Scientific Communication and the Dematerialization of Scholarship

by Douglas Brown

Abstract

Many scientific research fields are becoming massively computationally intensive, handling and mining enormous datasets, a trend that is opening up possibilities for new methods of discovery, transdisciplinary and problem-centred investigation, and very large scale collaboration. Simultaneously, research practices at the frontier are changing rapidly as scientists and engineers are moving towards a research process of continuous refinement - writing, annotating and revising in near real time using the Internet - a tendency that may be further encouraged by the emergence of new, informal writing platforms and collaborative tools. These and related developments of the last decade may be contributing to the transformation of a system of scholarly research communication, based on the printed scholarly journal and the research article, that has been in place essentially unchanged for over three centuries. Following a backward glance at the beginnings of modern scientific communication, this article draws attention to this sudden, apparently dramatic shift, reviewing moves towards the development of 'cyberinfrastructure', a vision of a 'natively digital' scholarly communication system, the proliferation of open access institutional repositories, and the possibility of entirely new forms of scholarly communication as science itself shifts into a new phase in the 21st century.

Introduction: Beginnings

We have circuits or visits of divers principal cities of the kingdom; where, as it cometh to pass, we do publish such new profitable inventions as we think good. And we do also declare natural divinations of diseases, plagues, swarms of hurtful creatures, scarcity, tempests, earthquakes, great inundations, comets, temperature of the year, and divers other things; and we give counsel thereupon, what the people shall do for the prevention and remedy of them.

Bacon, *The New Atlantis* (1627)

The Prophet of the Scientific Revolution

Sir Francis Bacon
http://en.wikipedia.org/wiki/Francis_Bacon

Sir Francis Bacon left behind a fragmentary but highly significant Utopian fantasy called *The New Atlantis*, (1) which seems to have been composed in his final years, following his calamitous fall from political grace in 1621. Through a tale of English mariners lost in the South Seas and encountering a remote but highly evolved island civilization, Bacon presents a vision of an ideal commonwealth, based around a cooperative scientific research establishment which takes as its purview the entire physical creation, adopting an experimental approach to fields ranging from astronomy to mechanics, agriculture and medicine. Famously wedded to ‘utility’ and opposed to ‘armchair’ metaphysical speculation of any kind, Bacon shows himself here as a practical visionary: at a time when proto-scientific investigation was widely seen as an occupation for isolated eccentrics, he envisages scientific research as the essential collaborative craft or industry, to be carried out on a grand scale; and, crucially, his knowledge-based society depends upon his college of the sciences circulating and actively publicizing its findings for the material benefit of all.

From Script to Print

As is well known, Bacon’s ideas were the major inspiration for the formation of the Royal Society of London in 1660, an event which is now generally seen to mark a critical moment in the intellectual history of the West, as part of a fundamental epistemological shift away from reliance upon received authority and tradition. The Society was dedicated to the experimental method and open circulation of information and ideas, imagining an ‘Empire of Learning’ across national and linguistic barriers.


Sir Henry Oldenburg
http://en.wikipedia.org/wiki/Henry_Oldenburg

In a Baconian light, it is interesting to consider that a key figure in establishing the Society’s early influence was the first Secretary, Henry Oldenberg (c.1617 – 1677), a German-born diplomat and natural philosopher whose social network included some of the finest philosophical and scientific minds in Europe. After several years of meticulously copying, annotating, translating, forwarding and re-forwarding the private correspondence of Society members, Oldenberg gained license to turn to the printing press in 1665. *Philosophical Transactions,* a monthly serial of which he now became founding editor and publisher, began life as a miscellany of speculative observations and early experimental findings, in a form designed to communicate not only with the membership but with all who ‘delight in the advancement of Learning.’ (2) The philosopher of science Thomas Kuhn has argued that the published exchange over Isaac Newton’s experiments in optics, which Oldenberg mediated in the early 1670s, gave birth to ‘professional consensus science’, (3) the principles of which have been built upon ever since.

Oldenberg’s Shadow

In carefully devising formative methods for establishing what would now be called intellectual property rights of authors, as well as precedence in conducting research and presenting findings, standards of impartiality and a measure of quality control through an embryonic form of ‘peer review’, Oldenberg’s collaborative spirit and entrepreneurial zeal can be credited with laying much of the groundwork for modern scholarly research communication. The letter eventually evolved into the more formal research article, and as knowledge grew increasingly specialized the following centuries saw the scholarly journal play an active role in helping to shape emergent disciplines.

Just two months before the first number of *Philosophical Transactions* was pulled from the press a cognate scholarly publication called *Journal des Sçavans* appeared in France; it is frequently remarked that from Newton's day, even through the rise of 'Big Science' and the large-scale commercialization of scholarly publishing following post-World War II, the essential functions and processes of the print and paper model remained unchanged for well over three centuries.

