1

Changing ways of sharing research in chemistry

Henry S. Rzepa

Abstract

The challenges of sharing research in chemistry are introduced via the molecule and how its essential information features might be formalized. The review then covers a period of around 33 years, describing how scientists used to share information about the molecule, and how that sharing has evolved during a period that has seen the widespread introduction of several disruptive technologies. These include e-mail and its now ubiquitous attachment, the world wide web and its modern expression via blogs and wikis. The review describes how digital documents have similarly evolved during this period, acquiring in some cases digital rights management, metadata and most recently an existence in the cloud. The review also describes how the dissemination of digital research data has also changed dramatically, the most recent innovation being data repositories, and speculates what the future of sharing research via the latest disruptive technology, tablets, might be.

Introduction

Chemistry is widely considered to stand at the crossroads of many disciplines, with signposts to molecular, life, materials, polymer, environmental and computer sciences, as well as to physics and mathematics, and even art and design. To collaborate and share research data and ideas across these areas, research scientists must strive (and do not always succeed) to find common languages to express their intended concepts. In reality, even different scientific dialects can be a challenge, since the semantics of multi-disciplinary areas of research are rarely

defined accurately or fully enough for people to cope with the ambiguities and subtleties. The modern digital information age has promised a revolutionary approach to these challenges, the latest incarnations being the formation of social networks to facilitate the interaction. This chapter will present a perspective on some of these aspects from my personal point of view as a research chemist. There are many crossroads at which one could stand: I can only follow the sage advice of Yogi Berra: 'When one comes to a fork in the road one should take it'![1] Here I will take the fork to molecular sciences.

My starting-point is a molecule. There are about 65 million[2] that have been formally shared by scientists (and quite a few more that may not have been shared). How do we know this? Well, in the mid-19th century an enlightened scientist called Konrad Beilstein decided to create a molecular taxonomy. Far fewer molecules were known then, of course, but he had the vision to realize that their number was going to grow, very probably exponentially. This was because more and more chemists were sharing 'recipes' or protocols for producing new molecules and developing ways of describing the properties of the new entrants. Beilstein's taxonomic project was based on three steps:

1 identify a new molecule
2 classify its measured structural characteristics and properties
3 identify the researcher(s) who reported these properties (what we now know as a literature citation).

Nowadays, almost all aspects of this project are conducted with the help of digital tools. The first task is to formally convert the structure of a new molecule into a digital expression. This expression is called a *connection table* and attempts to define which atoms in any molecule are connected to other atoms in the same molecule. The molecule may in fact also comprise several components unconnected by bonds but nevertheless inseparable. It soon became clear that molecular scientists needed to define more carefully what they meant by a connection; and in fact they came to call this a (chemical) *bond*. It sounds simple enough: all that molecular scientists need to do is agree among themselves what a bond is. The first person to attempt this in modern terms was G. N. Lewis, in a famous article published in 1916[3] outlining how a bond could be defined in terms of

shared electrons. The development of quantum mechanics in the 1920s allowed these apparently simple definitions to be formalized mathematically, and physics played its part by showing how X-ray crystallography might provide experimental measurement of such bonds. There are of course grey areas, especially nowadays when odder and odder bonds are continually being discovered and require constant refinement of the definitions of a bond. But by the time that the modern digital era started in the 1960s, connection tables for essentially all new molecules could be produced. The task was so gargantuan, however, that only a small number of commercial organizations could afford the resources (mostly human at that stage) to do this.

Armed with a connection table, a research scientist was in a position to contemplate formulating a search for a specific molecule about which others had shared information. Until around 1979, scientists had to visit in person what was often the one library at their organization with the shelf space to store the collected chemical indices, and systematically hunt through each five-year volume on the basis of a systematic name. That name was derived from the connection table, and part of the chemist's skill was the ability to infer such names from a set of rules acquired during training. From my own experience, I can vouch that it would take around five hours to find research information on a molecule such as that represented (in two dimensions) in Figure 1.1. In reality, even for a relatively simple system, deriving its systematic name was often too great a challenge. Instead, the community would refer to it by what was known as a trivial name, information to be acquired by (sometimes serendipitous)

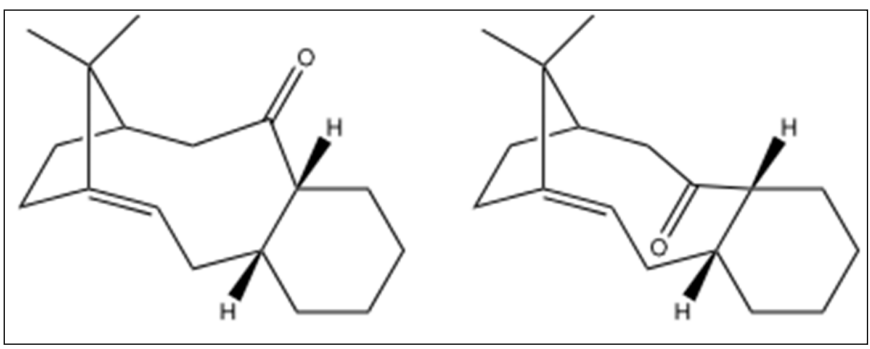


Figure 1.1 A molecule (and a stereoisomer) represented by a connection table, indicated by lines representing bonds

reading of books and articles, or indeed by talking to colleagues. And it would be difficult to stray too far from one's own specialist area.

