The Coming Revolution in Scholarly Communications & Cyberinfrastructure

GUEST EDITORS LEE DIRKS AND TONY HEY

1 Introduction
Lee Dirks and Tony Hey

FEATURED ARTICLES

5 The Shape of the Scientific Article in the Developing Cyberinfrastructure
Clifford Lynch

11 Next-Generation Implications of Open Access
Paul Ginsparg

19 Web 2.0 in Science
Timo Hannay

42 Incentivizing the Open Access Research Web: Publication-Archiving, Data-Archiving and Scientometrics
Tim Brody, Les Carr, Yves Gingras, Chawki Hajjem, Stevan Harnad, and Alma Swan

51 The Law as Cyberinfrastructure
Brian Fitzgerald and Kylie Pappalardo

32 Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication
Herbert Van de Sompel and Carl Lagoze

PERSPECTIVES

26 Reinventing Scholarly Communication for the Electronic Age
J. Lynn Fink and Philip E. Bourne

58 Cyberinfrastructure For Knowledge Sharing
John Wilbanks

67 Trends Favoring Open Access
Peter Suber
Introduction

By now, it is a well-observed fact that scholarly communication is in the midst of tremendous upheaval. That is as exciting to many as it is terrifying to others. What is less obvious is exactly what this dramatic change will mean for the academic world – specifically what influence it will have on the research community – and the advancement of science overall. In an effort to better grasp the trends and the potential impact in these areas, we've assembled an impressive constellation of top names in the field – as well as some new, important voices – and asked them to address the key issues for the future of scholarly communications resulting from the intersecting concepts of cyberinfrastructure, scientific research, and Open Access. All of the hallmarks of sea-change are apparent: attitudes are changing, roles are adjusting, business models are shifting – but, perhaps most significantly, individual and collective behaviors are very slow to evolve – far slower than expected. That said, each of the authors in this CTWatch Quarterly issue puts forward a variety of visions and approaches, some practical considerations, and in several cases, specific prototypes or break-through projects already underway to help point the way.

Leading off is Clifford Lynch's excellent overview ("The Shape of the Scientific Article in the Developing Cyberinfrastructure") – an outstanding entry point to the broad range of issues raised by the reality of cyberinfrastructure and the impact it will have on scientific publishing in the near-term. His paper is an effective preview to the fundamental shift of how scholarly communication will work, namely how the role of the author is changing in a Web 2.0 environment. A core element in this new world is the growing potential benefit for inclusion of data in submissions (or links to data sets). Lynch thoughtfully addresses the many implications arising in this new paradigm (e.g., papers + data) – and how policies and behaviors will need to adapt – most especially the impact this will have on the concept of peer review. He astutely raises the issue of the importance of software/middleware in this new ecosystem – namely in the areas of viewing/reading and visualization. This is a critical point for accurate dissemination to facilitate further research – and is also integral to discoverability as well as the ability to aggregate across multiple articles.

In his piece, “Next-Generation Implications of Open Access,” Paul Ginsparg provides an invaluable perspective on the current state of affairs – a "long view" – as one of the originators of the Open Access movement. Having in essence invented the Open Access central repository when he launched arXiv.org in 1991, Ginsparg’s brief retrospective and forward-looking assessment of this space is a useful look at the features and functionality that open repositories must consider to stay relevant and to add value in this changing environment. Indeed, it is a testament that arXiv.org has been able to remain true to its original tenets of remaining low-cost, selective, and complimentary/supplemental to other publishers or repositories. However, Ginsparg’s treatment hints at several new directions and areas for enhancement/improvement relating to the issues of (a) storage and access of documents/articles at scale, (b) the social networking implications for large-scale repositories as well as (c) a discourse on how to handle compound objects, data and other related supporting documentation. Also insightful are Ginsparg’s musings of the economics of Open Access, and he surfaces the important theme highlighted by several of the authors in this issue—the notion that a generational shift is required to enable the necessary behavioral change, and the recognition that our field(s) may not progress until this reality is brought about.

Lee Dirks
Director, Scholarly Communication
Microsoft Corporation

Tony Hey
Corporate Vice President, External Research
Microsoft Corporation
Timo Hannay's extremely useful survey of the Web 2.0 landscape is an especially valuable landscape map. In this environmental scan, Hannay takes a snapshot of the current state-of-the-art and provides not only definitions but also definitive examples/applications that demonstrate the reality, the potential, and the remaining hurdles faced by the social-networking phenomenon. Now that we've finally begun to realize the power and potential that had been promised us with the “web-as-platform” – we're also understanding the many benefits and the driving-force of the network effect: the more who participate, the richer the experience. (Yet, Hannay also points out the cruel truth that the scientific community has been miserably late to the game, when it should have been first – considering the Internet was initially constructed to facilitate the sharing of scientific data.) As exciting as it might be at this point in time, a core tenet of this article is to point out that – as a community – we have yet to realize the full potential of Web 2.0, as we are still so very early in the initial phase. Considering the very medium we are using changes/alters the methods we employ, Hannay stresses that is “impossible to predict” the future, but the hints he provides promise us a very exciting journey.

Lynn Fink and Phil Bourne's "Reinventing Scholarly Communication in the Electronic Age" is an especially compelling article in that it lays out examples of research and projects currently in progress to enact Web 2.0 principals. Echoing the irony from Hannay's paper, the authors note that scientific and scholarly articles have not evolved at the same pace as other developments on the Internet. In an effort to change that, the University of California, San Diego is undertaking two projects to catalyze developments in this space: (1) the "BioLit" project relating to the semantic mark-up of journal articles enhanced during the authoring stage – not after the fact – and, (2) the launch of "SciVee.com", a new online resource for augmenting scientific papers with brief video presentations. Also striking is that the article raises a theme that appears in several of the other papers related to the "generational change" that is a crucial underlying factor to the success of these types of projects. There is clearly agreement among the authors that the newer, younger generation is going to carry the scientific world forward in a way that the existing, established community cannot. So, changes that are being implemented now will begin to have greater and greater impact as this new generation of scientists and scholars shift the behavior of the community. It is an exciting prospect – but the groundwork to ensure this occurs is only being laid now with enabling research projects such as these from UCSD.

Herbert Van de Sompel and Carl Lagoze's "Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication" is an exciting look at the seminal work now underway related to the Open Archives Initiative's "Object Reuse & Exchange" (ORE) project. Building upon a concept initially referenced in Hannay's paper, they point out the need to understand "the change in the nature of the unit of scholarly communication" – meaning we can no longer effectively think about an article or paper as the primary vessel of conveying knowledge within the academic world or across scientific disciplines. The transmission of knowledge has grown, broadened, and now spans a range from atomic data to entire datasets, from a single paragraph to a series of articles about specific concept, to presentations, videos, or other modes/formats related to the dissemination of a given concept. Since our ability to communicate information has exploded, we must likewise evolve our effort to describe, find, and utilize these new "compound objects." This paper is an in-depth "nuts-and-bolts" presentation, and Van de Sompel and Lagoze explain why there is a crucial need to re-architect scholarship on the internet and how they propose to enact this to ensure that we maximize the intent and the value of what is made available for scholars and researchers. Accompanying their article is an illustrative, online demonstration of their prototype ORE implementation in the form of a screencast referenced in the Appendix of their article; this companion piece provides useful context and a quick
overview with examples to their work in progress. The screencast can be found here: http://www.ctwatch.org/quarterly/multimedia/11/ORE_prototype-demo/.

In the article by Stevan Harnad et al. entitled “Incentivizing the Open Access Research Web: Publication-Archiving, Data-Archiving and Scientometrics,” we see a bold proposal for employing a new application of research metric across multiple sources (defined as “scientometrics” – the collection, measurement and analysis of full-text, metadata, download and citation metrics) to drive author self-archiving, encourage data-archiving, and enhance research quality overall. Similarly referenced by other articles in this publication, this piece also notes that behavior and current practices of researchers and scholars need to change to match the potential that technology has provided academia. Namely, the authors propose a system to help accomplish this systemic and behavioral change – and a substantial transformation this would be. The system they present is based around three core components: (1) functionality of a network infrastructure, (2) established and agreed upon metrics to provide the necessary incentive(s), and (3) mandates from authoritative organizations that promote publishing into an Open Access system that maps directly into #1 and #2. In an effort to provide concrete examples of how such a system could evolve, the authors point to Citebase and the UK’s “Research Assessment Exercise” (RAE) as tangible case studies that could be modeled/expanded to achieve a vision of complementary components meshing together. Indeed, the tenet here being that, with this system in place, the “Open Access Impact Advantage” can be realized – where self-archiving driving citation impact becomes a virtuous cycle delivering more value than publication into closed/proprietary journals.

Brian Fitzgerald and Kylie Pappalardo’s important review of the legal options for treatment of open access implications is a focused piece honing in on how the law is working in a positive way to enable knowledge sharing in this new world of Open Access – which is made possible by the “collective endeavor through networked cyber-infrastructure.” Addressing the larger issue of “open licensing” models – Fitzgerald and Pappalardo provide multiple case-studies/examples of how legal concepts have been applied to complex issues to produce simple tools (like Creative Commons licenses) to enable researchers and academics to protect themselves and their work. The easier the system is, the more likely it is to be used and promulgated. Based on the uptake of the Creative Commons license – it is clear that scholars are leveraging this resource to protect their intellectual property – but with the spirit of sharing it as broadly as possible in the process. This is certainly a welcome trend and one that promises to further encourage others in the process.

This issue also includes two special articles in the “Perspectives” section:

In John Wilbanks’ piece, “Cyberinfrastructure for Knowledge Sharing,” we see an intriguing outline of the many painful issues currently faced in achieving true scientific research in our current information environment. The core thesis behind Wilbanks’ article is that “…we aren’t sharing knowledge as efficiently as we could be” – meaning that even though the potential is there, we are not yet realizing the full potential presented to us by cyberinfrastructure. The content is there, the data is there, but the entire system and network is not yet fully “wired” and functioning for optimal efficiency. Indeed, Wilbanks posits, we’re not even close. To address this opportunity space, Scientific Commons was created to overcome hurdles related to (1) access to literature, (2) access to experimental materials, and (3) to encourage data sharing. Wilbanks describes some projects currently underway (e.g., Neurocommons), but also charges the community to address the challenge and make the most of the tools around them to push forward faster.
And finally, Peter Suber’s “Trends Favoring Open Access” article is a personal closing look at where things stand with Open Access. From the opening lines of this strongly-opinionated piece, you glean the fact that that Suber clearly views the Open Access movement as a joint endeavor (evidenced by the many mentions of “us” across the article), a community undertaking with this as a mid-term report card. In this assessment, he acknowledges the short-comings and inadequacies to date (namely, low deposit rates), but he also calls out the significant achievements – specifically the solid momentum and high-arching trajectory of Open Access. What is undeniable is the dramatic progress that has been made on all nearly all fronts – progress that would have been nearly unimaginable 5-10 years ago. That said, Suber is also quick to point out the areas where for-profit companies are continuing to consolidate their positions. There is a battle raging and a victory for Open Access is not yet assured—although the tide would appear to be turning.

Evidenced by the breadth of trends and topics addressed in this issue and the progress we are seeing in the environment overall, we are obviously at an inflection point in the world of Open Access + Cyberinfrastructure. Perhaps not at the fabled tipping point – yet – but we have clearly summated the crest of one range and find ourselves peering anxiously at the next (last?) set of mountains to conquer. There can be no question: it is no longer IF, but WHEN…and some would argue the coming revolution has already arrived.
The Shape of the Scientific Article in the Developing Cyberinfrastructure

Introduction

For the last few centuries, the primary vehicle for communicating and documenting results in most disciplines has been the scientific journal article, which has maintained a strikingly consistent and stable form and structure over a period of more than a hundred years now; for example, despite the much-discussed shift of scientific journals to digital form, virtually any article appearing in one of these journals would be comfortably familiar (as a literary genre) to a scientist from 1900. E-science represents a significant change, or extension, to the conduct and practice of science; this article speculate about how the character of the scientific article is likely to change to support these changes in scholarly work. In addition to changes to the nature of scientific literature that facilitate the documentation and communication of e-science, it's also important to recognize that active engagement of scientists with their literature has been, and continues to be, itself an integral and essential part of scholarly practice; in the cyberinfrastructure environment, the nature of engagement with, and use of, the scientific literature is becoming more complex and diverse, and taking on novel dimensions. This changing use of the scientific literature will also cause shifts in its evolution, and in the practices of authorship, and I will speculate about those as well here.

A few general comments should be made at the outset. First, I recognize that it is dangerous to generalize across a multiplicity of scientific disciplines, each with their own specialized disciplinary norms and practices, and I realize that there are ample counterexamples or exceptions to the broad trends discussed here; but, at the same time, I do believe that it is possible to identify broad trends, and that there is value in analyzing them. Second, as with all discussions of cyberinfrastructure and e-science, many of the developments and issues are relevant to scholarly work spanning medicine, the biological and physical sciences, engineering, the social sciences, the humanities, and even the arts, as is suggested by the increasingly common use of the more inclusive term "e-research" rather than "e-science" in appropriate contexts. I have focused here on the sciences and engineering, but much of the discussion has broader relevance.

Finally, it's crucial to recognize that the changes to the nature of scholarly communication and the scientific article are not being driven simply or solely by technological determinism as expressed through the move to e-science. There are broad social and political forces at work as well, independent of, but often finding common cause or at least compatibility with, e-science developments; in many cases, the transfigured economics and new capabilities of global high-performance networking and other information technologies are, for the first time, making it possible for fundamental shifts in the practices and structures of scholarly communication to occur, and thus setting the stage for political demands that these new possibilities be realized. Because the same technical and economic drivers have fueled much of the commitment to e-science, these other exogenous factors that are also shaping the future of scholarly communication are often, at least in my view, overly identified with e-science itself. Notable and important examples include the movements towards open access to scientific literature; movements towards open access to underlying scientific data; demands (particularly in the face of some recent high-profile cases of scientific fraud and misconduct) for greater accountability and auditibility of science through structures and practices that facilitate the verification, reproducibility and re-analysis of scientific results; and efforts to improve the collective societal return on investment in scientific research through a
recognition of the lasting value of much scientific data and the way that the investment it represents can be amplified by disclosure, curation and facilitation of reuse. Note that in the final area the investments include but go beyond the financial; consider the human costs of clinical trials, for example.

Scientific Articles and their Relationships to Data

The vast majority of scientific articles present and/or analyze data. (Yes, in mathematics, and some parts of theoretical physics, they do something else, and if, when and how this particular sub-genre of writings will be transformed is a fascinating question that deserves an extensive discussion in its own right. But that is another question, for another time.) As this data becomes more complex, more extensive, more elaborate, more community-based, more mediated by software, the relationships between articles and the data upon which they are based is becoming more complex and more variable. And recognize that implicit in these relationships are a whole series of disciplinary norms and supporting organizational and technical cyberinfrastructure services.

To what extent should articles incorporate the data they present and analyze, and to what extent should they simply reference that data? The issues here are profoundly complex. First, there's the question of whether the data is original and being presented for the first time; certainly it is commonplace to draw upon, compare, combine and perhaps reinterpret data presented in earlier articles or otherwise made available, and here the right approaches would presumably be citation or similar forms of reference. Repeated publication of the same data is clearly undesirable.

For newly publicized data there are a range of approaches. Some journals offer to accept it as "supplementary materials" that accompany the article, but often with very equivocated commitments about preserving the data or the tools to work with it, as opposed to the article proper. Not all journals offer this as an option, and some place constraints on the amount of data they will accept, or on the terms of access to the data (e.g., subscribers only).

For certain types of data, specific communities — for example crystallographers, astronomers, and molecular biologists — have established norms, enforced by the editorial policies of their journals, which call for deposit of specific types of data within an international disciplinary system of data repositories, and have the article make reference to this data by an accession identifier assigned upon deposit in the relevant repository. Clearly, this works best when there are well agreed-upon structures for specific kinds of data that occur frequently (genomic sequencing, observations of the sky, etc.); it also assumes an established, trustworthy and sustainable disciplinary repository system. Indeed, we have already seen the emergence of what might be characterized as a "stub article" that in effect announces the deposit of an important new dataset in a disciplinary repository and perhaps provides some background on its creation, but offers little analysis of the data, leaving that to subsequent publications. This allows the compilers of the dataset to have their work widely recognized, acknowledged, and cited within the traditional system familiar to tenure and promotion committees.

Another alternative is for the authors to store the underlying data in an institutional repository. While in some ways this is less desirable than using a disciplinary repository (due to potentials for economies of scale, easy centralized searching of material on a disciplinary basis, and for the development and maintenance of specialized discipline-specific software tools, for example) the institutional repository may be the only real option available for many researchers and for many types of data. Note that one function
The Shape of the Scientific Article in the Developing Cyberinfrastructure

(and obligation) of each institutional repository is to provide the depositing researcher with a persistent accession identifier that can be used to reference the data.

Recognize that over time individual researchers may move from institution to institution and will ultimately die; technical systems evolve, organizations change mission and responsibilities, and funding models and funding agency interests and priorities shift — any of which can cause archived data to have to be migrated from one place to another or reorganized. The ability to resolve identifiers, to go from citation to data, is highly challenging when considered across long time horizons. The research library community collectively has made a clear commitment to the long-term preservation of access to the traditional scientific literature; the assumption of similar ultimate responsibility for scientific and scholarly data today is still highly controversial.

Just because a dataset has been deposited into a repository does not automatically mean that other researchers (or indeed the public broadly) can have access to it. This is a question ultimately of disciplinary norms, of requirements imposed by funding agencies, of university policies, and of law and public policy. What the e-science environment does is to make these policies and norms much more explicit and transparent, and, I believe, to advance a bias that encourages more than less access and sharing. And there is still work to be done on mechanisms and legal frameworks — for example, the analogs of the Creative Commons type licenses for datasets are under development by Science Commons and others, but are at a much less mature stage than those used for journal articles themselves, with part of the problem being that the copyright framework that governs articles is much more consistent globally than laws establishing rights in datasets and databases.1 Also to be recognized here are certain trends in the research community – most notably university interests in technology transfer as a revenue stream, and the increasing overreach of some Institutional Review Boards in restricting the collection, preservation and dissemination of materials dealing in any way with human subjects – which run very much counter to the bias towards greater openness.

