J Comput Aided Mol Des (2012) 26:151–152
DOI 10.1007/s10822-011-9502-5

PERSPECTIVE

Silver threads

Wendy A. Warr

Received: 13 November 2011 / Accepted: 29 November 2011 / Published online: 9 December 2011
© Springer Science+Business Media B.V. 2011

It seems to be the season for silver anniversaries. As I write this in November 2011 I have just returned from the 7th German Conference on Chemoinformatics [1], where the Chemistry-Information-Computer (CIC) Division of the German Chemical Society (Gesellschaft Deutscher Chemiker) was celebrating the 25th anniversary of the CIC workshops [2]. By the time this article appears, not only will it be the silver anniversary of the *Journal of Computer-Aided Molecular Design*; it will also be 25 years since the first Noordwijkerhout International Conference on Chemical Structures [3], now known as "ICCS".

The 1987 ICCS meeting concentrated heavily on in-house and commercial systems for storing chemical structures and related data; on searching by substructures, and 3D and generic structures; and on chemical reaction retrieval, planning, and prediction. The first CIC workshop had similar themes. All the basic work on structure representation was complete by the 1990s but some of the same applications are still being addressed, albeit using approaches of the current decade. ICCS1 in 1987 featured a paper by Glaxo on an integrated system for handling chemical structures and associated data; the ChemAxon US Users meeting this year featured yet another GSK chemical registration system [4]. Chemical inventory systems are also being written and rewritten [4].

Using cheminformatics, chemical and pharmaceutical companies have been able to prevent duplication of synthetic effort and have been better able to handle patent information. In particular, the pharmaceutical industry has benefited from systems that enable it to keep track of its inventories and proprietary compounds, and from software that makes drug discovery significantly faster and more efficient.

Work continues on the special problems of structure representation: more unusual forms of stereochemistry, polymers, and generic structures, for example [5]. There was much research into the handling of generic structures (so-called Markush structures) in the 1980s, culminating in the launch of two proprietary systems: Thomson Reuters' Markush DARC and CAS' MARPAT [6]. Much more recently, interest in Markush searching has been revived [4, 6], especially now that Thomson Reuters has released its data for in-house use [4]. This could have implications for combinatorial libraries as well as for patent searching.

Patents are of fundamental importance to the pharmaceutical industry; its profits depend upon protecting its intellectual property and much research funding depends on the profits of the pharmaceutical industry. Unfortunately patents do not make chemical structures as patent (i.e., open) as some users might like. The writing of patents is a black art in which the agents who are expert in the field seek to obfuscate the nature of their inventions in complicated generic structures, albeit alongside a few representative complete structures. The generic structures are then encoded into expensive commercial databases which are searchable only by information professionals who are prepared to cope with the user hostility of the related proprietary systems. Although the whole archaic legal system and its paraphernalia are unlikely to change any time soon, the wind of change does seem to be blowing through the software.

One useful implication is that patents are published openly, whereas much of the scientific literature is under copyright. Patents are thus widely available for those who do research in text mining. There has been considerable interest in this field of late. Text mining [7–13] recognizes

W. A. Warr (✉)
Wendy Warr & Associates, Holmes Chapel,
Crewe, Cheshire CW4 7HZ, UK
e-mail: wendy@warr.com

chemical entities in journal articles or patents and extracts them for conversion to connection tables, and, in some cases, into IUPAC International Chemical Identifiers (InChIs) [5]. Related work converts images of chemical structures into connection tables [14–19].

Cynics, of course, have pointed out that there would be no need to mine the literature to claw back chemical structures and data if the structures had been published "live" and reusable in the first place, rather than converted into PDF images (and generic structures), and if the related data had been deposited into suitable repositories or databases in data centers. The fault lies not only with publishers; researchers themselves have proved reluctant to submit connection tables and other data files, unless mandated to do so. Later in 2012 I hope to do an article on the "information loss" problem and its technologies, legalities and economics.

Without stealing the thunder from that article, I need to add a few words to answer some of the questions that the editors asked in soliciting articles for this issue of the journal. I have no intention of looking 25 years ahead. In the IT arena, foolish predictions have been made over much shorter periods: "640K ought to be enough for anybody" is a memorable example. In the field of scientific publishing, Outsell predicts that the "five to watch" for the next 2 years are Application Programming Interfaces (APIs), tablets, social messaging, HTML5 and MySQL.

