# InChI Trust Project Director's Report

## February 2016

**Summary:**

Since the July 2015 report there continues to be good progress with InChI and the InChI Trust in a number of areas. Bio-Rad has rewritten parts of the InChI code to allow for threading (essentially parallel processing) so that the algorithm would run faster. This code has been tested and gives identical results in the large files it has been tested on. 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 structures, which resulted in their decision not to consider extending InChI for Markush structures at this time. A new InChI working group on chemical mixture composition, which is related to lab safety, has been initiated and is progressing well. In general, we need the working groups to 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, with Sigma-Aldrich joining in November. In addition, NIH has moved from Associate to become a full member of the InChI Trust. As mentioned numerous times in the past, in most organizations, since InChI works and it is not high on their immediate priority lists, real progress is slow without a dedicated champion within the organization.

As of January 2016

Existing Members and Associates: 15

Supporters: Remains at 48

**InChI RFP/Contracts**

The contract with Dr. Gerd Blanke (Germany) for taking forward the RInChI work that Jonathan Goodman and Chad Allen did at Cambridge University is progressing well. Testing of the code is now being undertaken as noted in the attached report from the RInChI working group.

**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 and upcoming work:

Planning InChI Software v. 1.05 release (Winter/Spring 2016)

Release v. 1.05 is a maintenance release.

It includes some improvements of existing code and functionality, as well as new features which do not suppose significant change of existing InChI/InChIKey's base (all other things not included here – extended/modified support for organometallics, tautomerism, Markush structures, etc. – are expected in version 2).

Probably, this is the last release for InChI Identifier version 1.

1. What's in the release:

1.1. General

Added more elements

* updated to current IUPAC confirmed elements list, elements 114 (flerovium, Fl) and 116 (livermorium, Lv) were added.

Extended support of large molecules

* added ability to deal with molecules beyond 1024 atoms limit

* item which is most desirable to be community-tested as pre-release

Added support of Molfile V3000 format

* in principle, intended for large (bio)molecules, enhanced stereochemistry and organometallics

Added support for simple polymers

* simple polymers (no cross-linked, etc.)

Made fixes for various minor bugs/issues

* e.g., some mistakes in AuxInfo printed for isotopic substituted molecules

Made extensive code re-factoring

* simplification,  self-explaining names, avoid conditional compilation, etc., etc.

1.2. Specific parts of the Software

1.2.1. InChI Library (API )

Added support of safe multi-thread execution

* allows to significantly increase speed of InChI/Key generation while running on multi-CPU (most of modern systems)

* code mainly supplied by Bio-Rad

* subsequent changes

* porting to Linux with help of PubChem team (Evan Bolton, Paul Thiessen)

Added native API support for direct Molfile->InChI conversion

* Mol/SD file parser is the same as in inchi-1 executable

* this ensures that any correct caller of this InChI Library procedure will give the same result as inchi-1 executable

Added new executable mol2inchi

* example app demonstrating newest API features

* may in future replace inchi-1 executable (to eliminate hard-to-maintain diversity of codebase)

API usage Python example updated

* shows an example of mol2inchi API call

* migrated to Python 3 (current standard)

Added new "extensible" API - IXA API (by Digital Chemistry)

* IXA stands for “InChI Extensible API”; adds new API procedures including low-level functions to deal with atoms, bonds, etc.

* code supplied by Digital Chemistry

* subsequent changes

* porting to Linux with help of PubChem team (Evan Bolton, Paul Thiessen)

1.2.2. inchi-1 executable

Added new options for inchi-1 executable

Generate specific “error-InChI” string on generation errors

* produce “InChI=1//” instead of just skipping output if InChI generation fails (was requested by users; useful for mass generation and testing)

Treat inchi2struct warnings as errors

* consider warnings related to inability of adequately restoring InChI as errors

Improved support for wildcard-indicated multiple input files under Windows

* implemented through in-program wildcard expansion (was through Windows own mechanism, had issues for huge number of files)

1.2.3. InChI Certification Suite

* updated and expanded functionality

* expanded ability to deal with large files

* migrated to Python 3 (current standard)

2. Current status of the release

2.1. Code & testing

Chiefly all the things work (with some reservation for polymers, at the moment).

Most current efforts are concentrated on:

- making all the things work altogether (some code changes were implemented as separate branches; being OK themselves, they must be finally merged which is not easy sometimes) and pass the tests altogether;

- finalizing polymers,

- testing large molecules.