Setting aside the broad issue of the future of peer review, a particularly interesting set of questions involves the relationships between traditional peer review and the data that underlies an article under review. It’s often unclear the extent to which peer review of an article extends to peer review of the underlying data; even when the policies of a journal are explicit on this, I think it’s likely readers don’t have well-established expectations. Will there be a requirement that reviewers have access to underlying data as part of the review process, even if this data may not be immediately and broadly available when the article is published? A recent examination of editorial and refereeing policy by Science in the wake of a major incident of data falsification suggests that at least some journals may take a more aggressive and indeed even somewhat adversarial position with the authors of particularly surprising or high-visibility articles.2 And post-publication, there’s a very formalized means of correcting errors in published articles (or even withdrawing them) that’s now integrated into the online journal delivery systems (though not necessarily other open-access versions of articles that may be scattered around the net). Data correction, updating and maintenance take place (if at all) through separate curatorial mechanisms that are not synchronized to those for managing the article literature.

Visual Presentation of Data in Scientific Articles

The chemist Peter Murray-Rust speaks passionately and powerfully of the ways in which traditional presentations of information in scientific articles, such as graphs and charts, actually obscure or destroy data, invoking scenes of readers employing rulers to


2 See Donald Kennedy’s editorial at http://www.sciencemag.org/cgi/content/summary/314/5804/1353 and the report itself at http://www.sciencemag.org/cgi/content/full/314/5804/1353/DC1
try to estimate the actual values of coordinates of points in a graph. Clearly, in a digital environment, it would be much better to be able to move directly and easily between the underlying table of numerical values and their graphical representation, for example. Note such a table really is, in my view, intellectually an integral part of the article rather than underlying data (it may in fact be a complex derivative or reduction of the underlying data, or it might duplicate it). While the example of a two-dimensional graph is fairly straightforward, one can imagine a wide range of more specialized visualizing tools operating on various forms of structured data.

To me, resolving this problem implies a somewhat “thicker” layer of software mediating between the machine-readable representation of articles in the cyberinfrastructure environment and the human reader. Today, articles are most typically delivered to readers in very unexpressive, semantically limited forms such as PDF or HTML, which are rendered by a PDF viewer or a web browser, respectively. As we build out the laboratories and virtual workspaces to house the activities of our virtual organizations within the cyberinfrastructure, I hope that we will see a new generation of viewing and annotation tools deployed, presumably working on semantically rich XML document representations that will allow us to move beyond the kind of practices that Peter Murray-Rust so appropriately highlights as impeding the progress of scholarship.

The issues here are not limited to graphs and charts. Let me just give two other examples to illustrate the range of potential opportunities.

The astronomer Robert Hanisch gives a talk that includes this compelling example: An article includes an image of an astronomical object captured at a specific frequency range; a reader would like to place this object into the context of the available archive of astronomical observations and see what additional observations might be available at other wavelengths, for example. The process of manually re-creating what is, in effect, the digital provenance of the published image proves to be quite arduous, though once the source image in the archive context is established, it’s easy enough to check for the availability of additional imagery for the same region. Clearly, it would be very desirable to generate the trail of digital provenance as the image is prepared for publication, and to make an appropriate representation of that provenance available in the article along with the image to facilitate exactly the kinds of exploration Hanisch describes.

Finally, we are coming to a more sophisticated understanding of photography. A photographic print, or its reproduction on a printed page (or, indeed, a digital simulacrum of a printed page), is in effect a reduction of the image stored in a photographic negative or, more recently, an image dataset captured by a digital camera. Adjustments in focus, dynamic range, contrast, and similar parameters can yield radically different images from the same dataset. Here, again, it would clearly be desirable to have software tools that allow one to toggle between the rendering of an image dataset and the underlying dataset, and to be able to manipulate the key parameters that control the rendering.

All of these examples share common themes and raise common questions. The source article becomes more richly structured and exposes its semantics more explicitly. At the same time, the article becomes more and more intractable for humans to read without the correct set of mediating software. This means that the requisite mediating software must be highly reliable, simple to use, and ubiquitously available. Specialized software for individual journals or individual publishers will not, in my view, reach the necessary critical mass.

It’s easy to underestimate the problem of maintaining the level of quality and flexibility inherent in today’s visually oriented presentations. Consider the humble graph
The Shape of the Scientific Article in the Developing Cyberinfrastructure

— it’s not just a table of coordinate pairs; it has a caption, labels and markings on the axes, a default scale, perhaps a layer of annotations on some of the regions of the graph. We don’t want to lose any of these capabilities.

In terms of deployability and acceptance there are also questions about the potentially increased workload involved in preparing this new generation of scientific articles. While most publisher databases are already using XML based representations of articles internally and are carrying an increasing amount of tagging that can be put to good use, a careful analysis of the distribution of responsibilities between authors and publishers in areas such as the preparation of various kinds of “illustrations” is called for, as well as consideration of the tools that might be available to help the author.

Scientific Literature that is Computed Upon, Not Merely Read by Humans

In the previous section, we explored a few of the ways in which human readers may expand and extend their interactions with the scientific literature through the mediation of a new generation of software. But the use of the corpus of scientific literature is already changing in other ways as well: not only do human beings read (and interact with) articles from the scientific literature one article at a time, but we are also seeing the deployment of software that computes upon the entire corpus of scientific literature (perhaps restricted by discipline, and perhaps also federated with proprietary and/or unpublished literature or auxiliary databases). Such computation includes not only the now familiar and commonplace indexing by various search engines, but also computational analysis, abstraction, correlation, anomaly identification and hypothesis generation that is often termed “data mining” or “text mining.”

The implications of this shift are extensive and complex, but I want to sketch a few implications here. First, there will be greater demand for the availability of scientific literature corpora as part of the cyberinfrastructure, and for these corpora to be available in such a way — both technically and in terms of licensing, legal and economic arrangements — so as to facilitate ongoing computation on the literature by groups of collaborating researchers, including groups (“virtual organizations”) assembled often fairly casually from across multiple institutions. The barriers here are formidable: most commonly, access arrangements for publisher-controlled literature are established on an institutional basis; these licenses often specifically prohibit large-scale duplication of the text corpora for this kind of computational use; and today most publishers do not provide technical provisions for arbitrary computation of the type envisioned here.

More important to the changing nature of the individual article as opposed to the literature as a whole, the computational techniques that are applied to the current literature base make extensive use of heuristics (as well as various auxiliary databases, dictionaries, ontologies and other resources). Basically, they use algorithms to guess (with increasingly good accuracy) whether “Washington” in a given bit of text refers to a person, a state, or a city (and if so which one), whether something is the name of a gene, a chemical compound, a species, or other entity of interest. As we create new literature going forward, it makes sense to specifically tag some of these terms of interest to allow more accurate computation on the articles. Clearly, also, there are interesting possibilities of retrospectively tagging older literature, or even running the current best heuristics to provisionally tag the older literature, and then selectively (and perhaps collaboratively) applying human analysis to review provisional tags that are most suspect (Greg Crane and his colleagues at the Perseus Project have run some fascinating pilots of this type in doing markup of classical texts). The questions here are what entities merit tagging, how standards are set for tagging such entities, and what combination of
The Shape of the Scientific Article in the Developing Cyberinfrastructure

author and publisher should take responsibility for actually doing the markup? There are delicate issues about how to manage the evolution of the tagging over time and also how to manage it across disciplines in such a way to facilitate interdisciplinary work. There’s a difference between viewing the presence of tags as conclusive positive information and being able to count on the absence of a tag as conclusive negative information, for example.

**Paths of Change and Adoption**

What’s sure to happen? What is contingent or uncertain? What’s already happening?

The linkages between articles and underlying data are a very real and active (though far from fully settled) area of development today. All the models — supplementary materials as part of articles, disciplinary repositories and institutional repositories — are in use today.

The problems raised by visual displays of data in scientific articles are more widely recognized and frustrate more readers each day. Some problems — graphs for example — are easily solved in prototypes. Others, such as the appropriate way to express data provenance, remain active and challenging research areas. But the real challenges here are about deployment, scale, adoption, and standards. We have decades of examples of systems that support the creation and use of digital documents that are far more powerful and expressive than those in common today, but for a wide variety of reasons these systems never reached critical mass in deployment and adoption, which has proven hugely difficult to predict or ensure at Internet scale.

Text mining is a reality today, at least on a limited basis, and producing some results of real value. Here, I suspect, the barriers to progress will be more around business models for those journals that aren’t open access (some open access journals actually package up a compressed archive of all their articles and invite interested parties to simply copy the files and compute away; clearly this is not going to be as straightforward for a commercial publisher).

One thing is clear: in the cyberinfrastructure, the future of the article will be shaped not only by the requirements to describe changing scientific practices, but also by the changing nature of the authoring and usage environments for the scholarly literature that describes this changing science.

**Further Reading**

There are now a wealth of publications dealing with data and e-science; particularly valuable is the article “The Data Deluge” by Hey & Trefethen; the NSF/NSB report on Long-Lived Datasets; the report of the NSF/ARL workshop “To Stand the Test of Time”; the current NSF Office of Cyberinfrastructure vision document; and, a UK perspective. On institutional repositories, see the Lynch article in the ARL Bimonthly Report for more information on the implications of literature mining, see for example the article in the book Open Access: Key Strategic, Technical, and Economics Aspects.
Next-Generation Implications of Open Access

Introduction

The technological transformation of scholarly communication infrastructure began in earnest by the mid-1990s. Its effects are ubiquitous in the daily activities of typical researchers, instructors and students, permitting discovery, access to, and reuse of material with an ease and rapidity difficult to anticipate as little as a decade ago. An instructor preparing for lecture, for example, undertakes a simple web search for up-to-date information in some technical area and finds not only a wealth of freely available, peer reviewed articles from scholarly publishers, but also background and pedagogic material provided by its authors, together with slides used by authors to present the material, perhaps video of a seminar or colloquium on the material, related software, on-line animations illustrating relevant concepts, explanatory discussions on blog sites, and often useful notes posted by 3rd party instructors of similar recent or ongoing courses at other institutions. Any and all of these can be adapted for use during lecture, added to a course website for student use prior to lecture, or as reference material afterwards. Questions or confusions that arise during lecture are either resolved in real time using a network-connected laptop, or deferred until afterwards, but with instructor’s clarification propagated instantly via a course website or e-mail. Or such lingering issues are left as an exercise for students to test and hone their own information gathering skills in the current web-based scholarly environment. Some courses formalize the above procedures with a course blog that also permits posting of student writing assignments for commentary by other students and the instructor. Other courses employ wikis so that taking lecture notes becomes a collaborative exercise for students.

Many of these developments had been foreseen a decade ago, at least in principle, though certainly not in all the particulars. When the mass media and general public became aware of the Internet and World Wide Web in the mid-1990’s, this new “information superhighway” was heavily promoted for its likely impact on commerce and media, but widespread adoption of social networking sites facilitating file, photo, music, and video sharing was not regularly touted. Web access is now being built into cell-phones, music players, and other mobile devices, so it will become that much more ubiquitous in the coming decade. People currently receiving their PhDs became fluent in web search engine usage in high school, and people receiving their PhDs a decade from now will have had web access since early elementary school. (My 3.5 year old son has had access to web movie trailers on demand since the age of 1, but is at least two decades from a doctorate.)

Many aspects of teaching and scholarship will remain unchanged. Web access will not fundamentally alter the rate at which students can learn Maxwell equations for electromagnetism, and many web resources of uncertain provenance, e.g., Wikipedia entries, will require independent expertise to evaluate. We’ve also already learned that having a vast array of information readily available does not necessarily lead to a better informed public, but can instead exacerbate the problem of finding reliable signal in the multitude of voices. Recent political experience suggests that people tend to gravitate to an information feed that supports their preexisting notions: the new communications technologies now virtually guarantee that it will exist, and moreover make it easy to find. In the below, I will focus on questions related to the dissemination and use of scholarly research results, as well as their likely evolution over the next decade or
so. The issues of generating, sharing, discovering, and validating these results all have parallels in non-academic pursuits. In order to guide the anticipation of the future, I’ll begin by looking backwards to developments of the early 1990’s.

1. Background

The e-print arXiv\(^1\) was initiated in Aug 1991 to accommodate the needs of a segment of the High Energy Physics theoretical research community. (For general background, see Ginsparg, P. “Winners and losers….\(^2\)” Due to its prior history of paper preprint distribution (for a history, see O’Connell, H. B. “Physicists Thriving…”\(^3\)), it was natural for this community to port its prepublication distribution to the internet, even before the widespread adoption of the World Wide Web in the mid 1990’s. A primary motivation for this electronic initiative was to level the research playing field, by providing equal and timely access to new research results to researchers all over the globe and at all levels, from students to postdocs to professors.

arXiv has since effectively transformed the research communication infrastructure of multiple fields of physics and will likely continue to play a prominent role in a unified set of global resources for physics, mathematics, and computer science. It has grown to contain over 430,000 articles (as of July 2007), with over 56,000 new submissions expected in calendar year 2007, and over 45 million full text downloads per year. It is an international project, with dedicated mirror sites in 16 countries, collaborations with U.S. and foreign professional societies and other international organizations, and has also provided a crucial life-line for isolated researchers in developing countries. It also helped initiate, and continues to play a leading role, in the growing movement for open access to scholarly literature. The arXiv is entirely scientist driven: articles are deposited by researchers when they choose – either prior to, simultaneous with, or post peer review – and the articles are immediately available to researchers throughout the world.

As a pure dissemination system, it operates at a factor of 100 to 1000 times lower in cost than a conventionally peer-reviewed system.\(^4\) This is the real lesson of the move to electronic formats and distribution: not that everything should somehow be free, but that with many of the production tasks automatable or off-loadable to the authors, the editorial costs will then dominate the costs of an unreviewed distribution system by many orders of magnitude. Even with the majority of science research journals now online, researchers continue to enjoy both the benefits of the rapid availability of the materials, even if not yet reviewed, and open archival access to the same materials, even if held in parallel by conventional publishers. The methodology works within copyright law, as long as the depositor has the authority to deposit the materials and assign a non-exclusive license to distribute at the time of deposition, since such a license takes precedence over any subsequent copyright assignment.

The site has never been a random UseNet newsgroup- or blogspace-like free-for-all. From the outset, arXiv.org relied on a variety of heuristic screening mechanisms, including a filter on institutional affiliation of submitter, to ensure insofar as possible that submissions are at least “of refereable quality.” That means they satisfy the minimal criterion that they would not be peremptorily rejected by any competent journal editor as nutty, offensive, or otherwise manifestly inappropriate, and would instead at least in principle be suitable for review. These mechanisms are an important – if not essential – component of why readers find the arXiv site so useful.

The arXiv repository functions are flexible enough either to co-exist with the pre-existing publication system, or to help it evolve to something better optimized for

\(^1\) http://arXiv.org/
These surveys were conducted independently by the CERN library during May/Jun 2007 and by the American Physical Society at its Mar 2007 and Apr 2007 meetings of condensed matter and high energy physicists, respectively. I thank S. Mele and M. Doyle for access to the results.


Some of the resources accessed were:
- http://adsabs.harvard.edu/ (The Smithsonian/NASA Astrophysics Data System),
- http://simbad.u-strasbg.fr/simbad/sim-fid (SIMBAD),
- http://nedwww.ipac.caltech.edu/ (NASA/IPAC extragalactic database (NED)),
- http://chandra.harvard.edu/pub.html (Chandra X-ray center),
- http://sci.esa.int/stri/mm/ (XMM-Newton),
- http://aladin.u-strasbg.fr/ (Aladin Sky atlas),
- http://heasarc.gsfc.nasa.gov/ (HEASARC missions and object catalogs),
- http://tdc-www.harvard.edu/uzs/ (Glinsky Catalog),

Next-Generation Implications of Open Access

2. Conservative user preferences

In recent polls, users of physics information servers rated as most important their breadth and depth of coverage, comprehensiveness, timeliness, free and readily accessible full-text content, powerful search index, organization and quality of content, spam filtering, non-commerciality, and convenience of browsing. They also emphasized the importance of notification functions (newly received articles, author/keyword/reference alerts), whether by e-mail, separate webpage listing, or RSS feed. Users have grown to expect seamless access to older articles, most conveniently from one universal portal. Considered less important were user friendliness and general ease of use, quality of submission interface, availability of citation analysis, measures of readership, multimedia content, personalization, keywords and classification, and other collaborative tools. Users also mentioned the expected future utility of citation and reference linkages to open access articles, comment threads, access to associated data and ancillary material (data underlying tables and figures, code fragments underlying plots), and various forms of smarter search tools. A majority reports being willing to spend at least a few minutes per day tagging materials for collaborative filtering purposes.

More than 80% of respondents reported a desire for article synopses and overviews, although these would likely be just as labor-intensive to produce as in the paper age. A variety of other navigation tools were suggested, some potentially automatable, including personalized table of contents, and flow diagrams representing the relations among research articles on a topic. Some suggested in addition a descriptive page for each research area including lists of recent reviews, cutting-edge research articles, experimental results, and primary researchers involved, with links to their own web pages. Many of the desired features are prompted by the increasing wealth of information available in bibliographic and other databases. Their realization will rely on tools for organizing a hierarchical literature search and that are able to preserve the context of retrieved items. Users also extol the potential utility of having conference presentations in the form of slides or video linked to articles, and more generally having linked together all instances of a given piece of research: notes, thesis, conference slides, articles, video, simulations, and associated data.

The connection between scientific literature and data in astronomy and astrophysics was illustrated by Kurtz. The new research methodology employed can involve multiple steps back and forth through a bibliographic database: following citations, references, looking for other articles by the same authors, or searching for keywords in...
abstracts of articles. It can also involve looking for related articles, or related objects, or articles about those objects in over a dozen different databases, e.g., finding reference codes of objects, finding the experimental source, examining the observation itself, finding publications at the website of experiment, following references to other external archives, including finding object catalogs and active missions, checking a sky atlas or digital sky survey at yet another site, checking an extragalactic database, cross-checking an archival catalog search, etc. While the possibilities enabled by these distributed resources are already quite impressive and result in fundamental improvements over prior methods, various manual steps in the search process are quite awkward due to the independent configurations of all the pieces of the system. Many navigation functions could be automated and usefully centralized if the separate databases were set up to interoperate more efficiently. Analogs of these inefficiencies exist in other fields of research.