Frontier Research

A New Age

Let us fast forward to 2003: a landmark report of the National Science Foundation on initiating the development of 'cyberinfrastrucure' suggests that 'a new age has dawned in scientific and engineering research'. (4) The report goes on:

The amounts of calculation and the quantities of information that can be stored, transmitted, and used are exploding at a stunning, almost disruptive rate. Vast improvements in raw computing power, storage capacity, algorithms, and networking capabilities have led to fundamental scientific discoveries inspired by a new generation of computational models . . . Powerful 'data mining' techniques operating across huge sets of multi-dimensional data open new approaches to discovery. Global networks can link all these together and support more interactivity and broader collaboration. (5)

Interactive map that shows nine of the world's largest computing Grids
http://www.pparc.ac.uk/Nw/GIN.asp

This new wave of developments in very high-performance computing is opening up entirely new possibilities for transdisciplinary or 'problem-centred' research, as emphasis shifts on a massive scale to the capture, processing and analysis of primary data, while networking greatly facilitates remote collaboration and pooled expertise; further, computation is becoming no longer simply an aid to scientific research, but integrated into its fabric, effectively changing the whole notion of a scientific instrument. The report notes that the rise of computationally intense, networked science is being accompanied by a changing pattern in the way scholars communicate: 'The traditional, linear, batch processing approach is changing to a process of continuous refinement as scholars write, review, annotate, and revise in near real time using the Internet.' (6)

Since the report was published, the explosion in the use of social writing platforms, such as weblogs and wikis, and the practices of creating 'mashups' and 'workflows' may be ramifying this trend towards 'informalisation' and 'massively distributed collaboration', perhaps shifting the World Wide Web somewhat closer to Berners-Lee's initial conception of it as a collaborative, editable, peer-to-peer space for scientists to develop their thinking. In line with the constellation of developments now known as ‘Web 2.0’, researchers across disciplines are increasingly expecting access to raw, primary data as well as finished documents, and the ‘journal article’ seems to be becoming part of a continuum of research communication, with the conventional distinction between formal and informal communication falling away.

e-Science, The Grid and The God Particle

In parallel with the development of the cyberinfrastructure initiative in the US, Research Councils UK (comprising the eight publicly-funded UK research councils), have focused on the idea of ‘e-Science’, which they define as “the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet…such collaborative scientific enterprises…will require access to very large data collections, very large scale computing resources and high performance visualization.” (7)

Large Hadron Collider at the particle physics laboratory CERN http://en.wikipedia.org/wiki/Large_Hadron_Collider

Increasing collaboration has been a trend in many branches of science throughout the latter half of the twentieth century, as a reflection of intensified specialization and the need for a mix of expertise, reinforced by improvements in technology. (8) Underpinning e-Science research is the development of grid computation, a group of technologies that will facilitate very large scale distributed data computation and storage, and enable flexible and coordinated resource sharing and problem solving. The best known application so far is probably the Human Genome Project; test-bed experiments include GridPP, led by the University of Glasgow, which will analyze enormous amounts of data produced by the Large Hadron Collider at the particle physics laboratory CERN, during the upcoming investigation into the fundamental nature of matter, scheduled for 2007. This massive experiment will initially involve 200 scientists in 150 academic institutions worldwide. The effort to isolate the elusive particle known as Higgs boson, sometimes called ‘The God Particle’, is expected to generate 10 terabytes of data for each 8-hour run of the Collider.

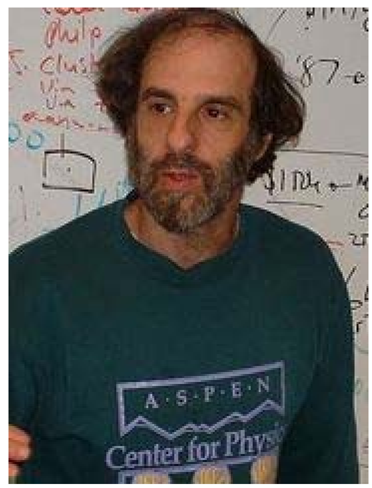
Towards a Natively Digital Model?

A New Vision

In concert with long-standing economic pressures, which have led academic librarians to speak of a ‘Serials crisis’ for many years, these mega-trends are now feeding into calls for structural reform of the formal scholarly communication system.