Publishers are opening up APIs so that applications can be built that incorporate the scientific literature into a scientist's workflow. The new dynamic Web will no longer deliver static documents but dynamic applications that interact with backend Web Services. Scientific "publishing" will morph into true scientific communication. In the early 2000s Murray-Rust and Rzepa coined the word "datument" [20] to define a compound information object, with not just text but live chemical structures and data. In the future (probably the near future) inclusion of audio and video in "datuments" will become commonplace.

What sort of chemical data will be proliferated and will it be of dubious quality? Olga Kennard has said [21] that the foundation of the Cambridge Crystallographic Data Center fulfilled a dream she shared with J.D. Bernal. They "had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments" [22]. Readers of this journal depend on the Protein Data Bank and the Cambridge Structural Database, and other resources, for high quality data.

I have talked about mining and the use of mining to create new knowledge but if you mine unchecked data you cannot rely on the outcome. The sharing and re-use of quality data undoubtedly lead to new science but the cultural and commercial complexities of capturing data and maintaining data centers are manifold. Bioinformatics and cheminformatics have considerable cultural differences and some people believe that chemistry lags behind bioscience. Unfortunately science and technology are much faster to change than people and their cultures. When it comes to culture and prejudices, we cannot expect much to change in the next 25 years.

References

1. 7th German Conference on Chemoinformatics (2011) November 6–8, Goslar, Germany https://www.gdch.de/veranstaltungen/tagungen/tagungen-2011/gcc.html. Accessed 12 Nov 2011
2. Gasteiger J (ed) (1987) Software-Entwicklung in der Chemie I: Proceedings des Workshops "Computer in der Chemie" Hochfilzen/Tirol, 19–21 November 1986 (Software Development in Chemistry 1. Proceedings of the workshops on the computer in chemistry, Hochfilzen/Tirol, November 19–21, 1986). Springer, Berlin
3. Warr WA (ed) (1988) Chemical structures: the international language of chemistry. Springer, Berlin
4. Warr WA (2011) ChemAxon US User Group Meeting, San Diego, September 26–29. http://www.chemaxon.com/ugm-presentations/2011-us/#meeting-report. Accessed 12 November 2011
5. Warr WA (2011) WIREs Comput Mol Sci 1(4):557–579
6. Downs GM, Barnard JM (2011) WIREs Comput Mol Sci 1(5): 727–741
7. Banville DL (2006) Drug Discov Today 11(1 & 2):35–42
8. Corbett P, Murray-Rust P (2006) High-throughput identification of chemistry in life science texts. In: Berthold MR, Glen RC, Fischer I (eds) Computational life sciences II. Springer, Berlin, pp 107–118
9. Kolarik C, Hofmann-Apitius M, Zimmermann M, Fluck J (2007) Bioinformatics 23(13):i264–i272
10. Klinger R, Kolarik C, Fluck J, Hofmann-Apitius M, Friedrich CM (2008) Bioinformatics 24(13):i268–i276
11. Banville DL (2009) Curr Opin Drug Discov Devel 12(3):376–387
12. Banville DL (ed) (2009) Chemical information mining: facilitating literature-based discovery. CRC Press, Boca Raton
13. Zimmermann M, Fluck J, Thi LTB, Kolarik C, Kumpf K, Hofmann M (2005) Curr Top Med Chem 5(8):785–796
14. McDaniel JR, Balmuth JR (1992) J Chem Inf Comput Sci 32(4):373–378
15. Simon A, Johnson AP (1997) J Chem Inf Comput Sci 37(1): 109–116
16. Zimmermann M (2007) Nachr Chem 55(10):997–1000
17. Filippov IV, Nicklaus MC (2009) J Chem Inf Model 49(3): 740–743
18. Park J, Rosania GR, Shedden KA, Nguyen M, Lyu N, Saitou K (2009) Chem Cent J 3:(online)
19. Valko AT, Johnson AP (2009) J Chem Inf Model 49(4):780–787
20. Murray-Rust P, Rzepa HS, Williamson MJ, Willighagen EL (2004) J Chem Inf Comput Sci 44(2):462–469
21. Kennard O (1996) Bernal's Vision: from Data to Insight. J. D. Bernal Lecture 1995. Birkbeck College, London
22. Kennard O (1997) Session 6: access to scientific data repositories. From private data to public knowledge. In: Butterworth I (ed). The Impact of Electronic Publishing on the Academic Community. Portland Press, London. http://www.portlandpress.com/pp/books/online/tiepac/session6/ch2.htm Accessed 23 November 2011