“Next generation” implications depend significantly on what the next generation of users will expect. Some clues in this respect are available from surveying the generic functionality of current commonly used websites. Many top-level browsing features are held in common among sites devoted to scholarship on the one hand, and to those designed for shopping, entertainment and popular file-sharing on the other, including: browse by groups, categories, subject area, most recent, or by a variety of “popularity” measures including recently featured, most viewed, top rated, most discussed, top favorites, most linked, most honored, most shared, most blogged, or most searched. The convergence of features in the parallel realms is also evident in pages devoted to specific items, including standard descriptive metadata (title, author(s), submitter), links to browse “related items,” “more from this user,” “related keywords” (both local at the site or linked to 3rd party aggregators), and in collaborative features including functions to add tags and labels, rate items, flag as inappropriate, save to favorites, add to groups, share/e-mail to friend, blog item, post to 3rd party site, and add or read comments and responses. Some features specific to publisher sites are links to retrieve full text and supplemental data, show references/citations, addenda/corrigenda, related web pages, export citation, cite or link using DOI, alert when cited, find same object in 3rd party database, search 3rd party database, or find similar articles by various flavors of relatedness (text similarity, co-citation, co-reference, co-readership). Many sites include a subscribe function, with the option to be alerted to new issues or when specific keywords appear, the possibility to upload content, and as well there are various forms of personalization, including provisions for a private library of “my articles” with view, and add/subtract functions. These private libraries can optionally be made open to other users, providing a potential collective enhancement, though with possible privacy issues.

All of the above can be expected to comprise the core functionality of future sites intended for scholarly participation. A recently implemented example is Nature Precedings, a free service from the publishers of Nature that permits researchers in the biological and life sciences to share preliminary findings, solicit community feedback, and stake priority claims. It distributes preprints as does arXiv, but it also accepts other document types such as posters and presentations. The site advertises its intention to help researchers find useful content through collaborative filtering features including tagging, voting and commenting. It is possible that preprint and file sharing have historically been impeded in the biological and life sciences communities due to the perception that such prior appearance could prevent later appearance of the work in premier journals. Since it is connected to a premier journal, Nature Precedings could significantly foster acceptance of dissemination of prepublication and ancillary materials within these important communities.
In another example of use of new technologies to port social research networking to a more distributed form, arXiv began accepting blog trackbacks in 2005. This technology permits blog software to signal to arXiv.org that a blog entry discusses a specific article, and a link added from the relevant article abstract at arXiv.org back to the blog site can then facilitate discovery of useful comment threads by readers. Blogging can require a substantial time commitment, but nonetheless a number of serious researchers have joined in and provided rich content, and provided links to other informative resources that wouldn’t otherwise be readily discoverable. arXiv.org now points back from about 7000 articles to about 2000 blog entries, still a small volume. The underlying idea is to replicate in some on-line form the common experience of going to a meeting or conference, and receiving from a friend/expert some informal recent research thoughts and an instant overview of a subject area. Though without the in-person contact, the blog links provide some semblance of the above discussion framework and are moreover available to all, helping to level the playing field just as does the open article dissemination. It is not yet known whether the useful blogger lifetime will be months, years, or decades, but more researcher bloggers are currently joining in than are dropping out. Perhaps trackbacks from a heavily used archival site such as arXiv will provide some additional incentive for bloggers, giving comfort that they’re not just typing into the wind. While the current number of bloggers remains a minuscule percentage of the number of authors, it is possible that externally moderated discussion fora, where people can post occasional comments non-anonymously without having to maintain their own dedicated long-term blogs, will be the most important long-term usage.

Before expecting too rapid a pace of change, however, it is useful to consider as well some current habits of the high energy and condensed matter physicists surveyed above. A large majority reports using arXiv as its primary information source, and a large majority has personal web pages. But only a small percentage (< 10%) uses RSS feeds and less than 20% listens to podcasts. Only a small percentage (< 10%) follows blogs regularly, a smaller percentage participates in blog discussions, and an even smaller percentage (< 1%) maintains its own blogs. Less than 10% have tried any social bookmarking sites, and only 1% found them useful. To be fair, many of these new resources have become widespread only in the past 3-4 years, so there may be an adoption lag for people already past their PhD’s and already focused on research. Past generations of users can be expected to expand their repertoires as new features become commonplace at Internet commerce and other non-research sites, but can’t be relied upon to anticipate the most useful future features.

3. Open Access: inevitable, impossible, mandatory or backdoor?

There is currently much discussion of free access to the on-line scholarly literature. It has long been argued that this material becomes that much more valuable when freely accessible, and moreover that it is in public policy interests to make the results of publicly funded research freely available as a public good. It is also suggested that the move to open access could ultimately lead to a more cost-efficient scholarly publication system. There are recent indications that the U.S. and other governments may become directly involved, by mandating some form of open access to research funded by government agencies. The message to legislators is deceptively short and simple: “The taxpayers have paid for the research so deserve access to the results.” The counter-argument is somewhat more subtle and takes paragraphs to elucidate, so the U.S. congress can be expected to legislate some form of open access, beginning with a requirement that articles based on certain forms of federally funded research be deposited in public repositories within one year of publication. It may seem


non sequitur to force researchers to act in what should be their self-interest, while the general public spontaneously populates file sharing sites such as photobucket and YouTube, but such is the current politics of scholarly publication.

The response of the publishing community is essentially that their editorial processes provide a critical service to the research community, that these are labor-intensive and hence costly, and that even if delayed, free access could impair their ability to support these operations. In short, it costs real money to implement quality control by the time-honored methodology. If that methodology is maintained, the requisite funds must continue to flow to the same, or perhaps more cost-efficient, intermediaries between authors and readers. If the flow of funds is reduced, then a different methodology for quality control and authentication of the materials will be required. The basic tension is that authors and readers want some form of quality control, but the most efficient mechanism for providing it, and for paying for it, is still unclear. The problems of trust, integrity, and authentication mentioned earlier for the web at large remain critical to the scholarly communities from which it sprang.

A complicating factor is that the current costs of doing business vary significantly from publisher to publisher, as do the profits. One proposal is for authors to pay for publication once an article is accepted, making it free for all to read without subscription costs. As a side-effect, this proposal exposes the real costs of publication, ordinarily visible to the libraries paying the subscriptions but not to the researchers themselves. If this provides a mechanism to influence author choice of journals and if lowered profit margins necessary to attract authors persuade some of the more profitable commercial publishers to shift to other more lucrative endeavors, then other entities would still have to be available to fill the large gap in capacity left by their departure.

There are not only hierarchies of cost, but also hierarchies of funding from research discipline to research discipline. The average amount of research funding per article can vary from a few hundred thousand dollars per article in some areas of biomedical research, to zero for the majority of mathematicians who have no grant funding at all.11 The areas with higher levels of funding per article are more likely to be able to take advantage of an author-pays model. Another recent proposal12 within the High Energy Physics community is sponsorship of journals by large organizations, starting with major research laboratories. The concern is again the long-term sustainability of the commitment – while there is a loss of access for failure to pay subscriptions, there’s no immediate downside for failure to meet sponsorship commitments. The “we don’t do charity” sentiment expressed by librarians is also understandable. So with subscriptions already on the decline since long before the advent of Internet access, it is difficult to argue that journals should accept the transition to sponsored open access, not knowing whether or not it would be permanent. While some new open access journals13 are accepted by scientists, there are, as pointed out by Blume,14 no examples of any journal of significant size that has been converted from subscription to open access, and few if any open access examples of sustained cost recovery.

Studies have shown a correlation between openly accessible materials and citation impact,15 though a direct causal link is more difficult to establish, and other mechanisms accounting for the effect are easily imagined. It is worthwhile to note, however, that even if some articles currently receive more citations by virtue of being open access, it doesn’t follow that the benefit would continue to accrue through widespread expansion of open access publication. Indeed, once the bulk of publication is moved to open access, then whatever relative boost might be enjoyed by early adopters would long since have disappeared, with relative numbers of citations once again determined by the usual independent mechanisms. Citation impact per se is consequently not a serious argument for encouraging more authors to adopt open access publication. A
different potential impact and benefit to the general public, on the other hand, is the greater ease with which science journalists and bloggers can write about and link to open access articles.

A form of open access appears to be happening by a backdoor route regardless: using standard search engines, over a third of the high impact journal articles in a sample of biological/medical journals published in 2003 were found on non-journal websites. Informal surveys of publications in other fields, freely available via straightforward web search, suggest that many communities may already be further along in the direction of open access than most realize. Most significantly, the current generation of students has grown up with a variety of forms of file and content sharing, legal and otherwise. This generation greets with dumbfounded mystification the explanation of how researchers perform research, write an article, make the figures, and then are not permitted to do as they please with the final product. Since the current generation of undergraduates, and next generation of researchers, already takes it for granted that such materials should be readily accessible from anywhere, it is more than likely that the percentage of backdoor materials will only increase over time, and that the publishing community will need to adapt to the reality of some form of open access, regardless of the outcome of the government mandate debate.

4. What will Open Access Mean?

There is more to open access than just free access. True open access permits any 3rd party to aggregate and data-mine the articles, themselves treated as computable objects, linkable and interoperable with associated databases. The range of possibilities for large and comprehensive full text aggregations are just starting to be probed. The PubMed Central database, operated in conjunction with GenBank and other biological databases at the U.S. National Library of Medicine, is a prime exemplar of a forward-looking approach. It is growing rapidly and (as of Jun 2007) contains over 333,000 recent articles in fully functional XML from over 200 journals (and additionally over 683,000 scanned articles from back issues). A congressionally mandated open access policy for NIH supported publications would generate an additional 70,000 articles a year for PubMed Central.

The full text XML documents in this database are parsed to permit multiple different “related material views” for a given article, with links to genomic, nucleotide, inheritance, gene expression, protein, chemical, taxonomic, and other databases. For example, GenBank accession numbers are recognized in articles referring to sequence data and linked directly to the relevant records in the genomic databases. Protein names are recognized and their appearances in articles linked automatically to the protein and protein interaction databases. Names of organisms are recognized and linked directly to the taxonomic databases, which are then used to compute a minimal spanning tree of all the organisms contained in a given document. In yet another “view,” technical terms are recognized and linked directly to the glossary items in the relevant standard biology or biochemistry textbook in the books database. Sets of selected articles resulting from bibliographic queries can also have their aggregated full texts searched simultaneously for links to over 25 different databases, including those mentioned above. The enormously powerful sorts of data-mining and number-crunching, already taken for granted as applied to the open access genomics databases, can be applied to the full text of the entirety of the biology and life sciences literature, and will have just as great a transformative effect on the research done with it.

On the one decade time scale, it is likely that more research communities will join some form of global unified archive system without the current partitioning and access
restrictions familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge. The genomic and related resources described above are naturally interlinked by virtue of their common hosting by a single organization, a situation very different from that described earlier for astronomy research. For most disciplines, the key to progress will be development of common web service protocols, common languages (e.g., for manipulating and visualizing data), and common data interchange standards, to facilitate distributed forms of the above resources. The adoption of these protocols will be hastened as independent data repositories adopt dissemination of seamlessly discoverable content as their raison d'être. Analogs of the test parsings described above have natural analogs in all fields: such as astronomical objects and experiments in astronomy; mathematical terms and theorems in mathematics; physical objects, terminology, and experiments in physics; chemical structures and experiments in chemistry, etc., and many of the external databases to provide targets for such automated markup already exist.

One of the surprises of the past two decades is how little progress has been made in the underlying document format employed. Equation-intensive physicists, mathematicians, and computer scientists now generally create PDF from TeX. It is a methodology based on a pre-1980's print-on-paper mentality and not optimized for network distribution. The implications of widespread usage of newer document formats such as Microsoft's Open Office XML or the OASIS OpenDocument format and the attendant ability to extract semantic information and modularize documents are scarcely appreciated by the research communities. Machine learning techniques familiar from artificial intelligence research will assist in the extraction of metadata and classification information, assisting authors and improving services based on the cleaned metadata. Semantic analysis of the document bodies will facilitate the automated interlinking to external resources described above and lead to improved navigation and discovery services for readers. A related question will be what authoring tools and functions should be added to word processing software, both commercial and otherwise, to provide an optimal environment for scientific authorship? Many of the interoperability protocols for distributed database systems will equally accommodate individual authoring clients or their proxies, and we can expect many new applications beyond real-time automated markup and autonomous reference finding.

Every generation thinks it's somehow unique, but there are nonetheless objective reasons to believe that we are witnessing an essential change in the way information is accessed, the way it is communicated to and from the general public, and among research professionals – fundamental methodological changes that will lead to a terrain 10-20 years from now more different than it was 10-20 years ago than in any comparable time period.
What is Web 2.0?

Perhaps the only thing on which everyone can agree about Web 2.0 is that it has become a potent buzzword. It provokes enthusiasm and cynicism in roughly equal measures, but as a label for an idea whose time has come, no one can seriously doubt its influence.

So what does it mean? Web 2.0 began as a conference, first hosted in October 2004 by O’Reilly Media and CMP Media. Following the boom-bust cycle that ended in the dot-com crash of 2001, the organisers wanted to refocus attention on individual web success stories and the growing influence of the web as a whole. True, during the late 1990s hype and expectations had run ahead of reality, but that did not mean that the reality was not epochal and world-changing. By the following year, Tim O’Reilly, founder of the eponymous firm and principal articulator of the Web 2.0 vision, had laid down in a seminal essay a set of observations about approaches that work particularly well in the online world. These included:

- “The web as a platform”
- The Long Tail (e.g., Amazon)
- Trust systems and emergent data (e.g., eBay)
- AJAX (e.g., Google Maps)
- Tagging (e.g., del.icio.us)
- Peer-to-peer technologies (e.g., Skype)
- Open APIs and ‘mashups’ (e.g., Flickr)
- “Data as the new ‘Intel Inside’” (e.g., cartographical data from MapQuest)
- Software as a service (e.g., Salesforce.com)
- Architectures of participation (e.g., Wikipedia)

The sheer range and variety of these concepts led some to criticize the idea of Web 2.0 as too ill-defined to be useful. Others have pointed out (correctly) that some of these principles are not new but date back to the beginning of the web itself, even if they have only now reached the mainstream. But it is precisely in raising awareness of these concepts that the Web 2.0 meme has delivered most value. Now, those of us without the genius of Jeff Bezos or Larry Page can begin to glimpse what the web truly has to offer and, notwithstanding the overblown hype of the late 1990s, how it really is changing the world before our eyes.

Initially the first item in the list above – the web as platform – seemed to have primacy among the loose collection of ideas that constituted Web 2.0 (see, for example, Figure 1 in [1]). The most important thing seemed to be that talent and enthusiasm in software development was migrating from traditional operating system platforms to the web. New applications were agnostic with respect to Unix versus Macintosh versus Windows and were instead designed to operate using web protocols (specifically, HTTP and HTML) regardless of the precise underlying software running on the server or client machines.

However, this view taken on its own overlooks one very important reason why that migration has happened: the web is more powerful than the platforms that preceded it because it is an open network and lends itself particularly well to applications that enable
collaboration and communication. With his usual eye for pithy phrasing, Tim O’Reilly described this aspect using the terms “architecture of participation” and “harnessing collective intelligence.” He pointed out that the most successful web applications use the network on which they are built to produce their own network effects, sometimes creating apparently unstoppable momentum. This is how a whole new economy can arise in the form of eBay, why tiny craigslist and Wikipedia can take on the might of mainstream media and reference publishing, and why Google can produce the best search results by surreptitiously recruiting every creator of a web link to its cause. In time, this participative aspect came to the fore, and these days “Web 2.0” is often seen as synonymous with websites that do not merely serve users but also involve them, thus enabling them to achieve that most desirable of business goals: a service that gets better for everyone the more people use it.

This brief survey will use a relatively broad definition of Web 2.0. So, while it will deal mainly with participative services and network effects, it will also cover other aspects of the original Web 2.0 vision that have particular relevance in science, including mashups and tagging.

Social software

If a cornerstone of the Web 2.0 meme is the web as a global, collaborative environment, how is this being put to use in perhaps the most global and collaborative of all human endeavors: scientific research? An irony often observed by those of us working in science communication is the fact that, although the web was originally invented as means for sharing scientific information, scientists have been relatively slow to fully embrace its potential. Blogging, for example, has become undeniably mainstream, with the number of bloggers somewhere in the high tens of millions (among a billion or so web users). Yet among a few million scientists worldwide, only perhaps one or two thousand are blogging, at least about science, and most of these are relatively young. By contrast, academic economists, for example, even very distinguished ones, seem to have embraced this new medium more enthusiastically.

Scientific blogging is still a niche activity, and what data there are suggest that it is not yet growing fast. For example, Alexa reports that ScienceBlogs, where many of the most prominent scientist-blogger post their thoughts, has shown little traffic growth over the last twelve months, and the scientific blog tracking service Postgenomic.com (created by an employee of Nature Publishing Group) shows the volume of posts from the blog in its index holding still at about 2,500 posts a week. Similarly, scientists appear reluctant to comment publicly on research papers. The blogging bug, it seems, has yet to penetrate the scientific citadel. This is a shame because blogs are a particularly effective means for one-to-many and many-to-many communication, and science no less than other spheres stands to gain from its judicious adoption.

Yet the participative web is about much more than blogging and commenting. The diagram below summarizes the manifold types of social software that exist online, all of them relevant in some way to scientific research.

Wikis: These have existed since the mid-1990s, but it took the astonishing rise of Wikipedia during the middle part of this decade for the potential of wikis to become widely appreciated. We can now see numerous examples of scientific wikis, from collaborative cataloguing and annotation projects like WikiSpecies and Proteins Wiki to open laboratory notebooks like OpenWetWare and UsefulChem. These all represent sensible uses of wikis, which are best employed to enable groups of geographically dispersed people to collaborate in the creation of a communal document with
an identifiable objective aim (as in Wikipedia, WikiSpecies and Proteins Wiki), or to allow individuals or small, real-world teams to share freeform information with others around the world (as in OpenWetWare and UsefulChem). In contrast, experiments at the *Los Angeles Times*\(^{21}\) and Penguin Books\(^{22}\) have demonstrated that wikis are not well suited to the creation of opinioned or fictional content – because the end goal cannot possibly be shared by all contributors at the outset. A particularly interesting recent development has been the launch of Freebase,\(^{23}\) the latest brainchild of parallel computing pioneer and polymath Danny Hillis. This takes a wiki-like approach to open contributions, but provides an underlying data model more akin to relational databases and the Semantic Web,\(^{24}\) allowing specific relationships between entities to be expressed and queried. Whilst Freebase is not aimed mainly at scientists, scientific topics are among those covered. It will be interesting to see how this approach fares over the less technically sophisticated but arguably less restrictive approach represented by traditional wikis.