2.2. Documentation to be delivered in early February.

3. Suggested timelines and pre-release testing

It is supposed that the Software will first be pre-released for tests by broad InChI community, with an announcement in inchi-discuss listserver

- from the Minutes of Boston 2015 Meeting: "… release would be made available to members of the SourceForge inchi-discuss list for comment, as part of the IUPAC review procedure. Alan McNaught would look again at the Heads of Agreement with IUPAC, add a fuller statement of the review procedure, and recirculate to the Board for comment. The relationship with IUPAC's new Committee on Publications and Cheminformatics Data Standards (CPCDS) should be defined."

## IUPAC InChI subcommittee & working groups

## IUPAC Committees

## Chemical mixture composition

Leah McEwen at Cornell University, the bright light of new working groups 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.

A revised project plan, entitled “InChI Extension for Mixture Composition” was submitted to IUPAC in January 2016. Leah gave a talk at Pacifichem on this (Paper 134 – A Chemical View on Research Data: Identifying Critical Structures for Management in the Machine Era”)

**Positional Isomers**
Considerable technical interest in positional isomers has developed in the past few months. Chris Steinbeck at EBI is 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 continues to work on this. He has a put together a beta test version with some infrastructure and some test content. I expect to have much more to report on this next time.

**Polymers**
As noted in Igor’s report, this work will be completed shortly..

**Reactions**
With the signing of a contract with Gerd Blanke this project is again moving ahead nicely. The RInChI working group met at the Boston ACS meeting and discussed the progress Gerd has made and what next steps are to be taken. The working group leader, Guenter Grethe, has provided a detailed report on the status of the project, which is attached to this report. There is an issue with how Google and other search engines index RInChIs.

Working group members are being asked to test the program with reactions from in-house databases or from RD files supplied on the web test site. After the successful conclusion of these

tests and the incorporation of modifications, a beta test will be carried out by members of the SourceForge group. Thereafter, the first release of RInChI is projected to be in early 2016.
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 Batchelor
Jonathan Goodman
Hans Kraut
Martin Schmidt
Keith Taylor

**Markush**
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 structures at the end of July 2015. As noted in the summary nothing resulted, yet again, from discussions with the Patent Office.

**Electronic States**
Don Burgess at NIST has developed plans for using InChI for Representations of Species at the Molecular Level. In 2014/2015 he 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. There has been no further developments here.

**Organometallics**
Colin Batchelor and his working group expect a final report in 2016.

**Inorganics**
The lack of a Chair for this working group has been an ongoing frustration. This working group proposal to IUPAC for funding was approved in 2013 but there has not been any progress report yet.
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

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

**Large molecules, biopolymers/Proteins/biological polymers/macromolecules/biomolecules, etc**.
Nothing has happened since the October 2014 working group meeting at NIH as waiting for the extensions of InChI past 1024 atoms. With this to be accomplished in v1.05 it is hoped progress will follow. There will be a Large Molecules meeting following the InChI Trust Board meeting in London in February 2016.

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 in 2015 with a plan to move forward.

Here is a short summary of his most recent report as of January 2016. His full report is attached.

It has to be noted that all following research activities, while spurred on by the IUPAC project, are conducted within Marc's group at the National Cancer Institute, funded entirely with NCI funds.

Since one of the starting points of the current initiative had been the proposal (by original InChI developer Dmitrii Tchekhovskoi) to base the tautomerism handling redesign on the scheme as implemented in the chemoinformatics toolkit CACTVS,NCI performed the following analysis of the approximate 20 default transforms in the toolkit: NCI took a large database of small molecules that are commercially available, the ChemNavigator/Sigma-Aldrich Market Select (AMS) database of about 6 million diverse samples aggregated from hundreds of original suppliers world-wide (see http://www.aldrichmarketselect.com/). For each of these rules, NCI analyzed its frequency in the AMS, as well as the number of cases for which the CACTVS rules declare two or more products in the AMS, labeled as different compounds in the database, to be really just tautomers or each other. NCI found over 31,000 such cases (the vast majority were tautomer pairs, but tautomeric multiplicity up to five-fold was found).

NCI selected about 140 such tautomer tuples (mostly pairs), covering as many of the CACTVS rules as possible in as unambiguous a way as possible. A total of 277 samples were received, providing us with 127 tautomer pairs and 5 triples. These samples have been analyzed by NMR, and a subset of them by mass spectrometry. The analytical goal is to determine if the two samples represent the same compound. The results will help us determine if the chemoinformatics rules – which said, "it's the same stuff" – were right, or if the vendors and AMS were correct in labeling these as different compounds. The analysis of the obtained spectra as far as conducted by now (it has been nearly completed) seems to indicate that the prototropic tautomer transforms NCI has tested do for the most part work better than the vendor representation, i.e. predict analytical identity of the samples shown with different connectivities. This seems to also hold for about half of the ring-chain transforms mentioned below.

