Internet-based tools for communication and collaboration in chemistry

Antony J. Williams

ChemZoo Inc., 904 Tamaras Circle, Wake Forest, NC 27587, United States

Web-based technologies, coupled with a drive for improved communication between scientists, have resulted in the proliferation of scientific opinion, data and knowledge at an ever-increasing rate. The availability of tools to host wikis and blogs has provided the necessary building blocks for scientists with only a rudimentary understanding of computer software science to communicate to the masses. This newfound freedom has the ability to speed up research and sharing of results, develop extensive collaborations, conduct science in public, and in near-real time. The technologies supporting chemistry, while immature, are fast developing to support chemical structures and reactions, analytical data support and integration to related data sources via supporting software technologies. Communication in chemistry is already witnessing a new revolution.

The wikisphere and the blogosphere

Many scientists have an urge to communicate either their own science or science in general, commonly with the intention of educating others, proliferating data or opinions, or connecting with others for collaboration or advice. Although scientific meetings have offered this socializing aspect, historically, the publication process has been the primary manner by which communication to an extended audience has been facilitated. This process can result in education, proliferation and collaboration, but puts many hurdles in the path of social and collaborative science, not least being the time associated with publishing data, the limited communication channels for facile feedback associated with a standard form of publishing, as well as the potential costs to the readers and authors of accessing the information.

A shift is already underway in terms of willingness and an interest in using web-based software tools to speed communication. Wikis and blogs are now common terms for the majority of users of the worldwide web. Most, if not all, readers of this article have at least browsed through Wikipedia to see what all the attention is about and probably have browsed a few blog sites, though not necessarily for the sake of science. Both wikis and blogs are fast becoming chosen platforms for the exchange of information between many scientists, not only as tools within their own research groups but also, more generally, with the public in general.

A blog or weblog is a website where entries are written in a chronological order and generally provide commentary or news on a particular subject (see http://www.en.wikipedia.org/wiki/Blog). A typical blog combines text, images and links to other blogs, web pages, and other media related to its topic (see Fig. 1 for an example of a chemistry blog written by a synthetic chemist). The ability for readers to leave comments and interact with the author is an important component of blogs (though such capabilities do exist in the wiki environment in terms of discussion groups) and is really the most interesting value proposition of the medium itself. The original blog posting remains untouched by the commenter, and is, therefore, less prone to abuse than the wiki environment, but readers are free to add their comments, generally in a mediated manner, where the blog host retains control over the postings. The number of chemistry-related blogs continues to grow dramatically and there have been efforts to provide a unified view into some of these (see http://www.scienceblogs.com/ and http://www.cb.openmolecules.net/) as well as provide an environment regarding 'how to' participate in scientific blogging (see http://www.wiki.scienceblogging.com/scienceblogging/). Of the chemistry blogs online, many are hosted by students and, in general, there is sadly a lack of contributors by teachers in this environment. There are blogs from members of the

E-mail address: antony.williams@chemspider.com.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2008.03.015

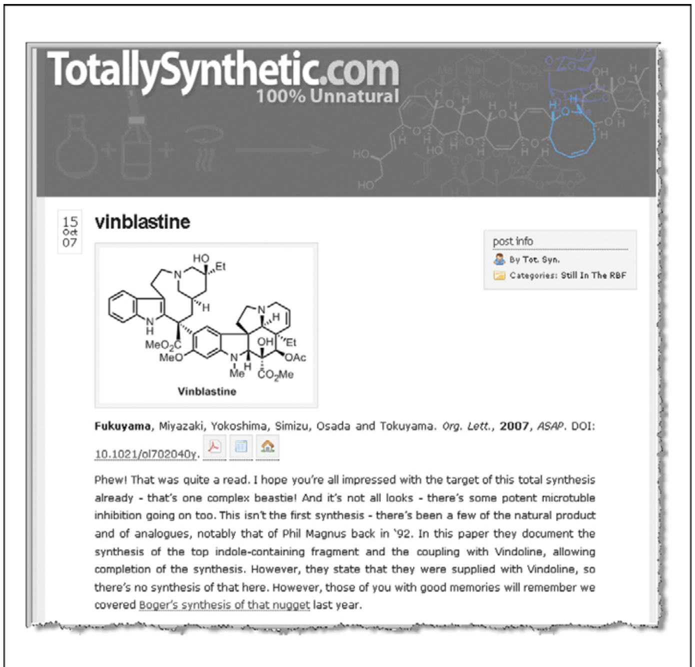

FIGURE 1

The TotallySynthetic.com blog. Paul Docherty discusses complex syntheses and offers readers an opportunity to comment, analyze and provide feedback. Many articles are labeled with InChIKeys to allow indexing by search engines (http://www.totallysynthetic.com/blog/).

pharmaceutical industry, from the cheminformatics world, from the open source chemistry software world and other willing participants in the 'blogosphere'. Many organizations are, appropriately, concerned with their employees blogging and, especially for nonacademic environments, there are likely to be company guidelines and policies in place.