However, from 1979 onwards access to such world collections of molecules started to go online, and an institution's library could now be expected to offer an online searching service. This involved booking an appointment with a specialist librarian (with around two weeks' notice, due to heavy demand). The librarian would be trained to understand how to formulate the required search syntax in such a diagram. For the first time, the scientist could realistically expect to search all the information on known molecules, rather than just the small subset determined by the amount of time he had available. Moreover, he could formulate a search based on a degree of similarity, rather than on exact matches, and so be far more adventurous in his searches.

At this stage a molecular query was formulated in terms of integer connectivity, such as 1 = a bond connection (mapping to one line in the above diagram) and 0 = pairs of atoms with no such connection. For many scientists, this description seemed too restrictive, and so in the early 1970s a project was initiated to collect and share the experimental data from which the bond connectivity could be quantified as a length. This initiative became in time the Cambridge Crystallographic Data Centre, which nowadays disseminates information on around 600,000 molecules on a commercial basis. It provides accurate 3D co-ordinates for all the atoms in a molecule, which allows much richer information about them to be inferred. I use the word *disseminates*, which does not mean quite the same as *shares*. The distinction relates to the difference between open and closed sharing of research, to which I shall return several times in this chapter.

By the early 1980s a more general scientific online presence was emerging, and central libraries no longer held the monopoly of access points to such information. Indeed a typical researcher might have one access point in his own building, maybe even reasonably near his office. In 1985 we reached another important fork in the road. Most individual researchers now regarded the software tools to describe molecules digitally as essential. This opened up a new paradigm for sharing chemistry. The date is quite specific, since it corresponds to the introduction of the Macintosh personal computer and two of the tools, in particular, that appealed to visually oriented chemists (including those who had difficulty naming molecules, see below). The first was

a mouse-driven sketching tool that could be used to represent the molecular connection table pictorially (and is still used to this day to draft diagrams such as shown in Figure 1.1). The second was the ability to transmit the diagram to a high-quality laser printer via a computer network. This network was built so that the cost of the printer (in those days a more expensive resource than the personal computer itself) could be shared among many researchers.[4] Although few realized the significance at the time, a spin-out benefit of the creation of such a network enabled something far more world changing than simply connecting a computer to a printer. Adding a so-called network router to the system also enabled an individual user's computer to be connected (via the somewhat unlikely printer port) to two new resources for sharing: e-mail and the then nascent internet. This in turn introduced researchers to entirely new paradigms for sharing their research and collaborating with others. A pictorial representation of the molecule (the natural language that had developed following Lewis's definition of the bond, see Figure 1.1) could now be immediately shared with any other researcher in the world with access to similar resources – admittedly in 1985 not very many. I say immediately, because the process involving the conventional, journal-based way of sharing was at that time often taking two years from start to finish, hardly an immediate process.

There was still another problem to be resolved: how to share the underlying data used to generate that picture. While e-mail was starting to allow two or more people to exchange information without delay, it was not yet recognized that reuse in a machine/software sense was also desirable.

E-mail as a content delivery mechanism for sharing research

E-mail became an increasingly popular tool for most scientists from 1985 onwards. This is significant because it introduced two components for dialogue: a loosely structured natural-language discourse and the document *attachment* facility associated with the process. The latter was a way of 'shrink wrapping' research ideas based in a standard document format, and particularly of the research data

underpinning the ideas. Defined by a standard specification called the MIME type, it allowed a labelling of the document to ensure that the recipient's e-mail program could correctly process what it received.[5] In fact, this allowed the research scientist (in principle anyway) to share his research data with others in a manner that would allow the recipient to invoke the appropriate software so as to add further layers of semantic meaning to the received data and information. In practice, this feature was never fully exploited with e-mail: to this day the MIME label is used to 'wrap' a relatively limited set of document types, such as word-processed files, numerical spreadsheets, graphic images and a format known as PDF or portable document format, itself a spin-out from the Postscript printer description introduced with the first laser printers.

Chemists (especially those whose activity centred on molecules) would more often than not share their research by simply sending an attachment comprising a chemical document to others. The recipient still had to put in informed effort to ensure the attachment was compatible with his particular computer and software. The document itself normally ended up in a non-hierarchical folder called 'attachments', with little information about its content available, because senders were not constrained by any particular naming convention for the file name. The 'metadata' describing these e-mail attachments are sparse (once they are ensconced in the attachments folder, their association with the MIME type is lost). The process of rescuing such information has been memorably described by scientists as *defrosting the digital library*.[6]

The web as a content-delivery mechanism for sharing research

The document deluge was about to be greatly increased by the next wave of mechanisms for sharing research. Starting in 1994, most of the world's scientific research publications and journals undertook a gradual journey online, promoted largely by the exponential adoption of the system known as the world wide web.[7] From the outset, it was apparent that this mechanism had rather different attributes, as compared with e-mail, for sharing information and data.