**Voting:** Slashdot\(^{25}\) and more recently digg\(^{26}\) have become staple information sources for computer nerds and web geeks everywhere. Their traffic, which ranks them among the top media organisations on the planet,\(^{27}\) belies their meager staff numbers (which, compared to a daily newspaper’s, are as near to zero as makes no difference). Like all good Web 2.0 sites, they exert their influence by getting readers to contribute: in this case by providing stories, links and comments – then other users to decide what’s most interesting by casting votes. In the case of digg, the users even decide which stories get elevated to the front page. Such sites, like search engines, are sometimes criticized for being parasitical on the mainstream media stories to which they link (after all, they generate no content, only link to it). But this is to misunderstand the value they add, which is to help people decide where to direct their scarce attention in an age of often oppressive information overload. They are no more parasitical on journalism than journalism is on the newsmakers themselves (after all, journalists don’t make the news, only report it – well, most of the time). Yet these services do have a very different feel to those in which the content is selected by an editor, and the optimum approach in some cases may be to marry the ‘wisdom of crowds’ (to highlight interesting stories) with professional editorial expertise (to provide a final selection and put these items in context). These systems are also vulnerable to the ‘tyranny of the majority’ and to
cynical gaming, so even while they save on traditional editorial staff, the operators of these sites do face other challenges in maintaining a useful service.

Of course, similar problems of information overload apply in science, so it is natural to ask whether it is possible to use these approaches to help scientists to help themselves. Sure enough, sites like ChemRank,28 SciRate29 and BioWizard30 have appeared. Nature Publishing Group has a few of its own experiments in this area, including: DissectMedicine,31 a collaborative news system for medics; Nature China,32 which includes summaries of the best Chinese research as submitted and voted on by readers; and Scintilla,33 a scientific information aggregation and personalization tool that employs user ratings in its recommendation algorithms. It is too early to say which of these scientific applications will prevail, but given the demonstrable success of this approach outside science, it seems almost inevitable that some of them will.

**File sharing:** This is one of those rare areas in which scientists – or at least some of them – have blazed a trail well ahead of the mainstream. Physicists (and a few others) have been sharing preprints (unpeer-reviewed manuscripts) through the arXiv.org server34 since 1991 (and even before that, they shared their findings with each other by email or post). Now, the web is replete with ways of sharing various types of content, from documents35 to videos36 to slides.37 And scientific services, too, have begun to diversify, from Nature Precedings,38 a preprint server and document-sharing service for those outside physics, and the Journal of Visualized Experiments,39 a way for scientists to share videos of experimental protocols.

**Social networks:** Perhaps the most obviously social of all social software are those that enable the creation of personal networks of friends and other like-minded people. The use of services like MySpace40 and Facebook41 has become almost ubiquitous among young people in many countries. The average age of users is now starting to grow as they break away from their core teenage and college student markets.42 Meanwhile, LinkedIn43 has become a favourite networking tool among business people. Once again, medics and scientists are following the mainstream with sites like Sermo44 for clinicians and Nature Network for scientists.45 These environments are not only for finding and contacting new people with shared interests (though they are good for that too and therefore have potential in everything from job-seeking to dating), they also enable the creation of discussion groups and allow users to efficiently follow the activities (e.g., in terms of posts and comments) of others whose views they find interesting. Correctly implemented and used, these services therefore have great potential to make scientific discourse more immediate, efficient and open. A major unanswered question, however, is the interoperability and openness of the services themselves. No one wants to have to register separately on multiple different sites or lock up their details in a system over which they have no control. Federated authentication technologies like OpenID46 and other approaches to interoperability hold promise, but it remains to be seen how enthusiastically they will be embraced by the operators of social networking services, and how receptive they will be to the idea of partial cooperation rather than outright competition.

**Classified advertising:** This may seem like a strange category to include here, but newspaper small ads are arguably the original grassroots participative publishing service. It is perhaps no coincidence, then, that they have been among the first areas of traditional publishing to fall victim to lean and radical Web 2.0 startups, most famously craigslist.47 Particularly among careers services, there is also keen competition to turn simple ads services into social networks, as epitomized by Jobster,48 and the distinction between social networks and career services is only likely to blur further. Though some very large employers, notably Britain’s National Health Service49 have established their own online jobs boards, effectively disintermediating their former advertising.

---

29 SciRate – http://scirate.com/
31 DissectMedicine – http://www.dissectmedicine.com/
33 Scintilla – http://scintilla.nature.com/
36 YouTube – http://www.youtube.com/
37 SlideShare – http://www.slideshare.net/
38 Nature Precedings – http://precedings.nature.com/
40 MySpace – http://www.myspace.com/
41 Facebook – http://www.facebook.com/
42 Gonzalez, N. Facebook Users Up 89% Over Last Year, Demographic Shift. TechCrunch, 2007.
43 LinkedIn – http://www.linkedin.com/
44 Sermo – http://www.sermo.com/
45 Nature Network – http://network.nature.com/
46 OpenID – http://openid.net/
47 craigslist – http://sfbay.craigslist.org/
48 Jobster – http://www.jobster.com/
49 NHS Jobs – http://www.jobs.nhs.uk/
outlets, this revolution has yet to hit the medical and scientific advertising realm with full force. One early sign of the changes to come was the switch by NatureJobs in late 2006 from an arrangement in which online ads were sold as part of a portfolio of products to a ‘freemium’ model\(^50\) in which simple online listings are provided free and other services such as rich, targeted or print advertisements are sold as add-ons. This reflects the different economics of operating online, where the marginal cost to serve an extra advertiser is low and the benefits of providing a single comprehensive jobs database high.

**Markets:** EBay\(^51\) is in some ways the definitive Web 2.0 company: it is a pure market in which the company itself does not own any of the goods being traded. Similarly, other services, such as Elance\(^52\) specialize in matching skilled workers to employees with projects that they wish to outsource. In the scientific space, the online trading of physical goods (such as used laboratory equipment) is not yet commonplace, though it might become so in the future. In contrast, the matching of highly trained people to problems does have some traction in the form of ‘knowledge markets’ such as InnoCentive.\(^53\) These are still at an early stage, and they are mostly used by commercial organisations such as pharmaceutical companies, but it is not hard to imagine academic research groups doing the same one day if (as it should) this approach enables them to achieve their goals more quickly and at lower cost.

**Virtual worlds:** By far the most prominent virtual world is Second Life\(^54\) (though others such as There.com exist too). What sets it apart from online role-playing games like World of Warcraft (which are orders of magnitude more popular) are the facts that they do not have predefined storylines or goals and that they give their users freedom to create and use almost whatever objects they choose, possibly making money in the process. In this sense, they represent a genuine alternative to the real world. Pedants might argue that Second Life is not really a Web 2.0 service because it is not technically part of the web (i.e., it does not use HTML and HTTP, though it can interact with the web in various ways). But at a more abstract level, the participative, user-generated environments that have grown inside Second Life are as good examples as exist anywhere of the ‘architecture of participation’ principle. The greatest scientific potential seems to lie in education and in conferences. Second Life provides an environment in which people from different locations can come together quickly and easily into a shared space that to some extent mimics the real world in important aspects of human communication such as physical proximity, gesture and the ability to seamlessly mix one-to-many with one-to-one communication (e.g., chatting to the person beside you during a lecture). As a result, educators have poured in – around 160 universities now have a presence in Second Life\(^55\) – as have some scientists. There is even a continent called the SciLands where a number of groups with scientific interests have congregated (though from a distance, and to my eyes, its administration appears dauntingly bureaucratic). Nature Publishing Group also has its own small archipelago – inevitably called Second Nature and consisting of three separate islands – on which a diverse group of scientists is building and maintaining educational features in evolutionary biology, genetics, cell biology, chemistry and earth sciences, among others. There are also meeting, presentation and poster display areas. The degree of activity and enthusiasm has, quite frankly, astonished us. True, Second Life and other virtual worlds are still at an early stage in their evolution, are clunky to use and require large doses of patience and practice to get the most out of them. But the same was true of the web during the early 1990s and look what happened there. One major factor working against Second Life’s rapid expansion is the fact that it is a proprietary ‘walled garden’ controlled by a single commercial organisation, Linden Lab. In this sense, it is more like early AOL than the early web. But conversations with staff at Linden Lab suggest that they understand this potential pitfall, and they have already released, as open source, the code to their client application.\(^56\) If the server side code is opened up

\(^50\) Wilson, F. The Freemium Business Model. [Link to Wilson’s blog post]
\(^51\) EBay – [Link to EBay website]
\(^52\) Elance – [Link to Elance website]
\(^53\) InnoCentive – [Link to InnoCentive website]
\(^54\) Second Life – [Link to Second Life website]
\(^55\) Nature Publishing Group estimate.
\(^56\) Linden, P. Embracing the Inevitable. [Link to Linden’s blog post]
Tagging and folksonomies

One class of social software that deserves special comment is social bookmarking tools. One of the earliest was del.icio.us, and its introduction of tagging – freeform keywords entered by users to facilitate later retrieval – soon gave rise to the concept of the ‘folksonomy,’ a kind of implicit collective taxonomy or ontology generated by the aggregate, uncoordinated activity of many people tagging the same resources. Some commentators, notably Clay Shirky and David Weinberger, have argued (convincingly in my opinion) that this approach, although anarchic, has certain advantages over traditional centralized taxonomic approaches (such as the Dewey Decimal System). In particular, traditional approaches have difficulty dealing with entities that belong in multiple categories (is Nature a magazine or a journal?), or about which our view changes over time (Watson & Crick’s 1953 paper reporting the structure of DNA is in the field of biotechnology, but that word did not exist at the time). Since such challenges are often particularly acute in science, which necessarily operates at the frontiers of human knowledge, it is tempting to wonder whether collaborative tagging can help in that domain too.

Nature Publishing Group has its own social bookmarking and reference management tool, Connotea, heavily inspired by del.icio.us but with certain features added with academic researchers in mind. As well as providing a way for researchers to store, organise and share their reading lists, we were also interested to find out how useful the resultant collective tag metadata could be in helping to automatically link together related online scientific resources. To that end, we developed code for the EPrints institutional repository software that enabled it to query Connotea for tag information and automatically derived related reading suggestions. The experiment proved a success and we have built tagging into many of the applications we have developed since then (e.g., Nature Network, Nature Precedings and Scintilla) with a view to implementing similar features when the data sets grow large enough.

Open data and mashups

Another area with huge potential – but one that I have space to deal with only cursorily here – is that of open scientific data sets and forms of interoperability that allow these to be transferred not only between scientists but also between applications in order to create new visualizations and other useful transformations. There are numerous challenges, but there is also progress to report on each front. Too often scientists are unwilling to share data, whether for competitive or other reasons, though increasingly funders (and some publishers) are requiring them to do so. Even when the data are available, they usually lack the consistent formats and unambiguous metadata that would enable them to be efficiently imported into a new application and correctly interpreted by a researcher who was not present when they were collected. Yet data standards such as CML and SBML are emerging, as are metadata standards such as MIAME. As software applications also adopt these standards, we enter a virtuous circle in which there are increasing returns (at least at the global level) to openly sharing data using common standards.

For a glimpse of the benefits this can bring, witness the work of my colleague, Declan Butler, a journalist at Nature. While covering the subject of avian flu, it came to his attention that information about global outbreaks was fragmented, incompatible,
and often confidential. So he took it upon himself to gather what data he could, merge it together and provide it in the form of a KML file, the data format used by Google Earth.\footnote{KML – \url{http://code.google.com/apis/kml/documentation/}} Shortly afterwards he overlaid poultry density data.\footnote{Butler, D. The spread of avian flu with time; new maps exploiting Google Earth’s time series function. \url{http://declanbutler.info/blog/?p=58}, 2007.} This not only meant the information was now available in one place, it also made it much more readily comprehensible to experts and non-experts alike. Imagine the benefits if this approach, largely the work of one man, was replicated across all of science.

**Wither the scientific web?**

Over the last 10 years or so, much of the discussion about the impact of the web on science – particularly among publishers – has been about the way in which it will change scientific journals. Sure enough, these have migrated online with huge commensurate improvements in accessibility and utility. For all but a very small number of widely read titles, the day of the print journal seems to be almost over. Yet to see this development as the major impact of the web on science would be extremely narrow-minded – equivalent to viewing the web primarily as an efficient PDF distribution network. Though it will take longer to have its full effect, the web’s major impact will be on the way that science itself is practiced.

The barriers to full-scale adoption are not only (or even mainly) technical, but rather social and psychological. This makes the timings almost impossible to predict, but the long-term trends are already unmistakable: greater specialization in research, more immediate and open information-sharing, a reduction in the size of the ‘minimum publishable unit,’ productivity measures that look beyond journal publication records, a blurring of the boundaries between journals and databases, reinventions of the roles of publishers and editors, greater use of audio and video, more virtual meetings. And most important of all, arising from this gradual but inevitable embrace of technology, an increase in rate at which new discoveries are made and exploited for our benefit and that of the world we inhabit.
Reinventing Scholarly Communication for the Electronic Age

Introduction

Cyberinfrastructure is integral to all aspects of conducting experimental research and distributing those results. However, it has yet to make a similar impact on the way we communicate that information. Peer-reviewed publications have long been the currency of scientific research as they are the fundamental unit through which scientists communicate with and evaluate each other. However, in striking contrast to the data, publications have yet to benefit from the opportunities offered by cyberinfrastructure. While the means of distributing publications have vastly improved, publishers have done little else to capitalize on the electronic medium. In particular, semantic information describing the content of these publications is sorely lacking, as is the integration of this information with data in public repositories. This is confounding considering that many basic tools for marking-up and integrating publication content in this manner already exist, such as a centralized literature database, relevant ontologies, and machine-readable document standards.

We believe that the research community is ripe for a revolution in scientific communication and that the current generation of scientists will be the one to push it forward. These scientists, generally graduate students and new post-docs, have grown up with cyberinfrastructure as a part of their daily lives, not just a specialized aspect of their profession. They have a natural ability to do science in an electronic environment without the need for printed publications or static documents and, in fact, can feel quite limited by the traditional format of a publication. Perhaps most importantly, they appreciate that the sheer amount of data and the number of publications is prohibitive to the traditional methods of keeping current with the literature.

To do our part to get the revolution turning, we are working with the Public Library of Science\(^1\) and a major biological database, the RCSB Protein Data Bank\(^2\) to destroy the traditional concept of a publication and a separate data repository and reinvent it as an integration of the two information sources. Here, we describe new authoring tools that are being developed to consummate the integration of literature and database content, tools being developed to facilitate the consumption of this integrated information, and the anticipated impact of these tools on the research community.

The Legacy of Scientific Publishing

Publications are the currency of research. They are the mechanism through which scientists communicate their results to their peers and the means through which we evaluate each other. This model is unlikely to change completely. However, the electronic age – the introduction of cyberinfrastructure – is introducing some differences in this paradigm, a phenomenon that has been observed previously.\(^3\)\(^4\)\(^5\)

One significant difference is the requirement of many publishers that the author deposit the data described in a publication in an appropriate public repository concomitant with publication of the manuscript. For example, macromolecular structure data must be deposited in the Protein Data Bank (PDB) when a manuscript describing the macromolecule’s structure is published. As part of the deposition process, a ref-

---

1. Public Library of Science – http://www.plos.org/
2. RCSB Protein Data Bank – http://www.pdb.org/
A similar scenario exists with publications that reference a record in one of these repositories. For example, if someone has used the structural data describing a protein from the PDB, this structure will be referenced by ID in the publication but this reference most times is not included in the PDB itself. Someone viewing this structure in the database will see only the citation of the publication and abstract describing the generation of the initial data. Much of the research performed subsequent to structure deposition concerns functional information about the protein, information that would surely be useful to anyone interested in that molecule but information that is not trivial to obtain.

One possible reason why a link between this secondary publication and the database is not made is that there is no value associated with database annotation. Scientists are valued on their peer-reviewed publications, not on their database annotations (which are not independently peer-reviewed). Furthermore, proper database annotation takes time and effort so there is little to gain in this endeavor.

Another barrier that contributes to disconnection between publications and data deposition is the delay in utilization of cyberinfrastructure by scientific publishers. Most publishers have at least an online presence and generally make the articles they publish available online for download, viewing, or printing but do little else to make use of the information they are communicating. Cyberinfrastructure is little more than a new means of distribution.

A significant issue that complicates extensive use of publication content is intellectual property rights, an issue that is currently quite controversial. Some publishers have risen to the challenge and adopted the open access philosophy, publishing their articles under a Creative Commons AL license. This means that the content is free to use and distribute as long as the original attribution is maintained. The Public Library of Science (PloS) and BioMedCentral (BMC) are good examples of publishers in the life sciences who have embraced the open access model. The articles published by these publishers and others are archived in a central repository, PubMed Central, which archives all open access articles in the life sciences. To date, there are over 54,000 articles in over 300 journals – hardly a major representation of the field, but still a solid beginning.

The centralization of open access articles is a significant step forward but, even more significant, is the storage of these articles in a standardized and machine-readable format: the NLM DTD. This document format allows all open access articles to be archived as XML files, which includes some semantic mark-up of the content and unique identifiers for the article itself and the objects (figures and tables) within. This format also allows these articles to be parsed for relevant information. Unfortunately, there is little value added to article content itself. To recall a previous example, referencing a protein structure in a manuscript, most authors referencing a protein structure do not include a link to the structural data on the PDB. In order to find a mention of a PDB ID, one would have to perform a full text search on the article content (including figure captions). Even this does not ensure that a successful search result actually references a PDB ID – the ID could belong to a different database or have an entirely different meaning, since there is no semantic context for that string of text. (This is not always true. Some papers do include direct references to the PDB using the xlink tag.)
Even if a link is included in the PDB to an article that mentions a PDB ID, it is not clear what the value of that reference is to the reader. Does the article describe the biochemical function of the protein or was the structure used in training a computational prediction algorithm? Rather than direct the reader to an article that may not be of interest, it would be useful to include some indication of the type of content of the article. Semantic mark-up of the article content is necessary. Using ontologies or controlled vocabularies within the framework of the NLM DTD would increase the usefulness of the article content dramatically.

