TONY DAVIES COLUMN

The ACS 2013 symposium on exchangeable data formats

Robert Lancashire[a] and Antony N. Davies[b,c]
[a]University of the West Indies, Mona Campus, Kingston, Jamaica
[b]Expert Capability Group – Measurement and Analytical Science, Akzo Nobel, Deventer, The Netherlands
[c]SERC, Sustainable Environment Research Centre, University of South Wales, Glamorgan, UK

In this edition Robert Lancashire has agreed to write up his experiences in being involved in helping to organise and preside over a successful two-day symposium for the Chemical Information Division (CINF) of the American Chemical Society at their recent 246th ACS National Meeting and Exposition, held over 8–12 September 2013 at the Indiana Convention Center in Indianapolis.

The idea

In early December 2012, Robert Lancashire had an e-mail from Tony Williams suggesting that they co-chair a symposium session at the September 2013 ACS meeting in Indianapolis. Jeremy Garritano had agreed to act as the Programme Chair of the CINF Technical Programme sessions and they coordinated through him. The symposium would centre on "Exchangeable Molecular and Analytical Data Formats and their Importance in Facilitating Data Exchange". It would include the challenges of exchanging molecular data formats and analytical data formats. The idea was to involve key developers and users of molfiles, InChI, JCAMP-DX, AniML etc. Within a fairly short time we had commitments from nearly 20 speakers and started to arrange the talks into sessions to cover the themes.

Spectroscopy relevant highlights—Wednesday & Thursday, 11–12 Sept

During the morning session on molecular data formats, Keith Taylor and Roger Sayle both noted that while a small number of formats were in common use (like the ubiquitous MOL file), some

The complete programme

Keith T. Taylor	Cheminformatics runs on molfiles and its siblings: There is a molfile for that
Roger A. Sayle	Reading and writing molecular file formats for data exchange of small molecules, biopolymers and reactions
Geoffrey R. Hutchison	Facilitating accurate chemical data interconversion using Open Babel: The good, the bad, and the painful
David Gosalvez, Alex Jewett, Phil McHale, Churl Oh, Rudy Potenzone, Chris Strassel	Exchanging chemical structures and ELN data: CDX, CDXML and other formats
Stephen Heller	InChI: Recent developments in the worldwide chemical structure identifier standard
Evan Bolton	Data exchange caveats and particulars, the devil is in the details
Sergio Rotsein, John Wise, Claire Ballamy, Michael Braxenthaler, Barry Bunin	Pistoia Alliance and the emerging HELM standard at the "dawn of the ADC informatics era"
Antony N. Davies, Robert J. Lancashire	30 years of JCAMP-DX formats and still going strong
Michael Boruta	Knowledge sharing or what I learned in first grade
Clemens Anklin	Long wait for exchangeable data formats vs the evolution of data
Stuart J. Chalk	Leveraging the AnIML specification for analytical data exchange
Robert M. Hanson, Robert J. Lancashire	JCAMP-MOL: A JCAMP-DX extension to allow interactive model/spectrum exploration using Jmol and JSpecView
Antony J. Williams, Colin Batchelor, Jon Steele, Valery Tkachenko	Importance of standards for data exchange and interchange on the Royal Society of Chemistry eScience platforms
Roger Sayle, Daniel Lowe, Noel O'Boyle	Extraction, analysis, atom-mapping, classification and naming of reactions from pharmaceutical ELNs
Richard Kidd	How standards helped RSC to create The Merck Index Online
Bin Chen, Bing He, Ying Ding, David Wild	Semantic mining and prediction for drug discovery

TONY DAVIES COLUMN

users did not conform to either the MOL V2000 or V3000 published standards. Geoffrey Hutchinson gave a description of the OpenBabel project (http://openbabel.org) that has produced a toolbox to read, write and convert over 110 chemical file formats and the difficulties this non-conformity has created.

Phil McHale, PerkinElmer: "CDX/CXML used as a standard file format by the US Patent Office, accepted by *J. Am. Chem. Soc.*, inconvertible with molfiles/SMILES/InChI, compatible with SciFinder, Reaxys, ChemSpider, Spotfire, eMolecules, KRMS, Vortex…"

In the presentation by Phil McHale, he noted that Perkin-Elmer was working on an Open XML format for export of data from Electronic Laboratory Notebooks (ELN). At present these were generally proprietary. He reviewed the CDX and CDXML formats as well, both of which have been widely accepted and utilised.