1 Whereas e-mail was a 'push' mechanism initiated by the content holder, the web was a 'pull' mechanism initiated by the content requester (which could be either a human or a software agent). The difference is subtle but important, in that it led directly to the era of the search engine.
2 The pull request was made using a standard known as a uniform resource locator (URL), a now familiar term.
3 Although URL itself was standard, it soon proved not to be permanent over a time-scale of years/decades. A mechanism known generically as a 'handle' was introduced to solve this problem. The handle system[8] was designed as a more permanent mechanism to identify a document, with the handle being resolved into a URL at the time of the request. The best-known implementation of such handle resolution is the DOI (digital object identifier).[9] Since 2005, virtually all publishers of scientific journals have fully implemented this mechanism (at the time of writing, 52,678,814 DOIs had been assigned).[10]
4 As a result of the wide adoption of these standards, researchers now tend to exchange these DOI identifiers between themselves, citing them in e-mails, in documents (both word processing and PDF), in web pages and embedded in other recent expressions of web pages such as blog, wikis and podcasts (see below).
5 Although most web-based journal articles may ubiquitously have an associated DOI, the naming conventions used for the DOI itself tend to be publisher-specific, often inscrutable, and in themselves tell little about the content of the article. Many documents, particularly those not associated with collaborating publishers, do not have such a unique handle.
6 In 1994 it was recognized that the MIME mechanism, already matured in the e-mail environment, could also help to identify the context of a web-based document. In the area of chemistry, specifically, a Chemical MIME label was introduced.[5] Some 50 types of chemical document were identified, a taxonomy that helped to define the types of data available to chemists.

These mechanisms allowed documents to be linked into web pages, and shared with chemists in a reusable manner. This concept introduced a

differentiation between *discourse*, of which the prime carrier was the journal article, and *data*, which itself could be contained in a document associated with a MIME type. The former was focused on the human reader, while the latter had a structured and standard form intended to be reused in conjunction with computer software. The latter could be used to transform the data into a visual representation to help scientists in their quantitative assessment of their models and associated interpretation, or as the input into further numerical analysis and model building.

The document type as a container for shared research

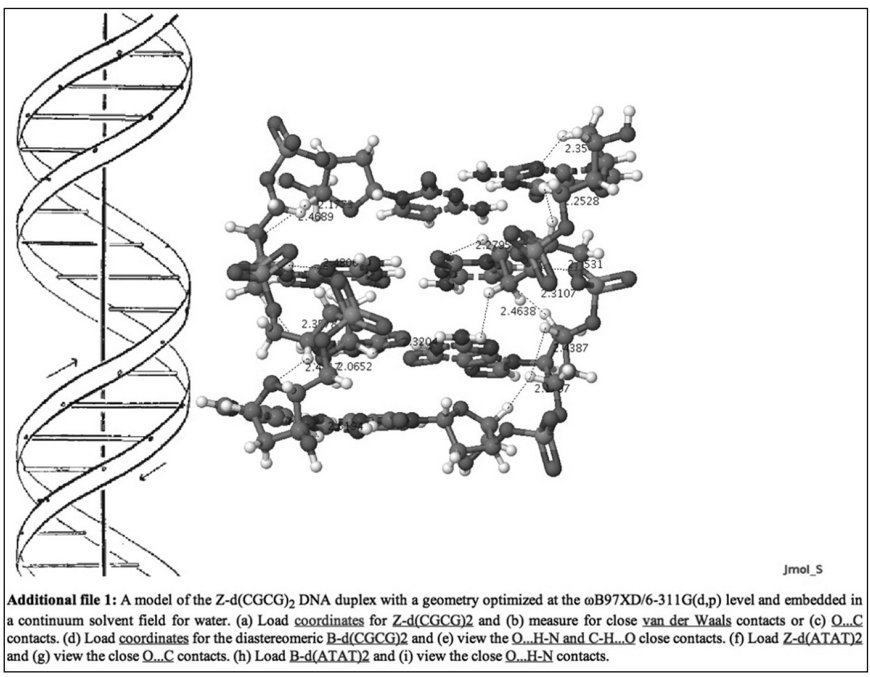
As the web was becoming established as the pre-eminent mechanism for delivering journal articles to their readers, so the wrapper for that content settled down into two principal digital formats. First, the portable document format or PDF (nowadays also generically referred to by the proprietary name Acrobat) represented essentially a printable version, which in appearance emulated the traditional look and feel of the bound journal article, complete with pagination, headers and footers. We might describe this as a format where the content and its style of presentation, its look and feel, are tightly integrated into a sealed and largely tamper-proof container.