All of these tools exist – the standardized document format, the ability to create hyperlinks in electronic documents, field-specific ontologies – but they have yet to be utilized to their full advantage. This may be due to the legacy of static manuscripts, which is largely perpetuated by scientists who did not have access to cyberinfrastructure during their formative years. However, today’s scientists do and it is time to make this happen.

**BioLit: Blurring the Boundary Between Publications and Data**

In order to initiate a community effort, we are developing a set of open source tools that will facilitate the integration of open literature and biological data, a project we call BioLit. Initially, these tools will be implemented using the entire corpus of the Public Library of Science (PLoS) and the Protein Data Bank (PDB) as testing platforms. The tools are being designed, however, to be generally applicable to all open access literature and other biological data.

The Public Library of Science (PLoS) is an ideal partner since it is leading the open access movement - a fundamental change in scientific publication, which represents a significant improvement in access to literature by the scientific community. Articles are published under a Creative Commons Attribution License\(^6\) whereby the community may use the article in any way they see fit, provided they attribute the original authors. Furthermore, the copyright to the material remains with the author. Once published, the article is available free in its entirety to anyone. This means that all PLoS articles, which are very high quality, are freely available, freely usable, and consist of a large body of text covering biology, medicine, computational biology, genetics, pathology and a variety of other fields.

The Protein Data Bank (PDB) is one of the oldest databases in biology and contains all publicly accessible three dimensional structures of biological macromolecules – currently over 44,000.\(^7\) The PDB is used by over 10,000 scientists every day and one structure is downloaded, on average, every second. Over the last decade or so, PDB structures appear in roughly 2% of all open access life science journal articles making the PDB an obvious target for an effort to integrate data with open literature.

Specifically, the BioLit tools will capture meta-data from an article or manuscript by identifying relevant terms and identifiers and adding mark-up to the original NLM DTD-based XML document containing the article. Terms relating to the life sciences are identified using ontologies and controlled vocabularies specific to this field such as the Gene Ontology\(^8\)\(^9\) and Medical Subject Headings (MeSH).\(^10\)

This meta-data is captured in different ways depending on the status of the article. A tool we are developing with Microsoft, which will be implemented as a plug-in for Word, will allow this information to be captured while the manuscript is being written. This strategy gives the author full and fine control over the exact meta-data that are

---

6 Creative Commons Attribution License - [http://www.plos.org/journals/license.html](http://www.plos.org/journals/license.html)


10 Medical Subject Headings - [http://www.nlm.nih.gov/mesh](http://www.nlm.nih.gov/mesh)
captured. The plug-in will prompt the author with choices or will allow the author to customize the meta-data if no appropriate matches are found in the resources that the plug-in has knowledge of. Cross-references to biological databases will also be detected and added to the meta-data, allowing the manuscript content to be more easily integrated with the database.

Articles that have already been published can be post-processed through a related tool that identifies the same types of meta-data and generates similar XML mark-up. The meta-data may not be as rich using this approach since the author has not had direct input, but the capture of any information is a significant advance. Processing all PLoS articles, and later all open access articles, with this tool will generate a considerable amount of meta-data, which will help establish the integration effort in the community.

The BioLit tools will effectuate a change in the authoring process that is nearly transparent to the author, but will capture significant new meta-data and establish an informative connection between the data and the article describing the data. Effective use of these tools will provide new views on the traditional literature and on biological databases. The literature will simply become another interface to biological data in a database, and the database can recall appropriate literature – not in abstract or complete paper size chunks, but knowledge objects that annotate the data being examined (see Figure 1 for an example).

Figure 1. A PDB entry enhanced with additional literature. Historically, PDB entries reference the original article in which the macromolecular structure was published (Primary Citation). The BioLit tools will add text and figures from other articles that reference this structure in order to provide data that describe other aspects of the protein (Additional Literature).
SciVee: Pioneering New Methods of Scientific Communication

In addition to the BioLit tools for authoring and data integration, we want to use cyberinfrastructure to the fullest advantage. Due to the increasing availability of high bandwidth and consumer-level video recording equipment, internet video is now wildly popular. We want to take advantage of this trend and use this medium to communicate science more effectively. It is important, however, to bear in mind the need for quality content. To this end, we have developed SciVee, which allows authors to upload an article they have already published (open access, naturally) with a video or podcast presentation (about 10 minutes long) that they have made that describes the highlights of the paper. The author can then synchronize the video with the content of the article (text, figures, etc.) such that the relevant parts of the article appear as the author discusses them during the video presentation. We call the result a pubcast. Figure 2 shows a typical view of a SciVee pubcast.

Figure 2. A SciVee pubcast. This figure shows how a video presentation is integrated with a published article. While the speaker is discussing a point from the article, the relevant figure or text is highlighted. The viewer can also download the original paper as a companion to the pubcast.

Anyone can visit SciVee and view the pubcast. It is similar to attending a conference to hear a particular speaker, except that the pubcast is available on demand, can be viewed any number of times, and explicitly refers to the content of the original article. Another important feature of SciVee is the ability of any user to add or read comments on pubcasts. This allows a community to be established around an article and encourages discussion about the results and their impact on the field. We believe this activity will transform what has traditionally been a static document into a dynamic exchange.

SciVee makes it easier and faster to keep up with current literature by delivering the key points of articles in a portable and enjoyable medium. A reader can interact with several articles using this website in the time it would take to read a single full article in the traditional way.
Conclusion

We believe revitalizing journal articles will have a significant impact on the scientific community. The traditional article format no longer effectively supports the research in the electronic age. The number of articles researchers need to read in order to keep up with their field has increased significantly in the last decades. In addition, there is an increasing number of articles that report data generated in a high-throughput manner, and the primary method of exploring these data is through a database, not through the article itself. In part, these phenomena are due to the increasing reliance on cyberinfrastructure to perform research. It is thus a natural response to use cyberinfrastructure to address this situation. Indeed, our initial tests have proven quite successful. A large group of students in the UCSD School of Pharmacy and Pharmaceutical Sciences were shown an eight minute pubcast of a recently published paper and were then quizzed on their comprehension. Their results were compared to students who were given the paper and eight minutes in which to read it. The pubcast group largely outperformed the paper group and, perhaps more importantly, greatly enjoyed the experience.

SciVee and the BioLit tools will complement similar efforts such as the Structured Digital Abstract,12 MutationFinder,13 and BioCreAtIvE.14 We hope that the scientific community will embrace these efforts and use cyberinfrastructure to its fullest capacity to make scientific communication more enjoyable and effective.

Acknowledgements

This work is supported by grants 0544575 and 0732706 from the National Science Foundation.

Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication

1. Introduction

Improvements in computing and network technologies, digital data capture, and data mining techniques are enabling research methods that are highly collaborative, network-based, and data-intensive. These methods challenge existing scholarly communication mechanisms, which are largely based on physical (paper, ink, and voice) rather than digital technologies.

One major challenge to the existing system is the change in the nature of the unit of scholarly communication. In the established scholarly communication system, the dominant communication units are journals and their contained articles. This established system generally fails to deal with other types of research results in the sciences and humanities, including datasets, simulations, software, dynamic knowledge representations, annotations, and aggregates thereof, all of which should be considered units of scholarly communication.¹

Another challenge is the increasing importance of machine agents (e.g., web crawlers, data mining applications) as consumers of scholarly materials. The established system by and large targets human consumers. However, all communication units (including the journal publications) should be available as source materials for machine-based applications that mine, interpret, and visualize these materials to generate new units of communication and new knowledge.

Yet another challenge to the existing system lies in the changing nature of the social activity that is scholarly communication. Increasingly, this social activity extends beyond traditional journals and conference proceedings, and even beyond more recent phenomena such as preprint systems, institutional repositories, and dataset repositories. It now includes less formal and more dynamic communication such as blogging. Scholarly communication is suddenly all over the web, both in traditional publication portals and in new social networking venues, and is interlinked with the broader social network of the web. Dealing adequately with this communication revolution requires fundamental changes in the scholarly communication system.

Many of the required changes in response to these challenges are of a socio-cultural nature and relate directly to the question of what constitutes the scholarly record in this new environment. This raises the fundamental issue of how the crucial functions of scholarly communication—registration, certification, awareness, archiving, rewarding—should be re-implemented in the new context. The solutions to these socio-cultural questions rely in part on the development of basic technical infrastructure to support an innately digital scholarly communication system.

This paper describes the work of the Object Re-Use and Exchange (ORE) project of the Open Archives Initiative (OAI) to develop one component of this new infrastructure in order to support the revolutionized scholarly communication paradigm—standards to facilitate discovery, use and re-use of new types of compound scholarly communication units by networked services and applications. Compound units are aggregations of distinct information units that, when combined, form a logical whole. Some examples of these are a digitized book that is an aggregation of chapters, where each chapter is an aggregation of scanned pages, and a scholarly publication that is an


aggregation of text and supporting materials such as datasets, software tools, and video recordings of an experiment. The ORE work aims to develop mechanisms for representing and referencing compound information units in a machine-readable manner that is independent of both the actual content of the information unit and nature of the re-using application.

2. Compound Information Objects

The new units of communication that are emerging from the modern research environment have a compound nature that does not have a direct parallel in traditional, paper-based publications or in the digital versions thereof (e.g., pdf, LaTex). They are aggregates of multiple distinct components that can vary according to semantic type (article, simulation, video, dataset, software, etc.), media type (text, image, audio, video, mixed), media format (PDF, XML, MP3, etc.), and network location (different components made accessible by different repositories). In addition, each aggregate carries an identifier associated with it by the information system that composed the aggregation, thereby establishing it as a logical unit of scholarly communication. In the remainder of this paper, we will refer to these aggregates as either compound information objects or compound objects (Figure 1).

![Figure 1. A compound information object composed by an information system.](image)

These compound objects are a fundamental building block of eScience and eScholarship, and support for them is an essential aspect of cyberinfrastructure. For example, the ImageWeb activity led by David Shotton’s BioInformatics Research Group at the University of Oxford explores the creation of so-called image webs that integrate cellular images held by publishers, research organizations, museums, and institutional repositories. Also, Gregory Crane, a leading scholar in the humanities, envisions the notion of recombinant documents. These documents have a number of features that differentiate compound documents from physical documents or their digital incunabula. They aggregate new information and existing fine-grained digital information. The aggregation can be human-author based, for example, as the result of a workflow within a so-called scholarly workbench, or machine-generated based, for example, on machine learning techniques and web crawling. The aggregation of an existing information unit into a compound object (re-use) is not due to the inherent nature of the aggregated unit, but is the result of the algorithmic design or the intention of the human that composed the compound object. Finally, these objects may be dynamic and grow over time based on usage patterns as well as social activity that provide additional context for the information within them.
3. Publishing Compound Objects to the Web

The layer cake metaphor is commonly used to describe information infrastructure, where new layers of functionality build upon existing layers. Tim Berners-Lee has, for example, used this model to describe the semantic web that consists of functionality built on the base web architecture. The work of ORE follows the same paradigm. It presumes the web architecture as the de facto foundation for interoperability and positions the ORE standards as a layer over this web foundation. Thereby the ORE work leverages the facilities provided by the web architecture, adding functionality related to compound objects that is not present in the web foundation layer.

This web foundation layer (Figure 2) defines architectural notions that allow information systems that compose compound objects to publish them to the web by associating a URI with each of the components of a compound object, thereby making the components URI-identified resources. Web services and applications, such as browsers and crawlers, can use these URIs to obtain representations of the resources via content negotiation.

When the components of a compound object are published as resources on the web, they may link to each other (e.g., S links to 2 in Figure 3), they may link to other
resources (e.g., 1 links to 8 in Figure 3), and other resources may link to them (e.g., 9 links to 2 in Figure 3). These links are the basis of the rich information environment that is the web. But, because they are generally un-typed (they are standard hyperlinks), or their types do not conform to any general standard, the links do not define the boundary relationship that exists among the resources that are components of a compound object (Figure 3). The logical whole that is the compound object disintegrates into a set of distinct resources that are indistinguishable from the other resource in the web graph.

Many information systems address this problem by expressing the compound object via a user-oriented html "splash" or "jump-off" page that lists links to all components of the compound object and to a variety of related resources. This is illustrated by Figure 4, where a splash page in the arXiv provides access to the various formats in which a document is available and also to external resources, such as citations. This splash page resource is also shown in Figure 3 as resource S.

![Figure 4. Splash page for an arXiv document](http://arxiv.org/abs/hep-th/0507171).

These splash pages have come to de facto represent the compound object "as a whole" on the web, and, as a result, a convention has emerged to use the URI of the splash page as the URI of the compound object itself. While this approach is useful for human users, it is problematic from a machine re-use perspective because:

- Machine interpretation of splash page information is difficult or impossible due to the lack of standards for its structure; and
- According to the Web Architecture the URI of the splash page merely identifies the resource that is the splash page, which is actually only a component of the compound object and not the compound object itself.

In addition to these problems related to identifying the compound object and defining its structure, the web architecture has no method to explicitly reference a resource within the context of a compound object. This functionality is important for scholarly communication because an existing resource can be re-used as a component of any number of compound objects. For example, a specific cellular image may be part of one image web illustrating confocal microscopy techniques and of another concerned with cancer therapy. For provenance tracking and citation, it is important to have the ability to reference a resource as it exists as a component of one or other specific compound object (e.g., the cellular image as a part of the cancer therapy image...
Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication

...because the exact meaning of the resource or of a reference to it can be dependent on the context provided by such object.

The goal of the ORE work is to address this shortcoming and define an infrastructure layer over the web architecture allowing interoperable use and re-use of and reference to compound information objects across a variety of networked applications. This ORE layer explicitly does not replace or redefine any core web architecture concepts. Indeed, it leverages them fully as part of the solution to the problem of expressing compound objects.

The ORE layer expresses the boundaries of a compound object in a manner that can be processed by machines and agents. This can be viewed as the specification of a machine-readable splash page that lists the components of a compound object, as well as their internal and, optionally, external relationships. Such a specification would support:

- Re-use of a compound object and its components across web-savvy applications.
- Reference to a compound object and its components in a manner that supports an understanding of their “compound object context.” This would allow reference to a resource as it exists in the context of a specific compound object, distinguishing that from a reference to the same resource as it exists in the context of another compound object, or as just a resource in its own right.
- Machine discovery of compound object information.

4. Towards a Solution for Compound Information Objects

The work to date on the formulation of an ORE interoperability layer over the web architecture consists of three components:

- Using Resource Maps for describing compound objects;
- Referencing compound object resources; and
- Discovering Resource Maps.

Similar to other efforts in the scholarly community to develop interoperable, machine-readable representations of research artifacts, this ORE work is inspired by activities in the semantic web community. Two notable influences are Linked Data and named graphs. We note however that the semantic web influence on ORE work does not mandate an implementation built on semantic web technologies such as RDF, RDFS, RDF/XML, and OWL. Rather, OAI-ORE might employ more lightweight technologies that are easier to adopt and that can be used in applications that typically do not require explicit semantic information (e.g., existing popular search engines, blogs). These simpler formats could then be transformed by mechanisms such as GRDDL to extract data for more complex semantic applications.

4.1 Using Resource Maps to Describe Compound Objects

As explained above, a major issue with representing compound objects on the web is the loss of the notion of the logical whole because the web architecture has no standardized facility for representing aggregations of resources. As such, a major focus of OAI-ORE is to add this missing logical boundary information and to do so in a machine-readable manner.

We are examining the use of named graphs as a model for expressing this boundary information. A named graph is a set of nodes and arcs. In this context, a node is a URI-identified web resource, and an arc is a directional link between two such...
resources typed by a URI that indicates a relationship type. The named graph itself is identified by means of a URI. This URI-identification means that the named graph itself is a logical unit and is an addressable resource on the web. As such, named graphs provide a mechanism for describing the basic relationship between a compound object and its components that is missing when a compound object is published to the web. In addition, named graphs provide a means of associating a proxy URI identity (the URI of the named graph) with the compound object. Finally, in addition to expressing this basic boundary-type information, these graphs can express semantically richer information because they may contain arbitrarily typed resources (nodes) and relationships (arcs) that, for example, meet the requirements of a specific application domain.

The ORE interoperability layer intends to leverage named graphs by publishing Resource Maps that describe compound objects. A Resource Map is a named graph in which the nodes are resources that correspond with a compound object and its components, as well as resources that are related to these (e.g., the citations of a scholarly paper). A Resource Map must unambiguously distinguish between those “internal” and “external” resources. The arcs of a Resource Map are typed relationships between those resources. We envision a core relationship ontology and the ability to extend this core with discipline-specific ontologies. For practical reasons, a Resource Map is identified by means of a protocol-based URI (e.g., HTTP URI). This makes it possible to obtain machine-readable representations of the Resource Map through content negotiation. As a result, a Resource Map is considered an information resource as per \[15\].

Published Resource Maps overlay the web graph and effectively become part of (are merged into) it. This is illustrated in Figure 5, where the top pane shows the web graph without the information contained in the Resource Map and the bottom pane illustrates how the boundary of the object and relationships among the components are now visible in the web graph. The URI of the Resource Map (R in Figure 5) provides a web-based handle to the aggregate of multiple resources and their inter-relationships in the Resource Map. This URI can be referenced by standard web applications.

4.2 Referencing Compound Object Resources

Re-use of a web resource depends on the ability to reference it. As explained earlier, referencing issues exist when publishing compound objects to the web. First, there exists no web-parallel to the identifier of the compound object, shown in Figure 1. Second, there is a need to reference a specific resource not just in its own right (i.e.,
Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication

by means of its URI), but in the manner that it appears in the context of a certain compound object.

Referencing the Compound Object as a Whole

The protocol-based URI of the Resource Map identifies an aggregation of resources (components of a compound object) and their boundary-type inter-relationships. While this URI is clearly not the identifier of the compound object itself, it does provide an access point to the Resource Map and its representations that list all the resources of the compound object. For many practical purposes, this protocol-based URI may be a handy mechanism to reference the compound object because of the tight dependency of the visibility of the compound object in web space on the Resource Map (i.e., in ORE terms, a compound object exists in web space if and only if there exists a Resource Map describing it).

