Publication at Source: Scientific Communication from a Publication Web to a Data Grid

J. G. Frey

Department of Chemistry, University of Southampton, Southampton, UK

D. De Roure, L. Carr

Department of Electronics and Computer Science, University of Southampton, Southampton, UK

Abstract

The Web has been used as a medium to facilitate the publication and dissemination of articles which summarise and interpret the outcome of scientific experiments. In this paper we focus on Chemistry to argue that the Grid will facilitate publication at source, the dissemination not only of experimental data, but also contextual information about the conditions, setup, running and ongoing analysis of experiments.

1. BACKGROUND

The accepted role of scientific and scholarly publication is to record research activity in a timely fashion, keeping others in the research community up-to-date with current developments. Until very recently, it has been the case that printed journals and conference proceedings were the most efficient method for the dissemination and archival of research results. Technical advances in the past decade have allowed the process of scholarly communication to take other forms, particularly in the dissemination and storage of articles via the World Wide Web.

It is worthwhile pausing to consider that technological proposals for improving the dissemination of scientific knowledge have been suggested for some sixty years. Immediately prior to the Second World War, the novelist and scholar H.G Wells proposed a microfilm-based index to all human thought and knowledge [9]. The experience of co-ordinating thousands of American scientists during that war led Vannevar Bush to propose a similar system complete with what we would now call "hypertext links" [2].

The technical advance which made this possible in reality was, of course, the World Wide Web. Developed at CERN to facilitate "instantaneous information sharing between physicists working in different universities and institutes all over the world" it gave publishers a new medium for making their journal archives available [5]. It also gave authors the means to break the so-called "Faustian bargain" and directly distribute their articles in pre- or post-publication form from their own Web pages [4] or in organised "eprint archives" [3].

However, it was not only the technical advance provided by the printing press in the late 15th century but the emergence of a reliable postal system and the development of the experimental method in the 16th century that led to the production of the first Scientific Journal in 1665 [8]. Similarly, it may not be simply the technical ability to reproduce and distribute articles electronically (e-publishing), but also the emergence of highly collaborative, large-scale investigations and analyses (e-science) that is likely to lead to significant change in the field of scientific communication and significant changes in the way such communications are produced, curated and disseminated [7].

2. THE MULTIMEDIA NATURE OF CHEMISTRY

Chemistry is a multimedia subject, 3D structures are key to our understanding of the way in which molecules interact with each other, dynamic images and movies are now required to portray adequately the chemist's view of the molecular world. This observation may dramatically change expectations of what a journal will

provide, but in reality the technology now available is only just beginning to provide for chemists the ability to disseminate the models they previously only held in the "minds eye".

FIGURE 1: Publication at Source: e-dissemination rather than simply e-publication of papers on a website

The databases and journal papers indicated in figure 1 link to reference data provided by the authors and probably held at the journal site or a subject specific authority. Further links back to the original data take you to the author's laboratory records. The extent and type of access available to such data will be dependent on the authors as will be the responsibility of archiving these data.

One of the most frustrating experiences for a chemist is reading a paper and finding that data necessary for your own analysis is in a figure that requires OCR to obtain the numbers. Even if the paper is available as a PDF the problems are not much simpler. In many cases the numeric data is already provided separately by a link to a database or other similar service. In many cases if the information required is not of the standard type anticipated by the author then the only way to request the information is to contact the author and hope they can still provide this in a computer readable form. We seek to formalise this process by extending the nature of publication to include these links back to information held in the originating laboratories. In principle this should lead right back to the original records (spectra, laboratory notebooks). It may be argued that for publicly funded research we have a responsibility to make all this information available. The immediate impact that many people may imagine on this is that it will make the detection of fraud much easier, but this is in fact a relatively minor issue. The main advantage will be the much greater use and re-use of the original data and the consequent checking of the data and different approaches to the analysis. The scientific process requires that as much as possible of the investigations are repeated and this applies just as much to the analysis as the initial data capture in the experiments

3. PAPER MANAGEMENT OR KNOWLEDGE MANAGEMENT

These issues are not new, but require careful reflection with the advent of high bandwidth distributed data grids. Careful curation of results is supposed to be one of the aspects of scientific research. Many laboratories have stocks of paper output from equipment containing the raw data, and in a well run laboratory these rolls will be annotated with all the necessary metadata. However when these are needed the printed information has often faded. Will electronic knowledge management help? Not without a fundamental shift in the way we record and save the information (*i.e.* with more regard to re-use by every one). The paper record may be more durable than the electronic. Properly kept paper will survive for hundreds of years but the 5 1/4 inch disks are hard to read because few readers are available — even more of a problem for other technology such as punched cards.