While the full text could now be digitally searched within the PDF document, it was not designed as an innovative format departing radically from its heritage of the printable page. One exception to this, of potential interest in the molecular sciences, was the introduction of a 3D enhancement to Acrobat. This allowed models containing 3D model co-ordinates to be embedded into the document. This, in turn, enabled interactive rotation in order to change the viewing angle of the object. Creating such documents is complex, and few scientists have chosen to share their research in this manner to date.[11] Moreover, this could be also regarded as a limitation to sharing since the model contains no accessible (structured) underlying data. In other words, it is a passive object suitable only for viewing by a human, but not for reprocessing or reusing in the manner appropriate for a scientific investigation.

Second, in parallel with the PDF document, research could also be shared by the publisher through the adoption of an HTML presentational format. Most journals offer both formats for their readers. In practice, the

HTML is generated automatically by a production workflow originating from a word-processor document and the original authors have little participation in its generation. Although in principle HTML, as a mark-up language, offers a non-proprietary and interactively rich environment for sharing research, the lack of author involvement in its preparation has limited the 'added value' of this mode of presentation. However, I would argue that more general mark-up languages have much, as yet unexploited, potential for enhancing the sharing of research.[12] The basis for this assertion is a series of experiments that we undertook to demonstrate this.[13] Figure 1.2 is just one example of such enhancement.

Of course you are viewing this figure in black and white, statically on the pages of a book: the original is in colour and fully rotatable and interactive. At the bottom, the original caption contains hyperlinks to scripts adding annotations in the form of measurements of a 3D object (a portion of the DNA molecule, in this example) or links to additional data. Such 'added value' can be accessed only through the original journal



Additional file 1: A model of the Z-d(CGCG)$_2$ DNA duplex with a geometry optimized at the ωB97XD/6-311G(d,p) level and embedded in a continuum solvent field for water. (a) Load coordinates for Z-d(CGCG)2 and (b) measure for close van der Waals contacts or (c) O...C contacts. (d) Load coordinates for the diastereomeric B-d(CGCG)2 and (e) view the O...H-N and C-H...O close contacts. (f) Load Z-d(ATAT)2 and (g) view the close O...C contacts. (h) Load B-d(ATAT)2 and (i) view the close O...H-N contacts.

Figure 1.2 An example of an enhanced figure as part of a scientific article. The diagram is fully rotatable and interactive

page. Unlike with an Acrobat 3D object, the user also has access to this data in the mode illustrated in Figure 1.3 and thus has a portal into further research exploration.

Such enhanced attributes of a journal article, however, raise an important new issue. Conventionally, most scientists and chemists are assumed to be familiar with a fairly standard set of tools that they use to share their research: a word processor and (for, e.g., chemists) the chemical structure drawing program (see Figure 1.1). These tools are, however, limited when it comes to handling the data that is so essential for enhancing an article in the manner shown in Figures 1.2 and 1.3. Few authors acquire the necessary skills, and it might be said that few have the motivation needed to handle such data. It may also transpire that incorporating enhancements such as are shown in Figure 1.2 into the journal production workflow might in turn result in greatly increased costs to both the author and the institutional library, in the form of increased subscription charges. We may get a glimpse of this in how many publishers already surcharge authors for incorporating colour plates into their discourse, or for making an article available via an Open Access (OA) licence. The article from which Figure 1.2 is derived is OA, and that in turn allows me to include a representation of it in this chapter.

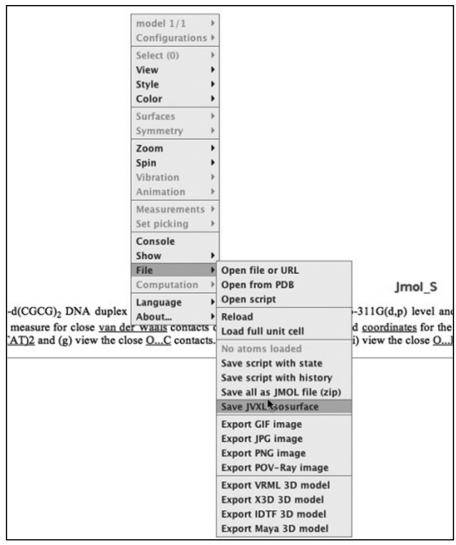


Figure 1.3 An illustration of how research data can be extracted from an enhanced journal figure