We note, however, two subtle points regarding the use of the URI of the Resource Map to reference the compound object. First, doing so is inconsistent with the web architecture and URI guidelines that are explicit in their suggestion that a URI should identify a single resource. Strictly interpreted, then, the use of the URI of the Resource Map to identify both the Resource Map and the compound object that it describes is incorrect. Second, some existing information systems already use dedicated URIs for the identification of compound information objects “as a whole.” For example, many scholarly publishers use DOIs, whereas the Fedora and aDORe repositories have adopted identifiers of the info URI scheme. These identifiers are explicitly distinct from the URI of the Resource Map.

These issues suggest that it should be possible to express in the ORE specifications, and therefore in the Resource Map and its representations, an additional URI – the identifier of the compound object itself. Once a URI to identify the compound object “as a whole” is introduced, it would play the prominent role in the Resource Map of identifying the resource that corresponds with the compound object and that has component parts (resource C in Figure 6).
Referencing Resources in Context

On the web, resources can unambiguously be referenced by means of their URI. As a result, each published component of a compound object, as well as a published Resource Map describing a compound object, can be referenced. As mentioned in the previous section, a compound object “as a whole” can be referenced if it is assigned a dedicated URI. However, as explained earlier by means of the cellular image example, there is often the need in scholarly communication to reference a resource as a component of a specific compound object.

Figure 7 illustrates a scenario in which resource U is used as part of two compound objects. To reveal boundary information regarding the upper compound object, Resource Map X is published; Resource Map Y is published to do the same for the lower compound object. Because U is part of both compound objects, both Resource Maps X and Y reference resource U. In order to accommodate the need to reference U as part of a specific compound object, OAI-ORE proposes to use an identifier pair that consists of the identifier of the resource itself and the identifier of the Resource Map that corresponds with the desired compound object, i.e. (U,X) to reference resource U as part of the upper compound object and (U,Y) to reference it as part of the lower compound object.

4.3 Discovering Compound Objects on the Web

Exposing compound objects on the web via Resource Maps is only part of the solution; the Resource Maps and its referenced resources need to be discovered to really become part of the web graph. OAI-ORE proposes two complementing approaches with this regard:

- Harvest type discovery, which consists of making batches of Resource Maps available through existing mechanisms such as RSS, Sitemaps, and OAI-PMH.
- Linked Data\textsuperscript{11} type discovery, which uses HTTP headers received in response to dereferencing the URI of a component of a compound object to point at the Resource Map(s) that correspond(s) with the compound object. This is shown in Figure 8 where a crawler lands upon splash page S and is pointed at the Resource Map R by means of an HTTP LINK header contained in the response to an HTTP GET request issued against S. From R, the crawler can obtain a Resource Map representation, and hence a list of all resources that are part of the compound object, as well as further internal and external relationships.

5. Conclusion

Compound information objects are becoming the norm rather than the exception in the new scholarly communication environment. As a result, it is essential to augment the existing technical communication infrastructure with an interoperable approach that allows using, re-using, referencing, and discovering them across the borders of scholarly disciplines and applications. The international OAI-ORE effort works towards a solution that fully leverages the web architecture and that consists of publishing Resource Maps that describe compound objects, referencing resources in their compound object context, and mechanisms to facilitate discovery of Resource Maps.

Although OAI-ORE has made significant conceptual progress since it started in September 2006, important questions remain unanswered. How will the solution deal with versioning? How can the trustworthiness of Resource Maps be assessed? Which kinds of relationship types should OAI-ORE define to support bootstrapping adoption, and which should be left to individual communities? Which technologies should be used to represent Resource Maps, and how does a choice affect potential adoption? Some of these questions will receive at least a preliminary answer by the end of September 2007, which is the deadline that OAI-ORE has set itself for the release of a public alpha specification. Following that release, OAI-ORE will encourage experimentation by various scholarly communities and solicit feedback from potential stakeholders worldwide. The insights gained from those activities will be taken into account for a version 1 specification that is planned for September 2008.

Acknowledgments
OAI-ORE is supported by the Andrew W. Mellon Foundation, the Coalition for Networked Information, Microsoft, and the National Science Foundation (IIS-0430906).

The authors acknowledge the contributions to the OAI-ORE effort from the ORE Technical Committee, Liaison Group and Advisory Committee. The authors also acknowledge the contributions of John Erickson (HP Labs) and Sandy Payette (Cornell Information Science).
Appendix

In the course of May 2007, the Digital Library Research & Prototyping Team of the Los Alamos Laboratory launched an experiment to explore the notion of Resource Map publishing as a means to expose compound object boundary-type information to the web. More particularly, the experiment explored whether an existing web application would be able to take advantage of published Resource Maps, without requiring any modifications to the application itself. The experiment pertained to archiving compound information objects as they evolve over time and the applications that were used were the Internet Archive's Heritrix toolkit that contains a web crawler and its Wayback Machine user interface.

The experiment’s optimistic scenario assumes that Resource Map publishing has become so commonplace that the Internet Archive starts to actively collect them. The experiment zooms in on two publishers that make Resource Maps discoverable via dedicated Sitemaps. When a Resource Map listed in a SiteMap changes, its associated SiteMap date-time is changed. When a new Resource Map is published, it is added to the SiteMap. The Internet Archive uses these Sitemaps and their contained date-times as a trigger to collect and archive Resource Maps as well as the resources they reference. As a result, the Wayback Machine now allows searching for a specific Resource Map of a specific date and for immediately seeing the version of the resources referenced by that Resource Map as they existed on that same date. Understanding that Resource Maps expose the boundaries of compound objects, the net result is in effect an archive of evolving compound objects, versioned by the datetime of the Resource Map that describes them.

A screencast that shows a walk-through of the various components involved in the experiment and follows the evolution of some Resource Maps over time can be found at http://www.ctwatch.org/quarterly/multimedia/11/ORE_prototype-demo/.
Introduction

The research production cycle has three components: the conduct of the research itself (R), the data (D), and the peer-reviewed publication (P) of the findings. Open Access (OA) means free online access to the publications (P-OA), but OA can also be extended to the data (D-OA): the two hurdles for D-OA are that not all researchers want to make their data OA and that the online infrastructure for D-OA still needs additional functionality. In contrast, all researchers, without exception, do want to make their publications P-OA, and the online infrastructure for publication-archiving (a worldwide interoperable network of OAI\(^1\)-compliant Institutional Repositories [IRs]\(^2\)) already has all the requisite functionality for this.

Yet because so far only about 15% of researchers are spontaneously self-archiving their publications today, their funders and institutions are beginning to require OA self-archiving,\(^3\) so as to maximize the usage and impact of their research output.

The adoption of these P-OA self-archiving mandates needs to be accelerated. Researchers’ careers and funding already depend on the impact (usage and citation) of their research. It has now been repeatedly demonstrated that making publications OA by self-archiving them in an OA IR dramatically enhances their research impact.\(^4\) Research metrics (e.g., download and citation counts) are increasingly being used to estimate and reward research impact, notably in the UK Research Assessment Exercise (RAE).\(^5\) But those metrics first need to be tested against human panel-based rankings in order to validate their predictive power.

Publications, their metadata, and their metrics are the database for the new science of scientometrics. The UK’s RAE, based on the research output of all disciplines from an entire nation, provides a unique opportunity for validating research metrics. In validating RAE metrics (through multiple regression analysis)\(^6\) against panel rankings, the publication archive will be used as a data archive. Hence the RAE provides an important test case both for publication metrics and for data-archiving. It will not only provide incentives for the P-OA self-archiving of publications, but it will also help to increase both the functionality and the motivation for D-OA data-archiving.

Now let us look at all of this in a little more detail:

Reasearch, Data, and Publications

Research consists of three components: (1) the conduct of the Research (R) itself (whether the gathering of empirical data, or data-analyses, or both), (2) the empirical Data (D) (including the output of the data-analyses), and (3) the peer-reviewed journal article (or conference paper) Publications (P) that report the findings. The online era has made it possible to conduct more and more research online (R), to provide online access (local or distributed) to the data (D), and to provide online access to the peer-reviewed articles that report the findings (P).

The technical demands of providing the online infrastructure for all of this are the greatest for R and D – online collaborations and online data-archiving. But apart from
the problem of meeting the technical demands for R and for D-archiving, the rest is a matter of choice: if the functional infrastructure is available for researchers to collaborate online and to provide online access to their data, then the rest is just a matter of whether and when researchers decide to use it to do so. Some research may not be amenable to online collaboration, or some researchers may for various reasons prefer not to collaborate, or not to make their data publicly accessible.

In contrast, when it comes to P, the peer-reviewed research publications, the technical demands of providing the online infrastructure are much less complicated and have already been met. Moreover, all researchers (except those working on trade or military secrets) want to share their findings with all potential users, by (i) publishing them in peer reviewed journals in the first place and by (ii) sending reprints of their articles to any would-be user who does not have subscription access to the journal in which it was published. Most recently, in the online age, some researchers have also begun (iii) making their articles freely accessible online to all potential users webwide.

Open Access

Making articles freely accessible online is also called Open Access (OA). OA is optimal for research and hence inevitable. Yet even with all of P’s less exacting infrastructural demands already met, P-OA has been very slow in coming. Only about 15% of yearly research article output is being made OA spontaneously today. This article discusses what can be done to accelerate P-OA, to the joint advantage of R, D & P, using a very special hybrid example, based on the research corpus itself (P), serving as the database (D) for a new empirical discipline (R).

For “scientometrics” – the measurement of the growth and trajectory of knowledge - both the metadata and the full texts of research articles are data, as are their download and citation metrics. Scientometrics collects and analyzes these data by harvesting the texts, metadata, and metrics. P-OA, by providing the database for scientometrics, will allow scientometrics to better detect, assess, credit and reward research progress. This will not only encourage more researchers to make their own research publications P-OA (as well as encouraging their institutions and funders to mandate that they make them P-OA), but it will also encourage more researchers to make their data D-OA too, as well as to increase their online research collaborations (R). And although the generic infrastructure for making publications P-OA is already functionally ready, the specific infrastructure for treating P as D will be further shaped and stimulated by the requirements of scientometrics as R.

First, some potentially confusing details need to be made explicit and then set aside: publications (P) themselves sometimes contain research data (D). A prominent case is chemistry, where a research article may contain the raw data for a chemical structure. Some chemists have accordingly been advocating OA for chemical publications not just as Publications (P) but as primary research Data (D), which need to be made accessible, interoperable, harvestable and data-mineable for the sake of basic chemical research (R), rather than just for the usual reading of research articles by individual users. The digital processing of publication-embedded data is an important and valid objective, but it is a special case and hence will not be treated here, because the vast majority of research Publications (P) today do not include their raw data. It is best to consider the problem of online access to data that are embedded in publications as a special case of online access to D rather than as P. Similarly, the Human Genome Database, inasmuch as it is a database rather than a peer-reviewed publication, is best considered as a special case of D rather than P.
Incentivizing the Open Access Research Web:
Publication-Archiving, Data-Archiving and Scientometrics

Here, however, in the special case of scienotmetrics, we will be considering the case of P as itself a form of D, rather than merely as containing embedded D within it. We will also be setting aside the distinction between publication metadata (author, title, date, journal, affiliation, abstract, references) and the publication's full-text itself. Scientometrics considers both of these as data (D). Processing the full-text's content is the "semiometric" component of scientometrics. But each citing publication's reference metadata are also logically linked to the publications they cite, so as the P corpus becomes increasingly OA, these logical links will become online hyperlinks. This will allow citation metrics to become part of the P-OA database too, along with download metrics. (The latter are very much like weblinks or citations; they take the form of a "hit-and-run." Like citations however, they consist of a downloading site – identified by IP, although this could be made much more specific where the downloader agrees to supply more identifying metadata - plus a downloaded site and document.) We might call citation and download metrics "hypermetrics," alongside the semiometrics, with which, together, they constitute scientometrics.

Scientometrics

The objective of scientometrics is to extract quantitative data from P that will help the research publication output to be harvested, data-mined, quantified, searched, navigated, monitored, analyzed, interpreted, predicted, evaluated, credited and rewarded. To do all this, the database itself first has to exist and, preferably, it should be OA. Currently, the only way to do (digital) scientometrics is by purchasing licensed access to each publisher's full-text database (for the semiometric component), along with licensed access to the Thompson ISI Web of Science\(^\text{10}\) database for some of the hypermetrics. (Not the hypermetrics for all publications, because ISI only indexes about one third of the approximately 25,000 peer-reviewed research journals published across all fields, nations and languages.) Google Scholar\(^\text{11}\) and Google Books\(^\text{12}\) index still more, but are as yet very far from complete in their coverage – again because only about 15% of current annual research output is being made P-OA. But if this P-OA content can be raised to 100%, not only will doing scientometrics no longer depend on licensed access to its target data, but researchers themselves, in all disciplines, will no longer depend on licensed access in order to be able to use the research findings on which they must build their own research.

Three things are needed to increase the target database from 15% to 100%: (1) functionality, (2) incentives, and (3) mandates.\(^\text{3}\) The network infrastructure needs to provide the functionality, the metrics will provide the incentive, and the functionality and incentives together will induce researchers' institutions and funders to mandate OA for their research output (just as they already mandate P itself: "publish or perish").

Citebase

As noted, much of the functional infrastructure for providing OA has already been developed. In 2000,\(^\text{13}\) using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (OAI-PMH),\(^\text{1}\) the Eprints\(^\text{14}\) group at the University of Southampton designed the first and now widely used free software (GNU Eprints) for creating OAI-interoperable Institutional Repositories (IRs). Researchers can self-archive the metadata and the full-texts of their peer-reviewed, published articles by depositing them in these IRs. (If they wish, they may also deposit their pre-peer-review preprints, any post-publication revisions, their accompanying research data [D-OA], and the metadata, summaries and reference lists of their books). Not only can Google and Google Scholar harvest the contents of these IRs, but so can OAI services such as OAIster,\(^\text{15}\) a virtual

---

\(^\text{10}\) Web of Science - [http://scientific.thomson.com/products/wos/](http://scientific.thomson.com/products/wos/)

\(^\text{11}\) Google Scholar - [http://scholar.google.com/](http://scholar.google.com/)


\(^\text{13}\) [http://www.dlib.org/dlib/october00/10inbrief.html#HARNAD](http://www.dlib.org/dlib/october00/10inbrief.html#HARNAD)

\(^\text{14}\) EPrints - [http://www.eprints.org/](http://www.eprints.org/)

\(^\text{15}\) OAIster - [http://www.oaister.org/](http://www.oaister.org/)
Incentivizing the Open Access Research Web: Publication-Archiving, Data-Archiving and Scientometrics

central repository through which users can search all the distributed OAI-compliant IRs. The IRs can also provide download and other usage metrics. In addition, one of us (Tim Brody) has created Citebase, a scientometric navigational and evaluational engine that can rank articles and authors on the basis of a variety of metrics.

Citebase's current database is not the network of IRs, because those IRs are still almost empty. (Only about 15% of research is being self-archived spontaneously today, because most institutions and funders have not yet mandated P-OA.) Consequently, for now, Citebase is instead focussed on the Physics Arxiv, a special central repository and one of the oldest ones. In some areas of physics, the level of spontaneous self-archiving in Arxiv has already been at or near 100% for a number of years now. Hence, Arxiv provides a natural preview of what the capabilities of a scientometric engine like Citebase would be, once it could be applied to the entire research literature (because the entire literature had reached 100% P-OA).

First, Citebase links most of the citing articles to the cited articles in Arxiv (but not all of them, because Citebase's linking software is not 100% successful for the articles in Arxiv, not all current articles are in Arxiv, and of course the oldest articles were published before OA self-archiving was possible). This generates citation counts for each successfully linked article. In addition, citation counts for authors are computed. However, this is currently being done for first-authors only: name-disambiguation still requires more work. On the other hand, once 100% P-OA is reached, it should be much easier to extract all names by triangulation – if persistent researcher-name identifiers have not yet come into their own by then.

So Citebase can rank either articles or authors in terms of their citation counts. It can also rank articles or authors in terms of their download counts (Figure 1). (Currently, this is based only on UK downloads: in this respect, Arxiv is not a fully OA database in the sense described above. Its metadata and texts are OA, but its download hypermetrics are not. Citebase gets its download metrics from a UK Arxiv mirror site, which Southampton happens to host. Despite the small and UK-biased download sample however, it has nevertheless been possible to show that early download counts are highly correlated with - hence predictive of - later citation counts.)

Figure 1. Citebase Ranking Metrics: List of current metrics on the basis of which Citebase can rank a set of articles.

---


18 http://www.citebase.org/help/

19 http://www.arxiv.org/

Citebase can generate chronometrics too: the growth rate, decay rate and other parameters of the growth curve for both downloads and citations (Figure 2). It can also generate co-citation counts (how often two articles, or authors, are jointly cited). Citebase also provides “hub” and “authority” counts. An authority is cited by many authorities; a hub cites many authorities; a hub is more like a review article; and an authority is more like a much cited piece of primary research.

![Figure 2. Citebase Sample Output:](image.png)

Citebase download/citation chronogram showing growth of downloads and growth of citations.

Citebase can currently rank a sample of articles or authors on each of these metrics, one metric at a time. We shall see shortly how these separate, “vertically based” rankings, one metric at a time, can be made into a single “horizontally based” one, using weighted combinations of multiple metrics jointly to do the ranking. If Citebase were already being applied to the worldwide P-OA network of IRs, and if that network contained 100% of each institution’s research publication output, along with each publication’s metrics, this would not only maximise research access, usage and impact, as OA is meant to do, but it would also provide an unprecedented and invaluable database for scientometric data-mining and analysis. OA scientometrics - no longer constrained by the limited coverage, access tolls and non-interoperability of today’s multiple proprietary databases for publications and metrics - could trace the trajectory of ideas, findings, and authors across time, across fields and disciplines, across individuals, groups, institutions and nations, and even across languages. Past research influences and confluences could be mapped, ongoing ones could be monitored, and future ones could be predicted or even influenced (through the use of metrics to help guide research employment and funding decisions).

Citebase today, however, merely provides a glimpse of what would be possible with an OA scientometric database. Citebase is largely based on only one discipline (physics) and uses only a few of the rich potential arrays of candidate metrics, none of them as yet validated. But more content, more metrics, and validation are on the way.

