ALPSP Future Watch Committee

White Paper 1

How is scholarly communication changing as a result of the Web?

Rapporteur:
Mary Waltham

Contributors:
Tony Hey, *Corporate Vice President for Technical Computing, Microsoft Corporation*
Clifford Lynch, *Executive Director, Coalition for Networked Information*

ALPSP
www.alpsp.org

First published 2006 by
Association of Learned and
Professional Society Publishers (ALPSP)
Registered address:
Blenheim House, 120 Chruch Street,
Brighton, BN1 1AG, UK
www.alpsp.org for contact details.

ISBN 978-0-907341-34-5

How is scholarly communication changing as a result of the Web? (or 'Daddy - what's a mash-up?')

Rapporteur:
Mary Waltham

Contributors:
Tony Hey,
Corporate Vice President for Technical Computing, Microsoft Corporation

Clifford Lynch,
Coalition for Networked Information

Context for the White Paper

The shift from print to online as the predominant publishing format for scholarly information is transforming both the economics and the operations of publishers at many levels. In turn the expectations from users of scholarly information have increased steeply as information that is published online can be linked, manipulated, imported and therefore used in a broad variety of ways which are distinctly different from print.

The ALPSP Future Watch committee invited two speakers uniquely positioned to review these transformations; what follows is a summary of their views of the scope, scale and direction of the transition now taking place. Brief biographies of the presenters - Tony Hey and Cliff Lynch - are included at the end of this White Paper.

Presentations

Tony Hey

Hey proposed that the strategic force behind e-science[1] is having a fundamental impact on scholarly communication. The vision for fully interlinked online information is becoming a reality in certain well defined disciplines and from the results of particular projects. Alongside this, and because of it, there are wider and more geographically diverse collaborative networks of scientists working on the same problem (see Fig. 1, page 2).

Powerful use is being made of computing resources to integrate, federate and analyse information from many disparate sources; the result is a new-found ability to access, move, manipulate, analyze and visualize the results of such work. Within this large pool of a variety of data and published information, the fraction that is described as 'scholarly' and has been peer-reviewed may be relatively small but it is of considerable significance.

[1] See http://en.wikipedia.org/wiki/E-Science

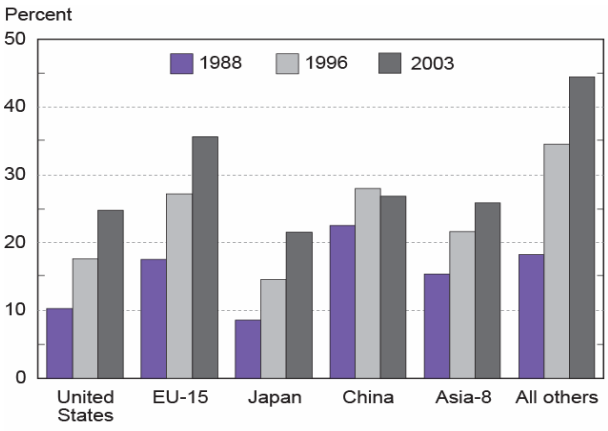

Fig. 1: *Trends in research collaboration globally: share of scientific and technical articles with international co-authorship, by country/region: 1988, 1996, and 2003*

Source: Science and Engineering Indicators 2006, NSF

Hey mentioned the following examples:

Neptune[2]

This application monitors movement of the tectonic plates in the N.W. Pacific Coast, and uses distributed storage and visualization typical of collaborative networks. 'Understanding and prediction depend upon being present throughout the ocean, all the time, with the capacity to detect, measure, sample and respond to these rapidly evolving, energetic events. Neptune will enable this type of remote, real-time, long-term interactive presence in the oceans.'

Sloan Digital Sky Survey (SkyServer)[3]

Data from many different instruments and times are brought together to create new knowledge. 'The ability to record and digest immense quantities of data in a timely way is changing the face of science. The Sloan Digital Sky Survey will bring this modern practice of comprehensive mapping to cosmography, the science of mapping and understanding the universe.' The astronomy community has been especially adventurous and successful in using digital formats for research.