The need for interactivity has, however, emerged from a different direction. Around 2005, electronic books, or e-books, started to have a significant commercial impact. Devices such as the Amazon Kindle or the Apple iPad began to demonstrate how portable electronic access to bookstores can transform a market. Although both devices come with their own proprietary format, a more open format, known as *epub* has also emerged.[14] This is in fact nothing more than the aforementioned HTML

wrapped into a compressed bundle and described by a manifest. The latest specification, epub3, adds the element of interactivity possible with, for example, Figure 1.2, and this in turn is based on the latest HTML standard, known as HTML5.[15] Even the presentation of the conventional static diagram is evolving. Images and diagrams are traditionally included in HTML documents, using bit-mapped formats such as JPG, and this is how most scientific journals present them to their readers. (It is also how Figure 1.2 was created and incorporated into this chapter.) But such a non-scalable image is not optimum for new generations of portable mobile e-book readers, which introduce the 'pinch-and-zoom' gesture, allowing instant magnification. A suitably scalable image format, an 'HTML for images', such as SVG (scalable vector graphics) is now more appropriate. Apple has also launched a rich interactive authoring environment (iBooks author)[16] which introduces a much more data-centric metaphor, as compared to the traditional word processor. As such tools mature, we may expect that scientists will be induced to use them in creating journal articles. Whilst in 2013 no journal or book publisher accepts submissions in such a format, we should look out for future developments with interest.

The importance of organizing the content and metadata

In the previous section, I reviewed how scientists and publishers had found the web an easy-to-use interface to search for information in journals, databases and other sources, and discovered how to use it to download documents to their own local computer for further reuse or analysis. As they did so the need for local capabilities to organize this content became increasingly apparent. Here I focus on a type of tool that emerged around 2008 for assisting this process, since it illustrates the increasing (and welcome) adoption of metadata as a content-organizing tool. This problem had in fact been already addressed in quite a different context: the music industry. In the early 2000s, the technology of digitally downloaded music was reinventing a creative industry in many ways not unlike scientific publishing in 2013. Apple Computers introduced iTunes as a new metaphor for a personal music library, and with it the concept of metadata to describe the attributes of the music (artist/author, date released/published, album/publisher, genre/scientific field etc.). One could in general copy a music track from a 'legacy device' (a music CD),

drop it into the iTunes library and then go online to acquire further metadata (including, e.g., album art and video). Playlists could be used to define a subset of the music, and copied onto portable listening/viewing devices for the listener's convenience.

Mendeley is an example of a program that adapted this music metaphor to scientific publishing and sharing contexts. Scientific articles, downloaded from a publisher's journal site, can be 'dropped' into the Mendeley article library. This triggers analysis of the metadata attributes of the article, either by pattern scanning to identify bibliographic information such as the authors, the title and so forth (succinctly summarized by the Dublin Core metadata schema)[17] or by inspection of any explicit metadata defined within invisible fields in the document itself. We have ourselves already described how such a harvesting process, using metadata stored directly within an Acrobat file as so-called XMP, can be aggregated and queried in a chemical context.[18] Programs such as Mendeley, which implement much of this concept, offer much more than just a convenient container for a personal library of scientific articles. Such an activated library can be used most simply in conjunction with a word processor as a citation and bibliographic tool when authoring new articles. A more innovative feature of Mendeley is that a selection of articles and the associated metadata can be uploaded to the user's online account and their metadata compared with the 'crowd sourced' content from other Mendeley users. This provides a seamless mechanism for identifying other scientists who may have published on similar topics. One can share such 'playlists' of articles with students and colleagues. Here, however, we see the first signs of the phenomenon of copyright assertion and digital rights management and the associated restrictions that this imposes upon the sharing of research. The implications are expanded upon below. Scientific playlist generation (scientists like to call these their publication lists) can even be automated: Symplectic Elements[19] is a software system that automatically garners all the scientific publications produced by an organization such as a university and organizes them according to the detected metadata. An individual scientist's personal publication record is automatically produced for them, and the system will even generate an h-index[20] as one purported metric of the esteem in which they are held by their colleagues.

The cloud and DRM

I have described above how the online metaphor has evolved between around 1995 and 2010, largely to replace the physical library as the primary mechanism for scientists to have access to shared research. Instead, scientists nowadays build their own personalized digital libraries on physical devices such as desktop computers, organized using metadata to help discoverability. Yet again we might look to the music industry to see how this metaphor might evolve. Most people now have multiple devices on which they can access content, ranging from static desktop computers to smaller portable laptops and to the always-on mobile device. There is no reason why scientists should not access their shared research in a similar manner across this entire device range. A concept known as 'the cloud' has evolved to deliver that content. At its simplest, this removes the user's local computer and storage from the centre of the hub, storing the content, in effect, on a central server-farm. The user purchases or inherits access rights to this content, which can, optionally, be encoded using a mechanism known as digital rights management (DRM).

While the DRM model is currently applied to creative content such as music, video and other forms of entertainment, there are signs that scientific journal articles are also now seen as belonging to the creative industries and subject to the copyright laws that apply to such industries. One such model already operating is known as *Secure Electronic Delivery (SED)*, from the UK British Library. A journal article can be delivered directly to the reader by e-mail as a DRM-enabled PDF attachment. This currently imposes some interesting restrictions.[21]

The recipient:

1 is allowed to make only a single paper copy of the article (it is not clear how enhancements such as Acrobat3D[11] could be invoked on paper), from which they may not make any further paper copies
2 may not convert the file into any other format
3 may not cut and paste or otherwise alter the text
4 may not forward the file to anyone else
5 and after printing the electronic document (once), must then delete it.