**The UK Research Assessment Exercise (RAE)**

The UK has a unique Dual Support System for research funding: competitive research grants are just one component; the other is top-sliced funding, awarded to each UK university, department by department, based on how each department is ranked by discipline-based panels of reviewers who assess their research output. In the past, this costly and time-consuming Research Assessment Exercise (RAE) has been based on submitting each researcher’s four best papers every six years to be ‘peer-reviewed’ by the appointed panel, alongside other data such as student counts and grant income (but not citation counts, which departments had been forbidden to submit and panels had been forbidden to consider, for both journals and individuals.).

21 [http://www.rcuk.ac.uk/aboutrcs/funding/dual/default.htm](http://www.rcuk.ac.uk/aboutrcs/funding/dual/default.htm)

22 [http://hero.ac.uk/rae/](http://hero.ac.uk/rae/)
To simplify the RAE and make it less time-consuming and costly, the UK has decided to phase out the panel-based RAE and replace it instead with ‘metrics’. For a conversion to metrics, the only problem is determining which metrics to use. It was a surprising retrospective finding (based on post-RAE analyses in every discipline tested) that the departmental RAE rankings turned out to be highly correlated with the citation counts for the total research output of each department (Figure 3).

Why would citation counts correlate highly with the panel’s subjective evaluation of researchers’ four submitted publications? Each panel was trying to assess quality and importance. But that is also what fellow-researchers assess, in deciding what to risk building their own research upon. When researchers take up a piece of research, apply and build upon it, they also cite it. They may sometimes cite work for other reasons, or they may fail to cite work even if they use it, but for the most part, a citation reflects research usage and hence research impact. If we take the panel rankings to have face validity, then the high correlation between citation counts and the panel rankings validates the citation metric as a faster, cheaper, proxy estimator.

New Online Research Metrics

Nor are one-dimensional citation counts the best we can do, metrically. There are many other research metrics waiting to be tested and validated: publication counts themselves are metrics. The number of years that a researcher has been publishing is also a potentially relevant and informative metric. (High citations later in a career are perhaps less impressive than earlier, though that no doubt depends on the field.) Total citations, average citations per year, and highest individual-article citation counts could all carry valid independent information, as could the average citation count (‘impact factor’)26 of the journal in which each article is published. But not all citations are equal. By analogy with Google’s PageRank algorithm, citations can also be recursively weighted in terms of how highly cited the citing article or author is. Co-citations can be informative too: being co-cited with a Nobel Laureate may well mean more than being co-cited with a postgraduate student. Downloads can be counted in the online age and could serve as early indicators of impact.
Citation metrics today are based largely on journal articles citing journal articles – and mostly just those 8000 journals that are indexed by ISI’s Web of Science. That only represents a third (although probably the top third) of the total number of peer-reviewed journals published today, across all disciplines and all languages. OA self-archiving can make the other two-thirds of journal articles linkable and countable too. There are also many disciplines that are more book-based than journal based, so book-citation metrics can now be collected as well (and Google Books and Google Scholar are already a potential source for book citation counts). Besides self-archiving the full-texts of their published articles, researchers could self-archive a summary, the bibliographic metadata, and the references cited by their books. These could then be citation-linked and harvested for metrics too. And of courses researchers can also self-archive their data (D-OA), which could then also begin accruing download and citation counts. And web links themselves provide a further metric that is not quite the same as a citation link.

Many other data could be counted as metrics too. Co-author counts may have some significance and predictive value (positive or negative: they might just generate more spurious self-citations). It might make a difference in some fields whether their citations are from a small, closed circle of specialists, or broader, crossing subfields, fields, or even disciplines: an ‘inbreeding/outbreeding’ metric can be calculated. Web link analysis suggests investigating ‘hub’ and ‘authority’ metrics. Patterns of change across time, ‘chronometrics,’ may be important and informative in some fields: the early rate of growth of downloads and citations, as well as their later rate of decay. There will be fast-moving fields where quick uptake is a promising sign, and there will be longer-latency fields where staying power is a better sign. ‘Semiometrics’ can also be used to measure the degree of distance and overlap between different texts, from unrelated works on unrelated topics all the way to blatant plagiarism.

Validating Research Metrics

The one last parallel panel/metric RAE, in 2008, will provide a unique natural testbed for validating the rich new spectrum of Open Access metrics against the panel rankings. A statistical technique called multiple regression analysis can compute the contribution of each individual metric to the joint correlation of all the metrics with the RAE panel rankings. Once initialized by being validated against the panel rankings, the relative weight of each metric can then be adjusted and optimised according to the needs and criteria of each discipline, with the panels only serving as overseers and fine-tuners of the metric output, rather than having to try to re-review all the publications. This will allow research productivity and progress to be systematically monitored and rewarded.27

This is a natural, ‘horizontal’ extension of Citebase’s current functionality but it does not need to be restricted to the UK RAE; once validated, the metric equations, with the weights suitably adjusted to each field, can provide ‘continuous assessment’ of the growth and direction of scientific and scholarly research. Not only will the network of P-OA IRs do double duty by providing access to research for researchers as well as serving as the database for the field of scientometrics, but it will also provide an incentive for data-archiving (D-OA) alongside publication-archiving (P-OA) for other fields too, both by providing an example of the power and potential of such a worldwide database in scientometrics and by providing potential new impact metrics for research data-impact, alongside the more familiar metrics for research publication-impact.

The Open Access Impact Advantage

Citebase has already been able to demonstrate that, in physics, OA self-archiving dramatically enhances citation impact (Figure 4a) for articles deposited in Arxiv, compared to articles in the same journal and year that are not self-archived. Lawrence had already shown this earlier for computer science. The advantage has since been confirmed in 10 further disciplines (Figure 4b) using the bibliographic metadata from the ISI Science and Social Science Citation Index (on CD-ROM leased to OST at UQAM) for millions of articles in thousands of journals, for which robots then trawled the web to see whether they could find a free online (OA) version of the full text. An OA/non-OA citation advantage – OA articles are cited more than non-OA articles in the same journal and year – has been found in every discipline tested so far (and in every year except the two very first years of Arxiv).

![Figure 4. Open Access citation advantage:](image)

Although only a small proportion of articles is currently being made Open Access, those articles are cited much more than those articles (in the same journal and year) that are not.

There are many contributors to the OA advantage, but currently, with spontaneous OA self-archiving still hovering at only about 15%, a competitive advantage is one of its important components (Figure 5). The University of Southampton, the first to adopt a self-archiving mandate, already enjoys an unexpectedly large “G-Factor,” which may well be due to the competitive advantage it gained from being the world’s first to mandate OA self-archiving: Figure 6) With the growing use of research impact metrics, validated by the UK RAE, the OA advantage will become much more visible and salient to researchers. Together with the growth of data impact metrics, alongside publication...
impact metrics and the prominent example of how scientometrics can data-mine its online database, there should now be a positive feedback loop, encouraging data self-archiving, publication self-archiving, OA self-archiving mandates, and the continuing development of the functionality of the underlying infrastructure.

Figure 5. Open Access Impact Advantage: There are many contributors to the OA Impact Advantage, but an important one currently (with OA self-archiving still only at 15%) is the competitive advantage. Although this advantage will of course disappear at 100% OA, metrics will make it more evident to researchers today, providing a strong motivation to reap the current competitive advantage.

Figure 6. Southampton Web Impact G-Factor: An important contributor to the University of Southampton's surprisingly high web impact 'G-Factor' is the fact that it was the first to adopt a departmental self-archiving mandate so as to maximise the visibility, usage and impact of its research output.30
The Law as Cyberinfrastructure

In almost everything we do, the law is present. However, we know that strict adherence to the law is not always observed for a variety of pragmatic reasons. Nevertheless, we also understand that we ignore the law at our own risk and sometimes we will suffer a consequence.

In the realm of collaborative endeavour through networked cyberinfrastructure we know the law is not too far away. But we also know that a paranoid obsession with it will cause inefficiency and stifle the true spirit of research. The key for the lawyers is to understand and implement a legal framework that can work with the power of the technology to disseminate knowledge in such a way that it does not seem a barrier. This is difficult in any universal sense but not totally impossible. In this article, we will show how the law is responding as a positive agent to facilitate the sharing of knowledge in the cyberinfrastructure world.

One general approach is to develop legal tools that can provide a generic permission or clearance of legal rights (e.g., copyright or patent) in advance (usually subject to conditions) that can be implemented before or at the point of use. This has become known as open licensing and will be discussed below in terms of copyright and patented subject matter.1

However, open licensing will not be adopted by everyone nor in every situation is it suitable. A generalisation is that it will be advocated in the context of publicly funded research producing tools and knowledge upon which platform technologies are built where considerations such as privacy are not an issue.

Where open licensing is not being used, the many parties to a collaborative endeavour will normally be required to map the scope and risk of their mutual endeavour through a contract. Contracts can take time to negotiate and, in many instances, promise to frustrate the fast paced and serendipitous nature of research fuelled by high powered cyberinfrastructure. To this end a number of projects throughout the world, for example The Lambert Project in the UK,2 the University Industry Demonstration Project (UIDP) in the USA,3 and (amongst other projects) the 7th Framework Project in the EU,4 have begun asking how we might be able to improve this situation. Suggestions include standard form or off the shelf contracts covering a variety of situations, a database of key clauses and, in the case of the UIDP project, a software based negotiation tool called the Turbo-Negotiator. Legal instruments that can match the dynamic of the technology and appear seamless and non-invasive are the goal. More work in this area is needed (and happening) and is critical to ensuring we have the law and technology of cyberinfrastructure working to complement each other.

In the remainder of this article we will focus on the open licensing model.


3 University Industry Demonstration Partnership - http://uidp.org/

Open Licensing

1. Open Content Licensing

From a legal perspective, one of the most significant responses to the technological advances that have revolutionized the creation and distribution of copyright materials during the last decade has been the development of new systems for licensing (or authorising) others to obtain access to and make use of the protected material. These new forms of licences – usually referred to as “open content” – are founded upon an acknowledgement of the existence of copyright in materials embodying knowledge and information, but differ from licences commonly used before the advent of the digital era in key respects. As well as being relatively short, simple and easy to read, they are standardised, conceptually interoperable with other open content licences, machine (computer) enabled and have the advantage that, since they are automated and do not require negotiation, they eliminate (or at least minimise) transaction costs. Running with the copyright material to which they are attached (thereby avoiding the privity issue where rights are conferred contractually), open content licences identify materials that are available for reuse and grant permissive rights to users, thereby facilitating access and dissemination.\
\
The most widely used of the open content licences are the Creative Commons licences. These licences attach to the copyright material and provide that anyone can reuse the material subject to giving attribution to the author of the material and subject to any of the optional conditions as selected by the licensor. The optional conditions are:

- non-commercial use;
- no derivative materials based on the licensed material are to be made; or
- share alike – others may distribute derivative materials based on the licensed material, but only under a licence identical to that covering the licensed material.

Creative Commons licences have more commonly been applied to publications than to research data. They have been particularly useful for academic authors depositing their publications in university or scholarly digital repositories or databases. Repositories help to make publications more accessible to the research and general communities. The advantage of a Creative Commons licence is that it tells people accessing the publication what they can and cannot do with the material, without the copyright owner having to deal with permissions on a case-by-case basis.

Below are two examples of scientific research publication projects that promote open access and reuse of material by utilising open content licensing models.

Example One – PLoS ONE

The Public Library of Science (PLoS) is a non-profit, open access, scientific publishing project that aims to create a library of peer-reviewed scientific and medical journals that are made available online without restrictions under open content licences. PLoS ONE is a peer-reviewed, scientific literature journal that enables scientific research to be published and disseminated within weeks, avoiding delays associated with traditional means of publication.
The features of PLoS ONE include:

- **rapid publication** – realising that the rapid publication and dissemination of research is one of the highest priorities, PLoS ONE ensures a streamlined electronic production workflow that ensures papers are published within weeks of submission;
- **freedom of use and ownership** – in accordance with the CC attribution licence, PLoS ONE enables users to read, copy, distribute and share papers freely without restrictions and formal permission, provided that the original author and source are cited; and
- **high impact** – PLoS ONE has been designed in light of the fact that papers published in OA journals are more likely to be read and cited given the lack of barriers to access.

### Example Two – Nature Precedings

Nature Precedings is an online database designed to allow scientific researchers to share pre-publication research, unpublished manuscripts, presentations, white papers, technical papers, supplementary findings and other scientific documents. Contributions are taken from biology, medicine (except clinical trials), chemistry and earth sciences. The database is free of charge to access and use, and is intended to provide a rapid means of disseminating emerging results and new theories, soliciting opinions and recording the provenance of ideas.

Nature Precedings aims to make scientific documents citable, globally available and stably archived. To this end, it can also be used as an archiving tool for scientists to store their work for their own future convenience.

Submissions made to Nature Precedings are screened by a professional curation team for relevance and quality, but are not subject to peer review. The database is designed to complement scientific journals by providing a more rapid and informal communication system, but submissions to Nature Precedings are not subject to the same rigorous and time-consuming reviews as submissions made to scientific journals.

The Nature Precedings website states that scientists should own copyright in a document and have permission from other copyright holders (e.g., co-authors), before they submit the document to Nature Precedings. Copyright then remains with the author. However, the website encourages scientists to release their work under a Creative Commons Attribution Licence so that content can be quoted, copied and disseminated, provided that the original source is correctly cited.

Authors who own copyright in their publication will be able to place a Creative Commons licence on their work, but if they have assigned copyright to their publisher or another party, they will need to ask permission from that party before they can attach a Creative Commons licence. A problem that often arises in this situation is that authors are unsure of whether they own copyright or their publisher owns copyright. Even when authors know that they have transferred copyright to their publisher, they may be reluctant to ask their publisher if they can attach a Creative Commons licence to their work for fear of jeopardising their relationship with the publisher.

These issues are best dealt with through established policies. Every research and academic institution should have in place policies relating to copyright management, including the licensing of copyright works. These policies should deal with the legal
impediments to making copyright material openly accessible, including determining who owns copyright, how to obtain necessary permissions from copyright owners and how to licence material in a way that grants the appropriate rights but retains the appropriate controls. The policies may also deal with non-legal issues, including how to get authors interested in open access repositories and how to assist authors in maintaining a positive relationship with their publisher while asserting additional rights.\footnote{For more information, see Kylie Pappalardo and Dr Anne Fitzgerald, “A Guide to Developing Open Access Through Your Digital Repository” (2007), available at http://www.oaklaw.qut.edu.au/node/32}

The Creative Commons open content principles have been extended to the sharing of scientific data and publications through the Science Commons Project.\footnote{Science Commons - http://sciencecommons.org/} As explained on the Science Commons website, Creative Commons licences can be used in relation to databases that attract copyright protection.\footnote{See http://sciencecommons.org/resources/faq/databases} An example of a database that uses a Creative Commons licence appears below.

Example Three – UniProtKB/Swiss-Prot Protein Knowledgebase

\begin{center}

\begin{tabular}{|l|}
\hline
UniProtKB/Swiss-Prot is a protein knowledgebase established in 1986 and maintained since 2003 by the UniProt Consortium. The UniProt Consortium is a collaboration between the Swiss Institute of Bioinformatics and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown University Medical Centre's Protein Information Resource. \\

The data held within UniProtKB includes protein sequences, current knowledge on each protein, core data (sequence data; bibliographical references and taxonomic data) and further annotation. The database is organised through a web interface that displays the data associated with each protein sequence. \\

The UniProt Consortium states that the public databases maintained by UniProt Consortium members are freely available to any individual and for any purpose. \\

A copyright statement on the UniProtKB website states: \\

\textit{We have chosen to apply the Creative Commons Attribution-NoDerivs Licence to all copyrightable parts of our databases. This means that you are free to copy, distribute, display and make commercial use of these databases, provided you give us credit. However, if you intend to distribute a modified version of one of our databases, you must ask us for permission first.}\footnote{UniProtKB - http://www.uniprot.org/terms/}

The UniProtKB open access system has been described as operating on an “honour system” on the basis that the user community is small and so accurately monitored by electronic tracking that non-compliance with the copyright licence would risk unacceptable costs in loss of reputation, peer pressure and possible denial of privileges.

\end{tabular}
\end{center}

2. Open Patent Licensing

Increased interest in sharing data also raises issues in relation to patents. Patents protect products and processes that are novel, useful and involve an inventive or innovative step. Patents must be registered and confer on the patentee the exclusive right to use or sell the patented product during a certain period of time (usually 20 years).
For researchers intending to seek patent protection for inventions derived from their research, a primary concern is whether they will be able to obtain a patent and whether disclosure of their data to other researchers could prevent them from obtaining a patent (because the product would no longer be “novel”). For researchers who do not intend to patent, a concern is whether another person could secure a patent over an invention that encompasses the researcher’s data.

Some researchers will be more interested in making their data openly available to advance research than in commercialising patented products or processes derived from their research. These researchers will not be concerned that public disclosure of their research data could prevent them from obtaining a patent because the invention is no longer novel or is obvious. However, disclosure of data, in itself, will not always be enough to prevent patenting. The problem arising from the public release of data is that it leaves the way open for another party to make improvements to the disclosed data and then make those improvements proprietary.

Claire Driscoll of the NIH describes the dilemma as follows:

“It would be theoretically possible for an unscrupulous company or entity to add on a trivial amount of information to the published…data and then attempt to secure ‘parasitic’ patent claims such that all others would be prohibited from using the original public data.”

Where information or data is used to develop a patentable invention, the subsequent patent rights may be broad enough to cover use of the actual data forming part of the invention. As Eisenberg and Rai explain:

“Although raw genomic data would not undermine claims to specific genes of identified function, annotated data might do so. A major goal of annotation is to identify coding regions in the genome and add information about the function of the protein for which the region codes.”

Consequently, some research projects have relied on licensing methods, similar to the open content copyright licences described above, in an attempt to keep the data “open,” rather than simply releasing the data into the public domain.

One example is the HapMap Project, which required anyone seeking to use research data in the HapMap database to first register online and enter into a click-wrap licence for use of the data. The licence prohibited licensees from filing patent applications that contained claims to particular uses of data obtained from the HapMap databases, unless that claim did not restrict the ability of others to freely use the data.