Key strategic requirements for such collaborative networks to be effective include:

- The components of the network must be interoperable.
- Data must be real and well documented.
- Data must be complete.

Clearly these requirements are not all met within many or even most research environments. However, the situation is changing; both Neptune and SkyServer provide useful living examples of such work.

There is a growing trend, both among national Government research funding agencies and in recommendations from influential groups such as OECD[4], towards policies that require scientists to deposit their data (not-yet-the resulting research articles) in a publicly available data repository. Increasingly, applications for research grant support must include a plan for managing the research data once the research has been completed.

Hey also commented on the need for more visualization of the results of online search. Some tools do exist to enable this, but there

[2] http://www.neptune.washington.edu/about/index.jsp?keywords=NEWRDS&title=New%20Era%20of%20Discovery

[3] http://cas.sdss.org/dr5/en/sdss/

[4] This article summarizes the report of an OECD sub-committee well: http://journals.eecs.qub.ac.uk/codata/journal/contents/3_04/3_04pdfs/DS377.pdf

will be more; it will be important for producers of information to describe their content clearly using metadata.

Certain A&I databases (e.g. Proquest's ABI Inform) distinguish 'peer–reviewed' content from non-peer-reviewed content. Peer review implies integrity, reputation and influence. Hey is intrigued by the role that social networks, based on the reputation of individuals, will increase in importance (e.g. Faculty of 1000 from BioMedCentral). He believes that a degree of anarchy will develop, simply because the online environment is implicitly anarchic and distributed.

In sum, e-science (or data-driven science) has the power to enable a unification of theory, experimentation, and simulation using data exploration and data mining techniques, which themselves are merging. Data is captured and then processed by software; the scientist then analyses the results within databases and files that are created using these software tools.

Current problems for the e-scientist include:

- Acquisition of data and information – can I find all I need?
- Import – is it in a format my software can read/explore?
- Annotation – is it clear what the data and information refers to?
- Provenance – where does the data come from and is it to be 'trusted'?
- Data storage – can I store what I need?
- Data quality – is this the best and most reliable data for my work?
- Curation and preservation of data – certainly a non-trivial question given the sheer size and growth of large data sets.

Since many of these issues are also those which concern scholarly publishers, interesting partnerships and collaborations are emerging. Hey mentioned that Microsoft is working with Johns Hopkins University Press to link data to publications – SkyServer is one example. (See also the ALPSP ILJS 2006 report by Kurt Paulus that includes references to Ray Everngam's concept of 'radial journals'[5].)

Hey commented on the following trends within publishing, based on his view from the 'outside':

- A global movement towards Open Access, through a variety of routes.
- Development of compliance with the Open Archive Initiative Protocol for Metadata Harvesting (not to be confused with the Open Access business model), which facilitates interlinking of repositories.

Within this context the JISC funded TARDIS[6] initiative takes account of OA and non-OA articles within the context of a repository. Portable PubMedCentral is now deployed in the US, China, UK, Italy and South Africa[7].

Microsoft is working on a number of tools to support information collection and dissemination, including:

- 'Data Workbench – a digital lab book to record methods and results of experiments that do not work, as well as those that do work and develop naturally into the results written up; negative results are also of value in developing knowledge.
- Conference Management Tool – since conference proceedings are often used by computer scientists as a fast track to publication.
- E-journal management tools to enable journal publishing.

5 http://www.alpsp.org/events/2006/TNP/iljs-kp.pdf

6 See http://www.jisc.ac.uk/index.cfm?name=project_tardis

7 'In its first phase, Portable PMC will mirror the PMC, but, over time, it will become more independent. A Japanese group, Microsoft, and the Wellcome Trust will be installing it. Other search engines can be hooked to the Portable PMC, giving the partners flexibility. There are hooks for different kinds of searching. Portable PMC does not come with its own search engine' Source: David Lipman: http://cendi.dtic.mil/minutes/pa_0205.html

Hey could see further development of social networks within the context of Web 2.0 such as wikis, blogs and RSS. Integration of online sources ('mash-ups') in new ways will add value by combining such services with more traditional ones.