Stephen Heller gave an update on the InChI representation and the InChI-Trust. Like Barcodes and QR codes, InChIs are not designed to be interpreted by humans but are produced by computer from structures drawn on-screen with existing structure drawing software. The original structure can be regenerated from an InChI with appropriate software. He noted that a number of videos had been produced to attempt to explain their application.

Stephen Heller, InChI Trust: "Too many 'Standards' actually slow things down and make getting to the information you want and need take a lot longer time and effort than it would with InChI"

Evan Bolton spoke about the new features in the PubChem data submission portal that supported a wide range of user-defined data and the need for data standards. Barry Bunin noted that there was no standard approach for a computer-based way of managing large molecules such as: peptides, antibodies, therapeutic proteins or vaccines. HELM (Hierarchical Editing Language for Macromolecules) was being introduced as an Open Source approach by Pfizer and was released into production in 2008. He then introduced the CDD (Collaborative Drug Discovery) vault as a hosted database solution for secure management and sharing of chemical and biological data.

For the afternoon session on spectroscopic data, the first presentation was a joint paper from Tony Davies and Robert Lancashire that gave some history on the JCAMP-DX data formats. Recognition was given to Paul Wilks, Bob McDonald as well as Jeannette Grasselli-Brown as pioneers in the publication of JCAMP-DX standards. We noted that the earliest published standard was for IR (version 4.24) and whereas we have seen some labelled 4.1, we are not aware of any labelled as version 1,2,3 or 4.0 data files! Since 1988 the standards for a wide range of techniques have been published and from 1995 they became the responsibility of IUPAC.

Michel Boruta followed by showing the transition from hand-written annotations on chart paper copies of spectra to electronic equivalents that could be stored in "knowledgebases". For example, ACD/Labs Spectrus Process includes separate knowledgebases for IR and Raman. The assignments can be exported as part of JCAMP-DX files.

Clemens Anklin identified the common data formats used for various techniques. In the case of NMR this was predominantly JCAMP-DX. He lamented the fact that whilst 2D NMR had existed from before any JCAMP-DX standards were published, the latest accepted standard for NMR was 5.01 published in 1999 and this only covered 1D. The version 6 format for 2D has been in draft since 2002 and has been implemented by vendors who could not wait for it any longer.

Clemens Anklin, Bruker: "Where is NMR Today?

- Data storage should be complete
- It has to be possible to recreate or rerun the data collection with the information stored
- This also applied to exchangeable data formats
- For NMR this means: Pulse program, parameters, extras such as decoupling sequences, lists, shaped pulses etc."

"ChemSpider:

- ChemSpider requires spectral data to be deposited in standard formats – JCAMP or images
- All spectra are available at http://www.chemspider.com/spectra.apsx
- Data are deposited on a regular basis
  - Students
  - Chemical Vendors
  - Growing collection now"

Stuart Chalk introduced the AnIML specification and highlighted the features and benefits of using an XML protocol that could be fully validated.

He noted that from 2003 it was designed to be a (backwards compatible) replacement for JCAMP-DX. The Task Group guiding the process set its charter: "to develop an analytical data standard that can be used to store data from any analytical instrument" and holds virtual meetings on a monthly basis to develop the specification. The first set of specifications are targeted to go through ASTM balloting in early 2014.

Bob Hanson finished this session with a proposal to have an extension to the JCAMP-DX standard whereby a single file could contain the molecular graphics data as well as the spectrum, together with annotations linking the two. This would allow interaction with cloud services such that a MOL file could be passed to a server and a simulated spectrum returned with sufficient information to apply all the required annotations to identify the peaks.

Conclusions

Robert and Tony Williams successfully put together a programme for the ACS which reads like a partial who's-who in the essential development, support and deployment of standardisation in chemical informatics. Well done to them.

continued on page 26

continued from page 23

References

The full CINF Technical Program is at: *Chemical Information Bulletin* **65(3)**, 39 (2013). http://bulletin.acscinf.org/node/483; the presentations will be available via the ACS CINF site soon

InChI videos

a. What on Earth is InChI? http://www.youtube.com/watch?v=rAnJ5toz26c

b. The Birth of the InChI http://www.youtube.com/watch?v=X9c0PHXPfso

c. The Googlable InChIKey http://www.youtube.com/watch?v=UxSNOtv8Rjw

d. InChI and the Islands http://www.youtube.com/watch?v=qrCqJ0o4jGs