The significance of this particular DRM model is that, since the only permitted action upon the received document is to print it (once), it cannot

be submitted to a program such as Mendeley to reap the benefits of metadata harvesting. Likewise, it would not be possible to harvest any (digital) data components of the document for electronic reuse (as data). The digital life-cycle or 'ecosystem' in such a model is, in effect, permanently destroyed. This is perhaps an extreme example of how a cloud-based, DRM-protected model may achieve little by way of sharing scientific research. Mechanisms such as this illustrate how critical the delivery mechanism will be to preserving the value of shared information, and how some models may be entirely inappropriate. For example, consider the article[13] in which I discussed how the data associated with that very article might be accessed and reused by readers. If such an article were to be DRM protected, such data components would be likely to be imprisoned by (i.e. to inherit) the DRM applied to the article as a whole, even though the data itself might not be covered by any copyright. Alternatively, one could envisage the different components of an article each having different degrees of DRM, and that this might differ from journal to journal, or between publishers. Would the original authors of an article and its data-based components have any control over how the article was accessed by its readers, via an open access buy-in or other mechanism? It is impossible to predict the answers to these questions, but they demonstrate the challenges ahead, and our need as scientists to keep as much of the world's shared scientific knowledge open as is possible.

The importance of data

The preceding discussion leads us to ask whether journals are still the best medium in which to place data intended for sharing. Data curation has traditionally been largely neglected by scientific journals. When those journals were exclusively printed, the additional (printing) costs of including an appendix or annex with the data frequently precluded its inclusion. Instead, the journal might encourage readers to contact authors directly for such information (assuming they were still contactable). The authors themselves then had to solve the problem of transferring the data into usable form. When electronic dissemination of journals started, authors were asked to include the data in a form that became known as ESI, or electronic supporting information. This was often presented as a single, monolithic PDF file containing a mixture of visual elements, and

tables of numbers intermingled with page footers and headers and other non-data. The task of adding semantics to the data fell to the (knowledgeable) reader. Unfortunately, a PDF document is a poor carrier of semantic information and data, and the irrelevant information present in such a document often made copying numbers out of a table an arduous task. That situation largely persists to this day.

Digital repositories

One solution to this problem has emerged in the form of digital data repositories.[22] These differ from the ESI/PDF formats noted above in several key regards:

1 They are OA; no institutional or personal subscription is required.
2 They carry formal and often complete metadata. This includes a date stamp that clearly shows when the data was deposited, with an assurance that it has not been subsequently modified. The metadata itself can be generated automatically from a scripted workflow, ensuring that it is error free and freeing the researcher from the otherwise often onerous task of manual insertion.
3 They carry provenance, in particular the name of the person who deposited the information.
4 They have an associated handle that can be quoted elsewhere, and, as with a DOI, it allows one-click access to the data.
5 The metadata can itself be searched; the data is easily discoverable.
6 The data collection can contain other appropriate identifiers. Thus, data associated with a specific molecule can have a derived and unique identifier known as an InChI key.[23] A digital repository provides an alternative to the scientific journal for scientists to share their data with others, and also a convenient method of claiming priority and ensuring provenance for the data. The handles (DOIs) for this information can themselves be inserted into tables, figures and other components of a traditionally published article. However, the use of a digital repository places the burden of creating this resource upon the scientists themselves, an infrastructure that many may not be willing or able to install.

Recently, however, open services such as Figshare[24] have started to provide an alternative. Among the claimed advantages of such an open repository are the following:

1 All deposited research data is citable (with a DOI).
2 It is cloud based (secure and accessible from anywhere).
3 It is taggable and easily filtered, making the research (data) easily found.
4 Negative results, traditionally difficult to publish in conventional journals, can be archived.
5 Private collaborative spaces to support projects between groups and scientists are available.
6 An API for programmers to interface with their own software is provided.[25]

While the use of digital repositories in this way is not yet common, it is expected to increase in the future.

Social networking mechanisms for sharing research

I have so far focused only on the scientific journal article as the mechanism that most scientists have traditionally used to share their research (another, the scientific conference, should also be noted, even if not here discussed). A highly respected figure in the field of chemistry, Whitesides,[26] urges scientists to explore other mechanisms for making their research shared and accessible, suggesting for example the addition of temporal components such as animations and movies (where appropriate) and that journals routinely support such features (note again the departure from the traditional printable format). It is clear that he also expects the scientific paper to evolve and change even further in the near future. Here I briefly explore the blog (= weblog) as an interesting new addition to the mechanisms for sharing research.