Dr. Laura Guasch, while postdoc in Marc's NCI group (she took up a new position late 2015), developed a set of new rules, encoded as CACTVS transforms in SMARTS format, that cover ring-chain (RC) tautomerism. NCI has been conducting a similar analysis as for the prototropic transforms to determine if RC tautomerism could be added to the types of tautomerism handled by InChI V2. NCI has analyzed by NMR 34 pairs of ring-closed and ring-open isomers that our new rules declare as tautomers of each other.

Parallel to these efforts, NCI has been working toward the creation of a database of tautomerism-related data from literature ("Tautomer DB"). So far, NCI has focused on experimental results such as obtained by various spectroscopic techniques. The work of extracting structures, tautomeric ratios or interconversion rates/barriers, the conditions of the experiments etc. was performed via a contract issued by NCI to a commercial company (Reverse Informatics,

Chennai, India). NCI identified a subset of about 200 papers out of approximately 1,000 possible references that NCI deemed as most interesting, and from which the company extracted structures plus comprehensive annotations.

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 prioritize 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.

The report from these two meetings is as follows:

Despite widespread advertising before and during the meetings at which the workshops were held, the attendance at the workshops was not high (seven and eleven people, respectively), and was predominantly people whom we had approached directly, and whom we knew to be motivated. Notwithstanding the low attendance, the discussions were fruitful and useful knowledge was gained.

The principal conclusions from the workshops were as follows:

* InChI-based labelling is not seen as particularly useful for large scale chemical manufacturers. They already have and provide large amounts of (often different) paperwork with shipments to different countries.

* Research organizations are VERY interested, particularly in the context of inventory control and management of health and safety, often in relation to tracking/storing/using small amounts of large numbers of different fine chemicals.

* The following additional information would be desirable: UN GHS or similar hazard code (bearing in mind country variability); unique bottle number; URL or mechanism to link to safety data sheets (with provenance/curation); transport classifications.

People in the second workshop were very interested in just how much data might practically be included in the QR code in addition to the InChI or InChI key. Given that InChIs can be quite long, the speculation was that InChI keys may be required, both in terms of limiting the amount of the QR code that would be required for the InChI information and being able to standardize the placement of other information.

The following key use cases were identified as needing further investigation/development:

* Inventory tracking and control (requires individual bottle coding)

* Location and compatibility (H&S information and linking to an institutional database of storage locations – each location barcoded?)

* Links to safety data sheets (inclusion of UN GHS, URL of H&S source, which raises questions of curation and provenance)

Recommended Future Direction for Consultation Project

Abandon third workshop in favor of targeted interviews of key people.

Develop scenarios for the amount(s) of data that could be encoded.

Identify and develop use cases for the QR InChI, using interviews with key people to provide additional information and insights into the possibilities.

**August 2015 – December 2015 activities**

**Meetings Attended; Talks/Posters Presented**

Lecture: Patterson-Crane Award lecture, September 8, 2015, Columbus OH

Poster: GDCh meeting in Fulda, Germany, November 2015. Poster presentation on InChI

Poster: Pacifichem, Honolulu, Hawaii, December 2015. Poster #1 1279: “Recent Developments in the Worldwide in the IUAPC Chemical Identifier Structure Standard”. Also co-authored the presentation with Leah McEwan mentioned above (Paper 134).

**Manuscripts**

The long overdue InChI technical manuscript was finally published in May 2015, but the html version had problems which were not fixed until just recently. 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: *J Cheminformatics*, December 2015, 7:23, First online: 30 May 2015, http://www.jcheminf.com/content/7/1/23/abstract

doi:10.1186/s13321-015-0068-4

http://www.ncbi.nlm.nih.gov/pubmed/26136848?dopt=Abstract&holding=f1000,f1000m,isrctn

Number of downloads/views (as of 1/17/16) –

Last 30 days: 286 accesses
Last 365 days: 4536 accesses
All time: 4536 accesses

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

It has received 5 citations:

http://citations.springer.com/item?doi=10.1186/s13321-015-0068-4

Altmetric rating – 15

The first major InChI article, published in 2013 continues to have considerable viewing:

11301 Accesses

InChI - the worldwide chemical structure identifier standard

Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, Igor Pletnev Journal of Cheminformatics 2013, 5:7 (24 January 2013)

Fabienne Meyers wrote a very positive article about InChI for the January 2016 issue of Chemistry International summarizing the articles in J. Cheminformatics (see above) and the article by Wendy Warr (see below). Here is what Fabienne wrote:

*Seminal InChI Publications*

Many InChIs and quite some feat—that is the greeting offered to all interested in the IUPAC International Chemical Identifier, InChI, in a review published by Wendy Warr in J Comput Aided Mol Des (August 2015, 29:8, pp. 681-694; http://dx.doi.org/10.1007/s10822-015-9854-3). The article details not only the history of the project, but also its recent and current development, and celebrates 15 years of work since the project was first launched in 2000. The 14-page paper includes 275 references.