Herbert Van de Sompel of Los Alamos National Laboratory has presented a case for a ‘natively digital’ framework which would ‘disaggregate’ the traditional functions of scholarly publishing (standardly characterized as registration, certification, awareness, and archiving) and create a record that would aim at more closely capturing 'the dynamics of scholarship.' (9) Simeon Warner of Cornell has suggested that a proper record can no longer regard datasets, software, simulations or 'rich media' as supplementary materials or add-ons. With Van de Sompel, he proposes that the traditional journal article should be extended into a new multi-faceted communicative unit he calls 'complex documents', where the (now often extremely large) dataset is referenced and seamlessly available to the reader, with the provenance of each component of the extended document being recorded and the integrity of each 'ensured and verifiable'. (10) The infrastructure here would be a range of heterogeneous, distributed digital repositories woven into a fabric by standards of interoperability, allowing for the flexible reuse and aggregation of the scholarly 'digital objects' and their various parts (text, images, data, software and so on), and the possibility of 'overlay journals'.

After Ginsparg



Paul Ginsparg at Cornell University
http://en.wikipedia.org/wiki/Paul_Ginsparg

Van de Sompel and Warner take account of changing research practices at the edge and the emergence of science mediated by networked resources and archives in a conscious effort to broaden discussion about Open Access to research beyond questions about the economics of electronic publication. They are inspired by the success of Paul Ginsparg's groundbreaking, physics-centred e-print server, founded at Los Alamos in 1991, now located at Cornell and known as arXiv. This site, which is mirrored in 17 countries, has now become the vital, global resource for access to results in medium and high energy physics (containing 400,814 e-print deposits in physics, mathematics, computer science, and quantitative biology as of January 2007). Researchers can deposit papers at any stage (pre-refereed working papers or formally peer reviewed 'post-prints'); most go on to appear in peer reviewed journals a year or two on. As a dissemination system, Ginsparg estimates that it operates at 'a factor of 100 – 1000 lower cost than a conventional peer-reviewed system'. (11)

Deposited papers are subjected to a minimal degree of screening before being made freely available. Some of arXiv's success may be predicated on the fact that high energy physics in particular enjoys a highly collaborative, communitarian pre-publication culture, in which findings are informally circulated or presented at conferences, and theses and working papers are subjected to rigorous internal review and external critique before submission for formal peer review. It is not unusual for collaborations here to involve over a hundred, even hundreds of, souls. One of the pioneers of Open Access, Stevan Harnad, has always taken the view that scientific inquiry is an ongoing process of continuous refinement, with the 'pre-publication continuum' by far the most interesting phase. (12) He has written: 'Learned inquiry is a continuum. Reports of its findings, both informal and formal, unrefereed and refereed, are milestones, not gravestones.' (13)

Institutional Repositories, and the Future

Institutional Repositories, Focused or Expansive

Professor Stevan Harnad
http://en.wikipedia.org/wiki/Stevan_Harnad

While acknowledging the crucial role of arXiv in developing a centralized, subject-based repository, Harnad is now a leading, tireless advocate of supplementary research self-archiving in institutionally-based e-print repositories. These have proliferated with the support of the Open Archives Initiative (OAI), which promotes standards of interoperability, and particularly the OAI Metadata Harvesting Protocol (2001), which has provided metadata tagging standards with the aim of allowing for a global network of repositories. Compliant software is now available open source, leaders being DSpace (developed at MIT) and EPrints (University of Southampton). Institutional repositories received an indirect boost from the appearance of the academic search engine Google Scholar in beta form in 2004.

While the majority of the 24,000 peer-reviewed scholarly journals in all fields now appear in electronic form, including a growing number of openly accessible 'born digital' journals, Harnad suggests that only around 15 per cent of researchers are currently making their work freely accessible on the Web through self-archiving (14). He is not concerned with reforming scholarly communication processes as such; his focus is squarely on research impact and usage, and the degree to which this has been lost since publication began by the high costs of print production and distribution. Harnad puts the case that, today, the deposit of a scholar's final, post-peer review version of an article in an institutional repository, in compliance with copyright law and with publisher policies, is the quickest and most effective means of ensuring that as many interested researchers as possible read and are led to cite the article; research suggests that open access may increase citation impact by 50-250%. (15) He submits that maximizing impact potentially benefits both the individual researcher and his or her institution, in making visible the quality and relevance of its intellectual output, and that ultimately society can benefit from increased return on investment and accelerated research progress.