The Microsoft strategy for e-science can be summarized as follows:

- Define open standards and/or interoperable high level services – work flows and tools.
- Assist the academic community in developing open scholarly communications.
- Work with publishers to explore new business models for scholarly publishing.

Clifford Lynch

Preservation

Lynch believes we are within a year or two of the inflection point when print will not be required for preservation of scholarly works to be viewed as 'safe', despite an apparent distrust of digital preservation by some sectors of the library community. This will be brought about through development of initiatives such as Portico, LOCKSS and the National Libraries' preservation policies (e.g. NLM in the USA and KB in the Netherlands), and a greater degree of confidence among stakeholders that there is a well articulated understanding of:

- Who does what and where responsibilities lie.
- The economics of preservation activities.

It is important to remember that 'even' print is not a perfect model for preservation (for example, JSTOR experienced considerable difficulty in locating complete runs of even quite well-known journals). Lynch believes that print will evaporate rapidly, except perhaps as a member benefit for society publishers. (Since many societies provide member copies at a net loss, this will mean that publishers will have to price their member copies more realistically; they may lose members [or member print subscriptions] as a result.)

Adding value through the editorial process

Peer-reviewers are becoming overloaded; this is leading to considerable reviewer fatigue, with reviewers withdrawing their support because they feel they are not given due credit or a sense of purpose. Publishers and their Editors have pushed peer review too far, and this is the result.

Submission patterns show a clear shift, as shown in Fig. 2, page 5).

More research is being submitted to English language journals by authors whose first language is not English. Publishers will need to respond to this, and doing so may have financial implications. Lynch believes that overall there is growing cynicism within the communities served about the service(s) provided by journals.

Social reviewing systems

These are often complex and contradictory. Despite invitations to 'comment' on articles, most authors rarely do; they are too busy, and there is too much literature. The need is more for services that 'point me to the stuff I should read'. In any case, not every article will ever be commented on! The current *Nature* open peer review trial[8] is an example of greater openness in peer review and the willingness of some scientists in some disciplines to put articles and comments in the public domain prior to publication.

Future issues of data and publishing

- Who does the data belong to? Authors will come under pressure from funding and public policy groups for greater data sharing (e.g. access to sequence data for avian 'flu.)

8 www.nature.com/nature/peerreview

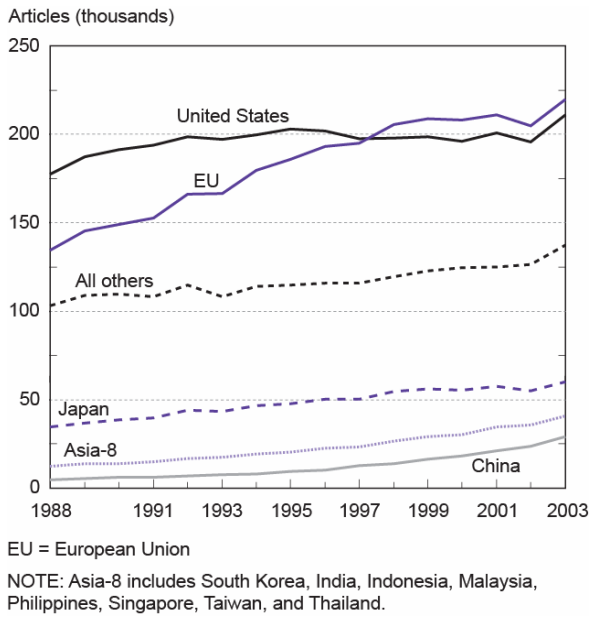

Fig. 2:
Trends in research productivity globally: Scientific and technical articles, by country/region: 1988-2003

Source: Science and Engineering Indicators 2006, NSF

- The relationship between scholarly literature and data. Data has a life of its own which will vary by discipline. Publishers only dabble in the role of curators of supplementary data; so although articles are kept alive, data may not be. Yet all the issues surrounding data are common to journal curation and preservation. For example: should it be discipline-specific? How will it be funded? The NSF Cyber-infrastructure Report[9] addresses this issue.
- Software will also need to be 'preserved'; there is little understanding yet of how to preserve complex software systems with very particular applications, and the associated data sets; data could be stranded in future.
- When to get rid of data? Unless some clear policies are established about this, considerable resources will be gobbled up curating old data! We need a systematic way of keeping or destroying data and this will prove hard to do. Textual information can be kept forever, because storage costs are dropping so rapidly that this component of data essentially costs nothing to store. But what about a listserv: when is it appropriate (or even desirable!) to delete the discussions that take place?
- Laboratory experimental software now includes many more features to enable 'data dumping'; it is clear that the data environment within a laboratory is getting much more complicated, and yet there is little in the way of cross-platform support.