Warr recalls that, back in 2000, the increasing complexity of molecular structures made conventional naming procedures inconvenient, and there was no suitable, openly available electronic format for linking chemical structures over the Internet. InChI was developed as a freely available, non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of data compilations and unambiguous identification of chemical substances. InChI was developed under the auspices of IUPAC, with principal contributions from the US National Institute of Standards and Technology

(NIST), and more recently, the InChI Trust. While the UPAC InChI SubCommittee has the responsibility for continued authentication of the InChI standard, the InChI Trust, in turn, develops and supports the non-proprietary InChI standard and promotes its use to the scientific community. Version 1 was launched in 2005and the current version (1.04) was released in September2011. CI readers have been regularly updated on InChI, its developments and applications (A. McNaught, Chem Int 28(6):12-14 (2006); J. Frey, Chem Int 28(6):14-15 (2006); S. Heller and A. McNaught, ChemInt 31(1):7-9 (2009); A. Yerin, et al, Chem Int 35(6):12-15(2013)).

A second article released slightly earlier, in May2015, by the core project task group, documents the design, layout, and algorithms of InChI, and was published in Journal of Cheminformatics (Heller et al. (2015)7:23; http://dx.doi.org/10.1186/s13321-015-0068-4).
This 34-page article is intended to provide a reasonably detailed description without being overlong fora journal article. For a more technical description, the reader is referred to the InChI Technical Manual and the free source code of the InChI software, available from the InChI Trust website at www.inchi-trust.org .In her conclusion, Warr reminds us that establishing a standard is not a quick process; and that we should be impressed with what InChI has achieved over 15years. She also declares that the InChI Trust does not intend to stop there. A new application programming interface (API) is being tested. Currently, the InChI algorithm can handle neutral and ionic organic molecules, radicals, and some inorganic, organometallic, and coordination compounds. Steps to expand it to handle more complex chemical structures are underway. Work continues on polymers and mixtures. Efforts on generic (‘‘Markush’’) structures are planned once funding is secured. In the longer term, consideration will be given to the InChI resolver, QR codes for InChI, macromolecules, positional isomers, crystal structures, and to extending the coverage of inorganic, organometallic and coordination compounds.

Volunteers willing to join the team will define the InChIs of the future. IUPAC and the InChI Trust continue to seek new members and people willing to help maintain and enhance the InChI standard. For more information, contact Steven Heller <steve@hellers.com>

Wendy Warr has published an excellent article on InChI:
“Many InChIs and quite some feat”
J. Computer-Aided Molecular Design
August 2015, Volume 29, Issue 8, pp 681-694
First online: 17 June 2015

**InChI Trust web site**

The Trust web site has left the IUPAC server and is now up on the InChI Trust cloud server.

**InChI Usage**

For lack of a better a better term, I use InChI Usage to refer to publications and blogs about InChI. There have been quite a number of publications using InChI, these to the web site. 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 – 804 (1/16); 728 (7/15); 629 views (12/14); 526 views (7/14)

The Googlable InChIKey – 1,037 (1/16) ; 915 views (7/15); 751 views (12/14); 597 views (7/14)

The Birth of the InChI - 1,084 (1/16); 984 views (7/15); 835 views (12/14); 687 views (7/14)

What on earth is InChI? - 3,331 (1/16); 2,956 (7/15); 2486 views (12/14); 1977 views (7/14)

An update on the Open Source InChI project – 1,790 (1/16); 1,702 views (7/15)
https://www.youtube.com/watch?v=F9XppyZg4E4

IUPAC InChI – 931 (1/16); 922 views (7/15)
https://www.youtube.com/watch?v=mH9fuspg_h0

Representing Chemical Structures on computer – 546 (1/16); 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 – 468 (1/16); 354 views (7/15)
https://www.youtube.com/watch?v=V9HHnRAS5BA

**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.

**Plans for the first half of 2016**

For 2016 my overall plans and goals are as follows:

1. Work to expand the current membership with two basic classes of members – Full and Associate, and add to the number of Supporters. Work to sign up more organizations for the Certification Suite.
2. Continue to attend meetings and give talks on InChI where useful and appropriate.
3. Attend ACS meeting in San Diego and meet with groups to discuss adoption and usage of InChI.
4. Attend the April BioIT meeting and chair a session on InChI for Large Molecules.
5. Work with Sigma-Aldrich to add InChIs and InChIKeys to their products.

Steve Heller