To achieve these ends Harnad campaigns for policies of institutional and funder-level self-archiving mandates; in the meantime, an apparent lack of willingness or incentive for faculty to self-archive, and alternative, more expansive conceptions of institutional repositories that resemble full digital libraries, have recently led some repositories to house a vast range of materials, including datasets, learning objects, presentations, technical reports, theses, video and sound files, teaching materials, and administrative and numerical data. All of these were previously unavailable campus-wide or to the wider research community. Enthusiasts for a possible new, active role for universities in management, preservation and stewardship of digital materials, and experiments with institutional repositories re-engineered as new university-based publishing platforms, seem to be setting off on a divergent path from Harnad's more focused position, which prioritises global access to peer reviewed papers. (16) A decade of technological flux has alerted librarians to issues of long-term accessibility and digital preservation, but for 'archivangelists' like Harnad who regard self-archiving of research as essentially supplementary to the version of record held by the publisher, concerns about preservation should not deter researchers from making their work immediately openly accessible.

Conclusion: Towards Another Scientific Revolution?

Uncertainty about the future role of institutional repositories is part of a much broader landscape of uncertainty. As scholarly publishers have begun over the last decade to test alternative business models and experiment with many varieties of Open Access, and alternatives to traditional peer review begin to be trialed, Blaise Cronin, a leading authority on scientific communication, has questioned whether it makes sense to speak of the scholarly communication system at all. Cronin predicts that cultural differences among disciplines will persist and combine with technological innovation and plenitude to discourage the emergence of any unified model of scholarly communication, and eventually encourage 'a much more heterogeneous and dynamic publishing ecosystem' than in the past, from which 'new hybrids' will evolve. (17) By contrast, other commentators see the developments reviewed here as symptoms of the wholesale disruption of a traditional, discipline-centred system and emergence of 'a new mode of knowledge production', the research infrastructure for which will require to be addressed and developed through 'an holistic approach' to the entire scholarly communication cycle across disciplines. (18)

Meanwhile, the sheer volume of data produced by scientific research is now said to be doubling with each passing year, and a recent report from some of the world's leading scientists suggested that the further evolution of computationally intense science will effectively result over the next decade or so in another scientific revolution. (19) Something that should remain unaltered is what John Willinsky has called a basic principle of science: 'that knowledge lives through the scope of its circulation, and thus through its very susceptibility to contention and alteration'. (20)

References

1. Bacon, F., The New Atlantis, English version, 1629 http://www.gutenberg.org/etext/2434
2. Oldenberg, H., *The Introduction,* Philosophical Transactions, vol.1 no.1, Monday, March 6, 1664/5
3. Quoted in Willinsky, J., The Access Principle: the Case for Open Access to Research and Scholarship, MIT Press, 2005, p.202
4. Atkins, D., et al, Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure (2003) http://www.nsf.gov/od/oci/reports/ExecSum.pdf
5. Atkins et al, ibid

6. Atkins et al, ibid
7. http://www.rcuk.ac.uk/escience
8. Odlyzko, A., *The Future of Scientific Communication* (2002) http://www.dtc.umn.edu/~odlyzko/doc/future.scientific.comm.pdf
9. Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., Warner, S., *Rethinking Scholarly Communication: Building the System that Scholars Deserve*, D-Lib Magazine vol.10 no.9, September 2004 http://dlib.org/dllib/september04/vandesompel/09vandesompel.html
10. Warner, S., *The Transformation of Scholarly Communication*, Learned Publishing vol. 18, no. 3, July 2005, p.181
11. Ginsparg, P., *As We May Read*, Journal of Neuroscience vol.26 no.38, September 2006, p.9606 http://www.jneurosci.org/cgi/reprint/26/38/9606
12. Harnad, S., *Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry*, Psychological Science 1, 1990, pp.342-343 http://www.ecs-soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html
13. Harnad, S., *The Invisible Hand of Peer Review* (1998) http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/nature2.html
14. Harnad, S., *Publish or Perish – Self-Archive to Flourish: the Green Route to Open Access*, ERCIM News 64 (2006) http://eprints.ecs.soton.ac.uk/11715/
15. Hajjem,C., Gingras, Y., Brody, T., Carr, L., and Harnad, S., *Open Access to Research Increases Citation Impact*, Technical Report, Institut des sciences cognitives, Université du Québec à Montréal (2005) http://eprints.ecs.soton.ac.uk/11687/
16. Poynder, R., *Open and Shut? Institutional Repositories and a Little Experiment* (2006) http://poynder.blogspot.com/2006/03/institutional-repositories-and-little.html
17. Cronin, B., *Scholarly Communication and Epistemic Cultures*, ARL, Washington DC, October 2003 http://www.arl.org/scomm/disciplines/Cronin.pdf
18. Houghton, J.W., Steele, C., Henty, M., *Research Practices and Scholarly Communication in the Digital Environment*, Learned Publishing vol.17, no. 3, July 2004, pp.247-248
19. http://research.microsoft.com/towards2020science
20. Willinsky, op.cit, p.205