New types of users of scholarly information

Users increasingly want to compute on large corpora of scholarly information. Workgroups are amassing their own digital libraries. Industry, in particular, invests heavily in this area; 'Big Pharma' is one example.

Web crawlers, not eyes, are used to extract information from the literature. Typical publisher license agreements do not address the requirement for large-scale downloading and storage for subsequent use; indeed, some licenses explicitly forbid it. Copyright law is

[9] Revolutionizing Science and Engineering through Cyber-infrastructure See: http://www.nsf.gov/cise/sci/reports/atkins.pdf

also murky around these areas: are these digital collections in fact derivative works? Plainly many of these difficulties disappear if both data and content are available under Open Access models.

Literature ubiquity

The scholarly literature needs to be available anywhere, not just tied to a physical location. There is much interest in and growing use of collaborative work environments, where virtual organizations are pulled together from all around the world (see Fig. 1, page 2); a huge advantage is that these can be set up and disbanded relatively easily. However, they face major problems because of the current focus on institutional subscriptions as the source of scholarly literature; clearly, the way scholarship works is increasingly at odds with the notion of an institutional site license - scholarly literature is seen to be 'balkanised' on publishers' sites. Open Access is very convenient because there are no concerns over the ability of a group to access all the same literature. Pay per View could be one solution; note, in particular, the results of the PEAK project[10] which helped to elucidate the delicate balance between time and cost to access individually selected research articles. As they become accustomed to the ease, convenience and low price of downloading music on sites such as iTunes, scientists are expecting their online experience to be just as good.

[10] See summary at http://www.library.yale.edu/~llicense/ListArchives/9909/msg00037.html

Summary of Committee discussion

- The infrastructure to deliver a complete and seamless online research environment is incomplete; it will be expensive to ramp it up to this level. This is therefore likely to take place over many years; it will continue to be discipline-driven and discipline-dependent.
- Data curation is costly. There is a lack of awareness of this within academia.
- Open Access, the Open Archiving Initiative and Open Source software/applications are frequently confused.
- There appears to be an increasing disconnect between academics and librarians. Author-side payment for publication of scholarly research, unlike library budgets, does scale with the increasing volume of research.
- The article-based economy is 'coming soon'.
- Institutions will drop print as the format they acquire within the next 5 years. Publishers need to be prepared for this with a thorough and detailed understanding of their print-specific, online-specific and shared print- and online-related costs and revenues.
- Publishers do need to face the question of how to change the online version to meet the needs of researchers. Adding value should not involve adding costs, as the journal is not an end in itself. If it is awkward and/or costly to find published research information, users can and will go elsewhere for their information needs.
- One critical point, during the transition that is now playing out, is the optimization of pricing in relation to changing business models.
- While data produced from research is made

available ahead of the published article, once online both are subject to analysis.

- Neither data nor published articles are read entirely or even mostly by individual humans, but rather by software.

- Specialised mark-up languages, that make it easier for the journal information to be 'read' by the software that is reading it now, could be a part of publishers' added value; however, this may be of limited benefit as automated text mining technology is improving rapidly.

- Social book-marking and tagging will play an increasingly important role; it is clear that tagging which is carried out by a large number of users is likely to be more effective (and less costly) than a publisher investing in the same process. For an example see NeuroCommons[11].

- We will see more development of the Amazon-type applications such as 'Show me more like this...' and 'People who bought (read: viewed/downloaded) this also bought this...'

- As more personalized online help tools develop, we are likely to see more tools to help users cope with information overload. One example could be a 'personal scout' that finds and retrieves information for the user and can be trained using neural network applications; see the Perseus project at Tufts University[12].