The blog

The first blog appeared in 1999 as an easy procedure for writing a web page. The facility for readers to leave comments is an important part of

many blogs. In 2013 this medium is now considered mature and is increasingly being adopted by both individual scientists and publishers as a means of both sharing their research and leaving opinions on that research. The blog can also be a rich carrier of data, and the two can be seamlessly merged into an attractive and enriched environment.[27] Some of the features that make it so include:

- support for citation management[28] and metadata harvesting by means of extensions that, e.g., can resolve a DOI (as defined above) into bibliographic metadata about the article referred to (most journals also include this feature in their production workflow);
- a suitable environment for expressing and rendering mathematical equations within a post;
- support for scalable graphics formats such as SVG (as described above);
- a rich environment for expressing and rendering molecule displays in both two and three dimensions (equivalent to Figure 1.2 above);
- style sheets for customizing the blog for optimal display on mobile devices and tablets;
- functionality that can *chemicalize* a blog. This is a way of identifying chemical terms and molecules contained with a post, and linking these to *pop-ups* that automatically translate, e.g., a chemical name to a chemical structure (Figure 1.1) or to a concept, to further explanation;
- statistics that provide information on post views and search engine terms used to find the post, and which indicate the impact of the blog;
- facility for instant publication;
- permanent archive using services such as WebCite.[29]

The blog provides a mechanism for a single author to share research and ideas. I have used this as a conduit for both my teaching and research activities for some four years now, during which time around 225 posts on diverse topics have appeared. This might be contrasted with my career total of some 330 peer-reviewed articles in scientific journals over a 40-year period. These numbers, of course, imply that the two genres are indeed rather different. A criticism often made of blogs is that they are not

peer reviewed, although this can be countered by the observation that posts can attract open comments (peer review is, after all, a closed process) from the community. It is this very feature that improves the science; a commentary on a post can either evince a response from the original poster, or indeed lead to a fully blown conventional article published in a traditional journal. In turn, this article can itself lead to commentaries on other blogs, thus completing the cycle. Seen in this light, the blog post becomes an integral part of the scientific cycle of sharing.

It would, however, be fair to say that most scientists would currently hesitate to use a blog as their primary mechanism for sharing research. Its strengths lie in commentary and discussion of articles found in journals (a form in fact adopted by many publishers who wish to attract a readership) as well as in its being a medium for reporting original research in conjunction with the use of digital repositories. Blogs also have a major pedagogic element, where modern developments are discussed and interpreted for a younger audience of students. Personally, I also find it a suitable medium for sharing whatever experience and knowledge I have acquired over my own career.

The wiki

Like blogs, wikis were initially envisaged as a simple way of creating and sharing a web-based article, albeit with a low learning curve for the authoring process. However, they came to public attention in the form of Wikipedia, a shared compendium of human knowledge with articles authored by more than a million contributors. Once it was adapted to carry rich, reusable chemical information and data[30] (in the manner[31] described above for blogs), we have also found that it is a popular medium for chemistry students to communicate their coursework.[32]

Conclusions

My review covers a period of around 30 years, a small fraction of the time since the first scholarly scientific journals were launched in 1665[33] to share research. During this period we have moved from the institutional or society library as the principal way to deliver printed journals to the research chemist, to a much more complex online environment. Printed journals are

by definition not interactive, and the cost of their production limits how much content can be shared on their pages, a limitation that frequently precludes the inclusion of full experimental information and data. In the electronic medium, these and other boundaries are largely removed. I hope that I have given a glimpse of the medium's rich new potential.

Along with this potential come many challenges to be solved. We have barely begun to address the restrictions of, e.g., DRM, and there is a clear need to encourage and educate researchers and teachers to share their science armed with this bewildering array of new tools. Were this review to have been written a mere 15 years into the future, its outcome and format would doubtless have been quite different (you would be unlikely to be reading it as a printed book, for example). Many of the mechanisms outlined above will have been replaced; perhaps even the written word itself will have been largely superseded by the spoken word. But I end as I started, with another apposite quotation from Berra: 'The trouble with our times is that the future is not what it used to be.'[1]

References

1 Berra, Y. (2010) *The Yogi Book*, Workman Publishing. See also www.yogiberra.com/yogi-isms.html.

2 The front page at www.cas.org/content/counter recorded 65,457,839 molecules on 7 March 2012, increased to > 68 million by 4 September 2012.

3 Lewis, G. N. (1916) The Atom and the Molecule, *Journal of the American Chemical Society*, **38**, 762–785. doi: 10.1021/ja02261a002.

4 Rzepa, H. S. (2011) *Computers 1967–2011: a personal perspective. Part 2. 1985–1989*, www.ch.imperial.ac.uk/rzepa/blog/?p=4578. (Archived by WebCite® at www.webcitation.org/65zW2zdhS.)

5 Rzepa, H. S., Murray-Rust, P. and Whitaker, B. J. (1998) The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World-Wide Web information exchange, *Journal of Chemical Information and Computer Science*, **38**, 976–82. doi: 10.1021/ci9803233.

6 Hull, D., Pettifer, S. R. and Kell, D. B. (2008) Defrosting the Digital Library: bibliographic tools for the next generation web, *PLoS Computational Biology*, **4** e1000204. doi: 10.1371/journal.pcbi.1000204.