The blog environment certainly provides more of a social networking environment than a wiki and can result in a more immediate response to a question, such as a call to collaborate on a funding proposal (see http://www.blog.openwetware.org/scienceintheopen/2007/11/22/e-science-for-open-science-an-epsrc-research-network-proposal/), an immediate path to communication of important issues (see http://www.wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=887), and in career-damaging dialogs across the community (see http://www.pipeline.corante.com/archives/2006/06/05/hexacyclinol_or_not.php).

A wiki is a type of computer software that allows users easily to create, edit and link web pages (see http://www.en.wikipedia.org/wiki/Wiki). A wiki enables documents to be written collaboratively, in a simple markup language using a web browser, and is essentially a database for creating, browsing and searching information. A defining characteristic of wiki technology is the ease with which pages can be created and updated. Certainly Wikipedia is the most well known option today, though there are many others already online and used within the confines of an organization to manage content.

There are active groups supporting the development of chemistry on Wikipedia and there are now thousands of pages describing small organic molecules, inorganics, organometallics, polymers and even large biomolecules. Focusing on small molecules in general, each one has a drug box (see http://www.en.wikipedia.org/wiki/Template:Drugbox) or a chemical infobox (see http://www.en.wikipedia.org/wiki/Wikipedia:Chemical_infobox). A drug box provides identifier information (chemical name, registry number and so on) and commonly the identifiers link out to a related resource. Chemical data, pharmacokinetic data and therapeutic considerations can also be listed.

Readers are probably aware of the criticisms of Wikipedia in terms of potential quality issues. This author has been highly impressed with the overall quality of the chemical information on the site and, when the author has found errors, has been able to make appropriate edits to the immediate benefit of the community. A curation project is presently underway to validate each of the chemical structures and associated information and should be complete by March of this year (see http://www.en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemicals#Good_news_-_fact_checking_of_chemboxes.2C_etc and http://www.chemconnector.com/chemunicating/dedicating-christmas-time-to-the-cause-of-curating-wikipedia.html). This is the true value of a wiki environment; mass participation to aid in the generation and the validation of the content.

Wikis have recently been used as the basis of Open Notebook Science [1], a term coined by Drexel University associate professor Jean-Claude Bradley (see http://www.reactivereports.com/51/51_0.html). The UsefulChem Wiki includes a series of experimental pages, for example as shown in Fig. 2 (see http://www.usefulchem.wikispaces.com/All+Reactions) commonly linked to related blog pages (see http://www.usefulchem-experiments1.blogspot.com/2006/05/exp-009.html) and also includes a thesis being written as the research proceeds (see http://www.usefulchem.wikispaces.com/Alicia+Holsey). Bradley has also utilized the wiki medium to challenge the publishing community (see http://www.usefulchem.blogspot.com/2007/04/wiki-paper-experiment-started.html) by writing a paper written on a public wiki and using links to experiment pages on a laboratory notebook wiki as valid references. Commonly, such public exposure of research before publication is frowned upon by publishers, but, as Bradley has pointed out, preprints are already hosted on institutional repositories, so a shift to wiki-based public chemistry, or Open Notebook Science, should not be that different. Differently from most researchers, he prefers to keep a portion of his future plans right out in public view on his 'To Do' list (see http://www.usefulchem.wikispaces.com/pending). He is not alone, as his Open Notebook Science efforts and the movement appears to be gaining momentum with the support of vocal advocates, such as Neylon (see http://www.blog.openwetware.org/scienceintheopen/2007/12/12/a-big-few-weeks-for-open-notebook-science/), Murray-Rust (see http://www.wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=671) and many others.