- Large applications for such tools within industry (such as business and competitive intelligence) can be gathered in real time by continual analysis of Blogs and Press releases.

- Licensing and technical standards to enable crawling of content are likely to become more widespread.

Summary – what it all means for publishers

Within scholarly publishing we will see wider collaborations between publishers and broader, non-traditional groups of stakeholders such as software creators and providers and database developers and curators. Publishers need to reach out to these non-traditional partners, to try and understand how scholarly content can best be integrated with, and thus used by, their target communities in ways that maximize research productivity. Text mining and personal software 'agents' will develop and grow according to the needs of particular disciplines and research areas; these exciting applications hold considerable opportunity for scholarly publishers and are worthy of exploration.

11 http://sciencecommons.org/data/neurocommons

12 http://www.perseus.tufts.edu/

Presenters' biographies

Tony Hey

Corporate Vice President for Technical Computing, Microsoft Corporation.

As Corporate Vice President for Technical Computing, Tony Hey coordinates efforts across Microsoft Corporation to collaborate with the global scientific community. He is a top researcher in the field of parallel computing, and his experience in applying computing technologies to scientific research helps Microsoft work with researchers worldwide in various fields of science and engineering.

Before joining Microsoft, Hey worked as head of the School of Electronics and Computer Science at the University of Southampton, where he helped build the department into one of the pre-eminent computer science research institutions in England. Since 2001, Hey has served as director of the U.K.'s e-Science Initiative, managing the government's efforts to provide scientists and researchers with access to key computing technologies.

Hey is a fellow of the U.K.'s Royal Academy of Engineering and has been a member of the European Union's Information Society Technology Advisory Group. He has also served on several national committees in the United Kingdom, including committees of the U.K. Department of Trade and Industry and the Office of Science and Technology. In addition, Hey has advised countries such as China, France, Ireland and Switzerland to help them advance their scientific agenda and become more competitive in the global technology economy. Hey received the award of Commander of the Order of the British Empire honour for services to science in the 2005 U.K. New Year's Honours List.

Hey is a graduate of Oxford University, with both an undergraduate degree in physics and a doctorate in theoretical physics.

Clifford Lynch

Executive Director, Coalition for Networked Information

Clifford Lynch has been the Director of the Coalition for Networked Information (CNI) since July 1997. CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and intellectual productivity. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation. Lynch, who holds a Ph.D. in Computer Science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science and the National Information Standards Organization. Lynch serves on the National Digital Preservation Strategy Advisory Board of the Library of Congress; he was a member of the National Research Council committees that published The Digital Dilemma: Intellectual Property in the Information Infrastructure and Broadband: Bringing Home the Bits, and now serves on the NRC's committee on digital archiving and the National Archives and Records Administration.

For links to Lynch's recent presentations see: http://www.cni.org/staff/clifford_talks.html

Rapporteur

Mary Waltham

Mary Waltham founded her own consulting company (www.MaryWaltham.com) in 1999 to help international scientific, technical and medical publishers confront the rapid change that the networked economy poses to their business models and to develop new opportunities to build publications that deliver outstanding scientific and economic value. Prior to creating her company, Mary was President and Publisher for Nature and the Nature family of journals in the US, and earlier was Managing Director and Publisher of *The Lancet*. Mary has worked at a senior executive level in science and medical publishing companies across a range of media, which include textbooks, magazines, newsletters, journals, and open learning materials. A graduate in biology from QMC, University of London, Mary received her education in business administration at Henley.
Mary lives in Princeton New Jersey.

ALPSP Future Watch Committee
White Paper 1

How is scholarly communication changing as a result of the Web?

Rapporteur:
Mary Waltham

Contributors:
Tony Hey, *Corporate Vice President for Technical Computing, Microsoft Corporation*
Clifford Lynch, *Executive Director, Coalition for Networked Information*

This White Paper is the first to be developed from the activities of the ALPSP Future Watch committee. It is drawn from presentations and materials discussed at a meeting of the committee at the *National Academy of Sciences* in Washington DC on June 7th 2006.

ALPSP
www.alpsp.org

ISBN 978-0-907341-34-5