The ability to move data round the Grid could eliminate these problems. The migration of data from one storage device to another could be handled by a central data store. For example RAL will store much of the data needed by the UK community from the LHC at CERN. This passes the problem from the individual to the community. But in many ways this is the opposite of the responsibilities of publication at source where because of the quantities, spread, and widely differing nature of all the data to be recorded (unlike large quantities of the same data from the LHC) it is the individual researcher (and laboratory) who needs to be responsible. We may need to distinguish between management of the information (responsibility of the researcher) and maintenance of the data (responsibility of the archivist).

Issues of access control, security and authentication need to be addressed to control sharing of such experimental data over geographically dispersed collaborations. These issues are common to many uses of the web/grid with the electronic availability simply increasing the degree of concern. In such an environment an audit trail is important, as it is even easier to misunderstand raw data than to misunderstand or misuse a traditionally published result.

If the Data Grid concept works well then it will not be necessary to copy the data resource but simply provide a link to the data (held back at the source laboratory); the relevant bi-directional audit trail will be clearer. This however places additional responsibilities on the authors and maintainers of the source data. It would allow the authors of the data to correct any problems, though this itself highlights the issue of version control to ensure that the ripple induced by such corrections in subsequent analysis can be observed and kept under control.

4. MULTIMEDIA COLLABORATION

A key issue for Chemists making use of the Grid will be the support it can provide for distributed collaboration. This includes video, multimedia as well as the traditional need we have for visualisation. We have already demonstrated the need for significant, real time, video interaction in the area of running a high throughput single crystal x-ray crystallography service. A demonstration Grid aware system allowing users to interact with the UK EPSRC x-ray crystallography service bases at Southampton has highlighted a number of QoS and security issues that a Grid system must encompass if it is to provide an adequate infrastructure for this type of collaborative interactions. For example the demands made on a firewall transmitting the video stream are very significant.

REFERENCES

- [1] Buckingham Shum, S., De Roure, D., Eisenstadt, M., Shadbolt, N. and Tate, A. (2002) CoAKTinG: Collaborative Advanced Knowledge Technologies in the Grid. Proceedings of the Second Workshop on Advanced Collaborative Environments, Eleventh IEEE Int. Symposium on High Performance Distributed Computing (HPDC-11), July 24-26, 2002, Edinburgh, Scotland.
- [2] Bush, V. (1945), As We May Think, Atlantic Monthly, July 1945
- [3] Ginsparg, P. (1996) Winners and Losers in the Global research Village. Invited contribution, UNESCO Conference HQ, Paris, 19-23 Feb 1996. <http://xxx.lanl.gov/blurb/pg96unesco.html>
- [4] Harnad, S. (1995) Electronic Scholarly Publication: Quo Vadis? Serials Review 21(1) 70-72 (Reprinted in Managing Information 2(3) 1995)
- [5] Hitchcock, S., Carr, L. and Hall, W. (1996) A Survey of STM Online Journals 1990-95: the Calm Before the Storm. In Directory of Electronic Journals, Newsletters and Academic Discussion Lists, sixth edition, edited by D. Mogge, (Washington, D.C.: Association of Research Libraries), pp. 7-32
<http://journals.ecs.soton.ac.uk/survey/survey.htm>
- [6] Lehn, J.M. (2001) "Some Reflections on Chemistry", Chapter 1 (page 1-7), in Chemistry for the 21st Century, E. Keinan and I. Schechter Eds. Wiley-VCH, Weinheim, ISBN 2-527-30235-2.
- [7] Lucier, R.E. (1990) Knowledge Management: Refining Roles in Scientific Communication, EDUCOM Review, Fall 1990
- [8] Schaffner, A. (1994) The Future of Scientific Journals: Lessons from the Past Information Technology and Libraries, Volume 13, Number 4, December 1994 (p.239-247)
- [9] Wells, H.G. (1937) World Brain: The Idea of a Permanent World Encyclopaedia, Encyclopédie Française, August 1937

© J. G. Frey, D. de Roure, L. Carr 2002