Wikis are now also being used as a corporate platform for information sharing also. The largest drug maker in the world, Pfizer (see http://www.pubs.acs.org/email/cen/html/090207084512.html), already has Pfizerpedia. Christopher Bouton, a scientist at the company used the MediaWiki platform, on which Wikipedia runs, as his platform. One year after its launch in 2006, Pfizerpedia was receiving 12,000 hits per month by 13,000 individual users across Pfizer worldwide. This pattern is sure to continue within drug and chemical companies, because of the ease of installing, running and collaborating on such a system. With certain extensions to the technology (*vide infra*), the platform could easily become an elec-

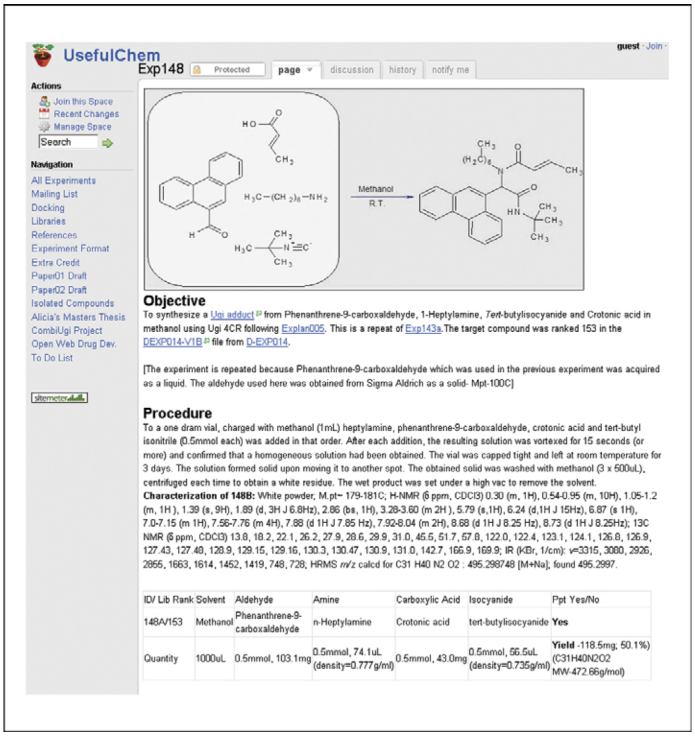

FIGURE 2

An example of UsefulChem wiki page (http://www.usefulchem.wikispaces.com/Exp148). This UsefulChem wiki page shows several important content items: (1) Links to the prior failed experiment. (2) Links to the docking results that justified making this compound. (3) Full characterization (spectroscopy and photographs) of an isolated product, with interactive NMRs (JSpecView/JCAMP-dx) of the starting materials. (4) In the discussion section a question is posed by Professor Bradley to his student, and then answered. The entire discussion history is captured. (5) A complete, detailed and dated log of the steps taken by the student. (6) In the tag section, InChIs of every compound used are provided for indexing by search engines.

tronic notebook platform, but would then be open to legal validation. Organizations will then be left with the choice as to whether to allow them to replace a paper notebook for the purpose of intellectual property protection and as a platform to reference during patenting. This issue too could be addressed with appropriate discussions and initiatives.

Although both wikis and blogs are very valuable for information exchange, what they enable in terms of text and image exchange is all but crippled in terms of searching by many chemists' additional query needs for chemical structures, reactions and data. For example, when a chemist is interested in a particular chemical structure, then the ability to search for that structure in a facile fashion is necessary. Although searching by name might return valuable information (i.e. only from resources where that compound has been named with that specific name) a single compound can have so many identifiers, so many trade names and so on, that such a search could be problematic. In an earlier article in this publication [2] I have discussed Open Chemistry Databases and how Wikipedia is a valuable resource for chemists, but at present this is limited to alphanumeric text searches: generally chemical names and registry numbers. Neither Wikis nor blogs, as yet, are enabled for the purpose of structure and substructure searching and, therefore, remain isolated, in general, from cheminformatics based search procedures. The InChI identifier, discussed in the next section, is starting to facilitate such searches and does offer hope for the future.

The semantic web

The semantic web (SW) is the evolution of the World Wide Web and offers a web where the content is available not only in a natural language format but also in a format to facilitate information sharing and integration via software. A defining article in Scientific American [3] communicated a vision of the web as a universal medium for data, information and knowledge exchange. Six years later, in the same magazine, an article describes the SW in action [4] concluding that SW technologies are already transforming drug discovery and health care. The SW is a philosophy (see http://www.w3.org/2001/sw/Activity), a set of specific design principles (see http://www.w3.org/DesignIssues/), the aggregate result of several collaborative working groups and a variety of supporting technologies. Several formal specifications already support the development of the SW (see http://www.w3.org/2001/sw/#spec) while organizations such as Science Commons (see http://www.sciencecommons.org/) help researchers post data on the Web and provide SW tools for attaching legally binding copyright and licensing information to those data. It is the execution on the vision of the SW, which will drive dramatic changes in how information will be shared and interrelated and this is already being delivered.

