? - 2,956 (7/15); 2486 views (12/14); 1977 views (7/14)

An update on the Open Source InChI project – 1,702 views (7/15)
https://www.youtube.com/watch?v=F9XppyZg4E4

IUPAC InChI – 922 views (7/15)
https://www.youtube.com/watch?v=mH9fuspg_h0

Representing Chemical Structures on computer – 390 views (7/15)
https://www.youtube.com/watch?v=uzXkJ9BsyHQ
(InChI part starts at about 14 ½ minutes into the video

Scott Wiedemann
Cheminformatics, Encodings SMILES & InChI – 354 views (7/15)
https://www.youtube.com/watch?v=V9HHnRAS5BA

Two other YouTube InChI videos are available:

Mcule video:
2012 San Diego ACS presentation: Registration system of mcule: InChI is the key
128 views (7/15); 122 Views (12/14 – uploaded 2012)

Audiopedia video
International Chemical Identifier
29 views (7/15) 5 views (12/14 – uploaded 11/14)
(1 minutes video):
http://youtu.be/MG49gn_CdUE

Published on/Uploaded to YouTube Nov 30, 2014:

"The IUPAC International Chemical Identifier (InChI /ˈɪntʃiː/ IN-chee or /ˈɪŋkiː/ ING-kee) is a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information and to facilitate the search for such information in databases and on the web. Initially developed by IUPAC and NIST during 2000–2005, the format and algorithms are non-proprietary. The continuing development of the standard has been supported since 2010 by the not-for-profit InChI Trust, of which IUPAC is a member. The current version is 1.04 and was released in September 2011.
Prior to 1.04, the software was freely available under the open source LGPL license, but it now uses a custom license called IUPAC-InChI Trust License."

Technical Issues

The mechanism to discuss and resolve technical issues continues to work well. Most issues seem to be able to be resolved by email and phone calls, but face-to-face meetings are still very critical as there are some very strongly held opinions that do not get resolved by emails. My regular meetings with NIH (PubChem , NCI, and FDA) staff have been very useful.

Steve Heller