7 Rzepa, H. S., Whitaker, B. J. and Winter, M. J. (1994) Chemical Applications of the World-Wide-Web, *Journal of the Chemical Society, Chemical Communications*, 1907.
8 *About the Handle System*, www.handle.net/factsheet.html. (Archived by WebCite® at www.webcitation.org/65zWFosJi.)
9 Paskin, N. (2005) Digital Object Identifiers for Scientific Data, *Data Science Journal*, **4**, 12–20. doi: 10.2481/dsj.4.12.
10 CrossRef.org, www.doi.org/factsheets/DOIKeyFacts.html.
11 Kumar, I. P., Ziegler, A., Ziegler, J., Uchanska-Ziegler, B. and Zeigler, A. (2008) Grasping Molecular Structures through Publication-integrated 3D Models, *Trends in Biochemical Science*, **33**, 408–12.
12 Murray-Rust, P. and Rzepa, H. S. (1999) Chemical Markup Language and XML: Part I. Basic principles, *Journal of Chemical Information and Computer Science*, **39**, 928. doi: 10.1021/ci990052b.
13 Rzepa, H. S. (2011) The Past, Present and Future of Scientific Discourse, *Chemoinformatics*, **3**, 36. doi: 10.1186/1758-2946-3-46.
14 Webb, J. *What to Expect in EPUB3*, http://radar.oreilly.com/2011/01/epub3-preview.html. (Archived by WebCite® at www.webcitation.org/5wo1jdC3L on 27-02-2011.)
15 Lawson, B. and Sharp, R. *Introducing HTML5*, http://introducinghtml5.com/. (Archived by WebCite® at www.webcitation.org/5wo1yKHbs on 27-02-2011.) See also http://dev.w3.org/html5/spec/Overview.html. (Archived by WebCite® at www.webcitation.org/5wo1o6RPx on 27-02-2011.)
16 *Apple Computer*, iBooks Author, www.apple.com/ibooks-author/. (Archived by WebCite® at www.webcitation.org/65zWoElsn.)
17 Dublin Core Metadata Initiative, http://dublincore.org/.
18 Casher, O. and Rzepa, H. S. (2006) SemanticEye: a Semantic Web application to rationalise and enhance chemical electronic publishing, *Journal of Chemical Information Models*, **46**, 2396–411. doi: 10.1021/ci060139e.
19 Symplectic, www.symplectic.co.uk/.
20 Hirsch, J. (2005) An Index to Quantify an Individual's Scientific Research Output, *Proceedings of the National Academy of Sciences*, **102**, 16569–16572. doi: 10.1073/pnas.0507655102.
21 British Library Document Supply Service, www.bl.uk/reshelp/atyourdesk/docsupply/help/receiving/deliveryoptions/electronic/sed/sedfaq/index.html. (Archived by WebCite® at

www.webcitation.org/65zX6HyeC).

22 Downing, J., Murray-Rust, P., Tonge, A. P., Morgan, P., Rzepa, H. S., Cotterill, F., Day, N. and Harvey, M. J. (2008) SPECTRa: the deposition and validation of primary chemistry research data in digital repositories, *Journal of Chemical Information Models*, **48**, 1571–1581. doi: 10.1021/ci7004737.

23 See www.iupac.org/home/publications/e-resources/inchi.html.

24 Hahnell, M., http://figshare.com/.

25 Rzepa, H. S., *Digital Repositories. An update*, www.ch.imperial.ac.uk/rzepa/blog/?p=7290 (Archived by WebCite® at www.webcitation.org/6AQKiyn3w). An example of such an automated deposition can be found at doi: 10.6084/m9.figshare.93114.

26 Whitesides, G., www.youtube.com/embed/NHuC5yZeHYQ.

27 Rzepa, H. S. *The Blog Post as a Scientific Article: citation management*, www.ch.imperial.ac.uk/rzepa/blog/?p=6341. (Archived by WebCite® at www.webcitation.org/65zXdiHXL.)

28 See, for example, http://knowledgeblog.org/kcite-plugin, and http://knowledgeblog.org/kblog-metadata.

29 Webcite, www.webcitation.org/.

30 Walker, M. A. (2010) Wikipedia as a Resource for Chemistry. In Belford, R., Moore, J. and Pence, H. (eds) *Enhancing Learning with Online Resources, Social Networking, and Digital Libraries*. ACS Symposium Series, **1060**, 79–92. doi:10.1021/bk-2010-1060.ch005.

31 Rzepa, H. S. *Jmol and WordPress: loading 3D molecular models, molecular isosurfaces and molecular*, www.ch.imperial.ac.uk/rzepa/blog/?p=8. (Archived by WebCite® at www.webcitation.org/660NLRmSP.)

32 Rzepa, H. S., Bearpark, M. J., Armstrong, A. and Hunt, P. Activating Computational Chemistry via an Online Presence, *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, 22–26 March*.

33 Hooke, R. (1665) An Accompt of the Improvement of Optick Glasses, *Philosophical Transactions*, **1**, 2–3. doi: 10.1098/rstl.1665.0003.