One of the key developments, which has already facilitated the SW for chemistry, is the InChI, the International Chemical Identifier (see http://www.iupac.org/inchi/). The InChI string is a textual identifier for chemical substances. It was 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. Chemical substances are expressed in terms of layers of information – the atoms and their bond connectivity, tautomeric information, isotope information, stereochemistry, and electronic charge information. This is illustrated in Fig. 3 (see http://www.en.wikipedia.org/wiki/Inchi and http://www.iupac.org/symposia/conferences/ga05/posters/31_mcnaught.pdf). The InChI string, unfortunately, has only partly delivered on the promise of facilitating web-based searches, because of unpredictable breaking of InChI character strings by search engines. To resolve this issue the InChIKey was introduced (see http://www.iupac.org/inchi/release102.html). The condensed, 25-character InChIKey is a hashed version of the full InChI and is not human-readable. The equivalent InChIKey for the InChIString of L-ascorbic acid is CIWBSHSKHKDKBQ-JLAZNSOCBT. The advantage of the key is one of the enabling web searches, but a lookup table to identify the associated structure, or reference to the original InChI string, is necessary.

FIGURE 3

The InChI string (top) and InChI key (bottom) for L-ascorbic acid.

The possibilities for utilizing this identifier for connectivities across the web have been discussed in detail ([5] and http://www.lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/). Casher and Rzepa have reported on SemanticEye [6], an application to enhance chemical electronic publishing and have utilized InChI identifiers for molecular structures to examine relationships between articles.

Although tens of millions of InChI strings and keys have been populated into databases, their value is still in its infancy. However, publishers have started to embed InChIs into their articles. The Royal Society of Chemistry (see http://www.rsc.org/) is presently pioneering a new publishing model, Project Prospect (see http://www.rsc.org/Publishing/Journals/ProjectProspect/), to demonstrate movement toward the SW for chemistry. The RSC is identifying compounds and ontology terms within their research papers and exposing through RSS feeds (see http://www.en.wikipedia.org/wiki/RSS_(file_format)) to facilitate computer discovery of relevant papers and identification of new compounds. Bloggers have started to use InChI strings and keys on their postings, and wiki-pages are being InChI-enabled to help the web become structure-searchable. The necessity of a central lookup facility for published InChIStrings will be necessary to facilitate substructure searching of the web but this capability is likely to be developed in the near future.

In terms of facilitating connections across chemistry, certain SW technologies are already being brought to bear. Willighagen has been utilizing the Resource Description Framework (RDF) (see http://www.en.wikipedia.org/wiki/Resource_Description_Framework and http://www.w3.org/RDF/) to 'RDF molecular space' (see http://www.chem-bla-ics.blogspot.com/2007/07/rdf-ing-molecular-space.html) and InChI strings are already aggregated onto a blog (see http://www.cb.openmolecules.net/inchis.php). The SW for chemistry is developing quickly and several passionate individuals are contributing. In particular, the Blue Obelisk (BO) group and its members (http://www.blueobelisk.sourceforge.net/wiki/Main_Page) are recognized as major contributors to the area [6]. They believe in the concepts of Open Data, Open Source and Open Standards (ODOSOS) and have vocally participated in the communication of the intention of the SW for chemistry [7–9], as well as in the development of tools to facilitate communication in chemistry. Various members of the BO group have participated in the development and delivery of markup languages for chemistry [10] (CML, https://www.sourceforge.net/projects/cml/), the implementation of InChIs for web-based searching [11], and the delivery of userscripts to aggregate information and computational results from different web resources [12]. Coupled with the collective efforts of many other contributors, developments toward the SW for chemistry are progressing well.

Conclusion

Is the future of modern science to include public exposure and Open Notebook Science on wikis and blogs for all to see-granting bodies, colleagues and competitors? It might appear so. However, it should be noted that the number of participants today in Open Science is but a small fraction of active scientists. Science can be very political and involves funding, career longevity, peer-review and, commonly, the 'old-boy network', so the progress toward a more open model for science may take time. Communication via technologies such as blogs and wikis can also put the traditional journal model and publishers business models at risk and may result in very interesting challenges as has been evident in the anti-PubChem sentiments (see http://www.osc.universityofcalifornia.edu/news/acs_pubchem.html). Publishers' business models are already at risk as a result of Open Access publishing, but some organizations do appear to be successfully navigating these challenges so Open Science will probably flourish despite potential challenges.

Will blogs and wikis influence drug discovery? This is likely to happen primarily by the adoption of wikis and blogs inside a corporation rather than by the sharing of information outside of an organization. As highlighted earlier by the growth in Pfizerpedia, wikis and derivative tools will become more prevalent for collaboration and information sharing. Except for particular cases blessed by the legal groups within a company neither wikis nor blogs are likely to be exposed to the outside world for anything more than general dialog regarding company activities. Certainly anything putting company intellectual property at risk will remain taboo in terms of making available for public consumption and commentary. Although there are popular blogs authored by participants in drug discovery (e.g. in the pipeline), such blogs are commentaries regarding activities in the industry, generic situations and challenges in the laboratory or discuss scientific publications, diseases or politics within the pharmaceutical industry. What these blogs do not expose are the internal politics, projects or activities of an organization.

Although this article has focused primarily on internet-based tools for communication and collaboration in chemistry through the use of wikis and blogs there are several efforts underway to facilitate collaboration by providing access to other capabilities. These include the ability to predict molecular properties online (see http://www.acdlabs.com/ilab/ and http://www.vcclab.org/), perform direct data analysis including the derivation of QSAR descriptors (see http://www.ceccr.ibiblio.org/c-chembench/modelbuilders.do) and the development of web services for a plethora of data analysis purposes (see http://www.chembiogrid.org/projects/proj_ws_all.html and http://www.chemspider.com/blog/?p=268). Such capabilities continue to expand and will greatly facilitate collaborative science. Enabling technologies such as these will continue to be delivered to support scientists, as well as to expose science as it happens. With the recent announcement regarding Semantic Wiki (see http://www.en.wikipedia.org/wiki/Semantic_wiki), and the future development of WiChempedia (http://www.chemspider.com/blog/wichempedia-is-now-on-its-way.html) on a platform to support direct chemical structure and substructure searching, communication and collaboration in chemistry will become easier. With continued efforts by the advocates and evangelists for ODOSOS, interoperability and the SW for chemistry, and the new generation of scientists adopting a more open approach to sharing their data and results we will see dramatic changes in the nature of collaborative science.

Acknowledgements

The author would like to acknowledge the active participation of the scientists in the blogosphere. Without their efforts to improve communication, stimulate dialog and openly share their opinions that space would be far less appealing. I also extend my thanks to Dr Jean-Claude Bradley for our conversations regarding Open Notebook Science and to Dr Martin Walker of the Wikipedia: Chemistry team for encouraging me to participate.

References

1 Bradley, J.C. (2007) *Open Notebook Science Using Blogs and Wikis*. http://www.precedings.nature.com/documents/39/version/1. doi:10.1038/npre.2007.39.1

2 Williams, A.J. (2008) A perspective of Publicly Accessible/Open Access Chemistry Databases. *Drug Discov. Today* 13, 495–501

3 Berners-Lee, T. *et al.* (2001) *The Semantic Web*, Scientific American Magazine – May

4 Feigenbaum, L. *et al.* (2007) *The Semantic Web in Action*, Scientific American Magazine – December. http://www.sciam.com/article.cfm?id=the-semantic-web-in-action

5 Coles, S.J. *et al.* (2005) Enhancement of the chemical semantic web through INChIfication. *Org. Biomol. Chem.* 3, 1832–1834

6 Steinbeck, C. *et al.* (2006) The Blue Obelisk – interoperability in chemical informatics. *J. Chem. Inf. Model.* 46, 991–998

7 Murray-Rust, P. and Rzepa, H.S. (2002) Towards the chemical semantic web. In *Proceedings of 2002 International Chemical Information Conference* (Collier, H., ed.), pp. 127–139, Infonortics

8 Casher, O. and Rzepa, H.S. (2006) SemanticEye: a semantic web application to rationalize and enhance chemical electronic publishing. *J. Chem. Inf. Model.* 46, 2396–2411

9 Murray-Rust, P. and Rzepa, H.S. (2003) Towards the chemical semantic web. An introduction to RSS. *Internet J. Chem.* 6 article 4

10 Murray-Rust, P. *et al.* (2001) Development of Chemical Markup Language (CML) as a system for handling complex chemical content. *New J. Chem.* 618–634

11 Murray-Rust, P. and Rzepa, H.S. (2004) Towards a semantic web for chemistry in lifescience. *W3C Workshop on Semantic Web for Life Sciences*, Cambridge, MA, USA

12 Willighagen, E.L *et al.* (2007) Userscripts for the life sciences. *BMC Bioinformatics* 8, 487 (published online on December 2007. http://www.biomedcentral.com/1471-2105/8/487)