Another approach – currently being practised by the CAMBIA project - is to obtain a patent and then open licence the use of the patented invention on certain conditions. Some argue that, in specific areas, effective open access will only be achieved by allowing a certain level of use of the copyright and patented material.

2.a The CAMBIA Approach

CAMBIA is an international, independent, non-profit research institute led by well known scientist, Richard Jefferson. CAMBIA was designed to “foster innovation and a spirit of collaboration in the life sciences.” This goal is achieved through four interconnected work products:
The Law as Cyberinfrastructure

- **Patent Lens**, which provides tools to make patents and patent landscapes more transparent;
- **Biological Open Source Initiative (BiOS)**, which advocates for the sharing of life sciences technology and data through a series of licences;
- **BioForge**, a research portal (or repository) that makes data and technologies openly available for others to use in new innovations, whether for research, commercial use, or humanitarian use; and
- **CAMBIA’s Materials**, new technologies developed by CAMBIA, particularly in the field of genetics, which CAMBIA makes openly available under a BiOS licence.

CAMBIA has also applied for and obtained twelve patents of biological material in different patent offices around the world. CAMBIA’s approach involves obtaining patents over products or processes, but then licensing the use of those inventions under open terms. A primary object of this is to ensure that the biological material is not patented by others under restrictive terms, which do not allow for open access and use by others. Another object is to encourage innovation. CAMBIA

Strives to create new norms and practices for dynamically designing and creating the tools of biological innovation, with binding covenants to protect and preserve their usefulness, while allowing diverse business models for wealth creation, using these tools.19

CAMBIA has developed two open licences relevant to data – the BiOS Plant Enabling Technology Licence and the BiOS Genetic Resource Technology Licence. Paragraph 2.1 of each licence gives licensees

A worldwide, non-exclusive, royalty-free right and licence to make and use the IP & Technology for the purpose of developing, making, using, and commercializing BiOS Licensed Products without obligation to CAMBIA, including a sub-licence…20

This gives licensees the right to sub-licence the material, as long as it is sub-licensed under the same terms as contained in the original licence agreement.

CAMBIA’s model allows researchers to obtain patents over inventions that build upon CAMBIA’s research data. However, instead of using patent licences to “extract a financial return from a user of a technology,” CAMBIA advocates using a patent licence to “impose a covenant of behaviour.”19

According to CAMBIA, the purpose of the BiOS licences is that:

Instead of royalties, BiOS licensees must agree to legally binding conditions in order to obtain a licence and access to the protected commons. These conditions are that improvements are shared and that licensees cannot appropriate the fundamental “kernel” of the technology and improvements exclusively for themselves. Licensees obtain access to improvements and other information, such as regulatory and biosafety data, shared by other licensees. To maintain legal access to the technology, licensees must agree not to prevent other licensees from using the technology in the development of different products.19

By making the licence cost-free, CAMBIA hopes to encourage what founder Richard Jefferson terms:

The most valuable contribution to the license community: “freedom to innovate.”19
CAMBIA is currently developing a new version of the BiOS licence, which to our understanding will remove any positive obligation to share improvements in return for some type of covenant not to enforce rights in relation to patented improvements against members of the CAMBIA community.

Conclusion

Any research project should adopt a “mission-driven approach.” The question to be asked is, “What do we want to achieve?” The goal may be commercial gain, may simply be the advancement of research for the public good, or both. Open access to research data and publications should always be considered, especially in the case of publicly funded research. The level of access to and reuse of research data and publications that is to be allowed should ideally be determined at the outset of a research project.

From the commencement of a research project, it is imperative to have appropriate policies and frameworks in place. Policies must cover copyright management and data management. Copyright management policies should deal with copyright ownership rights and how copyright protected material is to be shared. Researchers should consider the various open content licensing models that can be applied to their copyright material. Data management plans should deal with how data is to be generated, managed and stored; data ownership rights and legal controls that may apply to data (including patents); and how access will be provided to the data and how the data will be disseminated.

Interestingly, some argue that, while open access in terms of copyright material will allow us to read that material and potentially to reproduce and electronically communicate it to colleagues, it most likely will not provide permission to use or exploit related patented material. One of the challenges for the near future will be to consider to what extent open access to publicly funded knowledge (e.g., that makes up tools or platform technologies in biotechnology) requires an accompanying commitment to allow a certain level of use of patented material. In this regard, the CAMBIA project provides an interesting approach that deserves close attention in coming years.

As lawyers, we hope that the law can adapt to facilitate the very great potential cyberinfrastructure promises us. To this end, we need to think of legal tools as being part of the infrastructure and work towards providing innovative models for the future.
“Infrastructure never gets adequately funded because it cuts across disciplinary boundaries, it doesn’t benefit particular groups. Infrastructure is a prerequisite to great leaps forward and is thus never captured within disciplinary funding, or normal governmental operations. We need to revise radically our conception of cyberinfrastructure. It isn’t just a set of tubes through which bytes flow, it is a set of structures that network different areas of knowledge… and that is software and social engineering, not fiber optic cable. The superhighways of the biological information age should not be understood as simply physical data roads, long ropes of fiber and glass. They need to be structures of knowledge. The Eisenhower Freeways of Biological Knowledge are yet to be built. But that doesn’t mean the task isn’t worth starting.”

- James Boyle, William Neal Reynolds Professor of Law, Duke University Law School

Knowledge sharing and scholarly progress

Knowledge sharing is at the root of scholarship and science. A hypothesis is formulated, research performed, experimental materials designed or acquired, tests run, data obtained and analyzed, and finally a publication. The scholar writes a document outlining the work for dissemination in a scholarly journal.

If it passes the litmus test of peer review, the research enters the canon of the discipline. Over time, it may become a classic with hundreds of citations. Or, more likely, it will join the vast majority of research, with less than two citations over its lifetime, its asserted contributions to the canon increasingly difficult to find – because, in our current world, citations are the best measure of relevance-based search available.

But no matter the fate of an individual publication, the system of publishing is a system of sharing knowledge. We publish as scholars and scientists to share our discoveries with the world (and, of course, to be credited with those discoveries through additional research funding, tenure, and more). And this system has served science extraordinarily well over the more than three hundred years since scholarly journals were birthed in France and England.

The information technology revolution: missed connections and lost opportunities

Into this old and venerable system has come the earthquake of modern information and communication technologies. The Internet and the Web have made publication cheap and sharing easy – from a technical perspective. The cost of moving, copying, forwarding, and storing the bits in a single scientific publication approach zero.

These technologies have created both enormous efficiency gains in traditional industries (think about how Wal-Mart uses the network to optimize its supply chains) and radical reformulation of industry (Amazon.com in books, or iTunes in music). Yet the promise of enormous increases in efficiency and radical reformulations have to date failed to make similar shattering changes to the rate of meaningful discovery in many scientific disciplines.
For the purposes of this article, I focus on the life sciences in particular. The problems I articulate affect all the scientific disciplines to one extent or another – but the life sciences represent an ideal discussion case. The life sciences are endlessly complex and the problems of global health and pharmaceutical productivity such an enormous burden that the pain of a missed connection is personal. Climate change represents a problem of similar complexity and import to the world, and this article should be contemplated as bearing on research there as well, but my topic is in the application of cyberinfrastructure to the life sciences, and there I’ll try to remain.

Despite new technology after new technology, the cost of discovering a drug keeps increasing, and the return on investment in life sciences (as measured by new drugs hitting the market for new diseases) keeps dropping. While the Web and email pervade pharmaceutical companies, the elusive goal remains “knowledge management:” finding some way to bring sanity to the sprawling mass of figures, emails, data sets, databases, slide shows, spreadsheets, and sequences that underpin advanced life sciences research. Bioinformatics, combinatorial drug discovery, systems biology, and an innumerable number of words ending with “-omics” have yet to relieve the skyrocketing costs and increase the percentage of success in clinical trials for new drug compounds.

The reasons for this are many. First and foremost, drug discovery is hard – really, really hard. And much of the low-hanging fruit has been picked. There are other reasons having to do with regulatory requirements, scientific competition, distortions in funding, and more. But there is one reason that stands out as both a significant drag on discovery and as a treatable problem, one that actually can be solved in the short term: we aren’t sharing knowledge as efficiently as we could be.

Forget “Web 2.0” – what about “Web 1.0” for science?

Much of the functionality we take for granted on the Web comes from making the choice to make sharing information easier, not harder. A good example is the way that Google interacts with the scientific literature.

With few exceptions, we rank the importance and relevance of scientific articles the way we always have, with citations and “impact factors.” Citations are longstanding and important. Impact factors – the number of citations to the articles in a journal – are the dominant metric for journal quality. And for a long time, citations were clearly the best, and perhaps the only, statistical measure of quality of a journal. In a print world, a world without hyperlinks and search engines and blogs and collaborative filtering, citations are a beacon of relevance.

But we live in a different world now. We have the ability to make connection after connection between documents, to traverse easily from one page to another page. Hyperlinks are cheap and they’re everywhere. It was a conscious design decision made by Tim Berners-Lee to allow this functionality. Other competing systems thought it insane that the WWW would let just anyone link to just anything else – those links might be broken, leading to the dreaded “404 not found” – and that would obviously kill the WWW! It hasn’t worked out that way. The choice to allow users the right to make hyperlinks, to make hyperlinking easy and fast, not only did not kill the Web, it is a big part of what makes Google searching so powerful.

Google ranks pages by downloading enormous chunks of the Web and running software that analyzes the linkages between Web pages. The system quite literally depends on there being lots and lots of links, many of them perhaps useless on their
own, but which in aggregate provide hints of relevance. Thus, the number one Google search on the words "Science Commons" is the Web page analyzed with the words "Science Commons" that has the most links pointing to it. There's more complexity, obviously, but that's a big part of the idea.

If those Web pages were private, the page ranking system wouldn't work. The Web pages themselves are part of the infrastructure on which Google operates, on which millions of startup dreams are founded. In a world where every page was locked, where every Web designer had to ask permission to make an inbound link...we wouldn't have the sprawling value creation we associate with the Internet. It would look a lot more like Prodigy looked a long time ago: a closed network that can't compete in the end with the open networks.

Put another way, we have far more efficiency brought to bear on accelerating our capability to order consumer products than we do on accelerating our capability to perform scientific research. Biological reagents and assays are re-invented and reverse-engineered by readers of "papers" – years of laboratory work, data, living DNA and more compressed down to the digital equivalent of a sheet of dead tree.

We need the Web to work as well for science as it does for other areas. The capabilities now exist to integrate information, data, physical tools, order fulfillment, overnight shipping, online billing, one-to-one orders, and more. If we are to solve the persistent health problems of the world, of infectious disease in the developing world and rare disease in the developed world, the "web 1.0" efficiency is an obvious benefit to bring to the life sciences.

But these advances we take for granted in daily life, like Google's relevance based search of the entire Web, eBay's many-to-many listing and fulfillment, Amazon's one-click ordering, won't come to science accidentally. There's a significant collective action problem blocking the adoption of these systems and preventing the network effects from taking over in discovery.

But it's not just the Google issue, which simply forces us to forego existing technology and focus on citations as we have always done. Citations carry more constrictions as a search metaphor. You are likely to enter the citation search ranked world when you know what to search. But you might not know how to say it in the nomenclature of a related, but distinct, discipline.

It goes on. Citation linkages between papers are subject to enormous social pressures. One cites the papers of one's bosses, of course. Review articles can skew impact factors. And of course, a tried-and-true way to get a heavily cited article remains to be horrifically, memorably wrong.

And over the long term, the lack of more complex and realistic interconnections between articles – a web, a set of highways, an infrastructure connecting the knowledge – is that we can't begin to integrate the articles with the databases. That's because the actors in the articles (the genes, proteins, cells and diseases) are described in hundreds of databases.

And if we could link the articles not just to each other by a richer method than citations, but to the databases, we can inch closer to the goal of a Rosetta Stone of knowledge, the small element upon which we can begin to have truly integrated, public knowledge spaces. That would in turn allow us to begin automatically indexing the data that robots are producing in labs every day, to meaningfully extract actionable information from the terabytes of genomic data we are capable of producing.
You get those virtues only where you are dealing with the knowledge claims themselves, not the sub-component of them the people in the field thought it worthy to expose. Only a better infrastructure gets you there, just as the modern highway system in the US allowed for better efficiency than the evolved hodge-podge of state highways. Citation linkages are very useful (and a later version needs to cross reference them with these highways we propose – we didn't throw away the state highways, after all!). This is simply a different set of tasks, and one that can be accomplished if enough smart people have enough rights and time to work on the knowledge.

But sadly, no one – no one! – has the right to download and index with scholarly literature without burning years of time and money in negotiations. Google has spent years asking for the right to index a lot of the scholarly canon for its Scholar project, but that's not some open land trust for any researcher to work on. It's just for Google. And the fact that Google alone has the right to index articles for such a service means that the next Google, the next set of genius entrepreneurs with a taste for search coding away in the halls of the local university, can't apply their skills to the sciences.

Though we have the capability to drastically increase the sharing at a much lower cost through digital distribution, search, and more, the reaction has been instead to segregate knowledge behind walls of cost, technology, and competitive secrecy. The net result is that we're doing things the way we always did, but only somewhat faster. If we want to bring both efficiency gains and radical transformation to the life sciences, getting more knowledge online, with the rights to transform, twist, tag, reformat, translate, and more, is going to be part of the solution. We have to start allowing the best minds of the world to apply the newest technologies to the scientific problems facing us.

There isn't a single, open "Web" of content to search – it's owned by a group of publishers who prevent indexing and search outside their own engines, and who use copyright and contracts to keep it locked up. There isn't any easy way to find the tools of biological science – it's a complicated social system of call-and-response, of email and phone calls, of "are you in the club of scientists worth partnering with?" questions and answers. And there isn't a standard way to get your orders fulfilled, but instead a system in itself of materials transfer and ordering, university technology transfer, commercial incentives, deliberate withholding, and more. We don't have the Web working yet for science.

**Infrastructure for knowledge sharing: Science Commons**

I work on a project called Science Commons – part of the Creative Commons (CC) non-profit organization (CC is the creator of a set of legal tools for sharing copyrighted works on the Web using a modular set of machine-readable contracts. CC licenses cover more than 150,000,000 copyrighted objects on the Web, including such high-impact offerings as BioMed Central, Public Library of Science, Nature Precedings, Hindawi Publishing, and the Uniprot database of proteins.) Science Commons is building a toolkit of policy, contracts, and technology that increases the chance of meaningful discovery through a shared, open approach to scientific research. We're building part of the infrastructure for knowledge sharing, and we're also deploying some test cases to demonstrate the power of investing in this kind of infrastructure.

Science Commons isn't alone. Sharing approaches that address a single piece of the research cycle are making real, but painfully slow, progress. Open Access journals are far from the standard. Biological research materials are still hard to find and harder
to access. And while most data remains behind the firewall at laboratories, even those data sets that do make it online are frequently poorly annotated and hard to use. The existing approaches are not creating the radical acceleration of scientific advancement that is made possible by the technical infrastructure to generate and share information.

Science Commons represents an integrated approach – one with potential to create this radical acceleration. We are targeting three key blocking points in the scientific research cycle – access to the literature, experimental materials, and data sharing – in a unified approach. We are testing the hypothesis that the solutions to one problem represent inputs to the next problem, and that a holistic approach to the problems discussed here potentially benefits from network effects and can create disruptive change in the throughput of scientific research. I will outline how these approaches represent tentative steps towards open knowledge infrastructure in the field of neuroscience.

Knowledge Overload for Huntington's Disease

![Biological pathway for Huntington’s Disease](image)

Figure 1. Biological pathway for Huntington’s Disease.

Above is the biological pathway for Huntington’s Disease. This pathway is like a circuit – it governs the movement of information between genes and proteins, processes and locations in the cell. This one is a relatively simple pathway, as far as such things go. More complex pathways can have hundreds of elements in the network, each “directional” – not just linked like Web pages, but typed and directed links, where the kind of relationship and the causal order are vital both in vitro and in silico.

In this pathway, the problem is the HD gene in the middle of the circuit – if that gene is broken, it leads to a cascade that causes a rare, fatal disease where the brain degenerates rapidly. Although the genetic element has been understood for a long time, there is no cure. Not enough people get the disease for it to be financially worth finding a cure, given how expensive it is to find drugs and get them to market. That’s cold comfort to the tens of thousands of people who succumb each year and to their families who know they have a 50% chance of passing on the gene and disease to their children. But that’s the reality.
Years of research have led to an enormous amount of knowledge about Huntington’s. For example, a search in the U.S. government’s free Entrez Web resource on “Huntington’s” yields more than 6,000 papers, 450+ gene sequences, 200+ protein sequences, and 55,000 expression and molecular abundance profiles. That’s a lot of knowledge. The papers alone would take 17 years to read, at the rate of one paper per day (and that’s assuming no new papers are published in the intervening years). Yet Huntington’s actually provides a relatively small result. One of the actors in the pathway is called “TP53.” That brings up another 2,500 papers, but also brings up (in an indirect link to a page about sequences for this entity) that it has a synonym: “p53.” Entrez brings back 42,000 articles from that search string – 115 years to read!

It goes on and on. And having all of this knowledge is wonderful. But there are more than a few problems here. The first is something you might call “cognitive overload.” Our brains simply aren’t strong enough to take in 500,000 papers, read them all, build a mental model of the information, and then use that information to make decisions – decisions like, what happens if I knock out that CASP box in the pathway, with 27,000 papers?

The other problems stem from the complexity of the body. In what other circuits is each entity in the pathway involved? What about those tricky causal relationships above and below it in the circuit? What are the implications of intervention in this circuit on the other circuits?

Some of these entities, the boxes in the diagram, are metaphorically similar to the airport in Knoxville, TN. Knocking out that airport doesn’t foul up a lot of air traffic. But some of these – P53 for example – are more like Chicago. Interfering with that piece of the network reverberates across a lot of unrelated pieces of the network. That’s what we call side effects, and it’s one of the reasons drugs are so expensive – we know that we can impact this circuit, but we don’t realize how badly it affects everything else until we run the drug in the only model available that covers all possible impacts: the human body.

And this is just the papers. There are thousands of databases with valuable information in them. Each of them has different access privilege conditions, different formats, different languages, different goals, wasn’t designed to work with anything else, and are maintained at different levels of quality. But they have vital – or potentially vital, to the right person asking the right question – information. And if we could connect the knowledge around these knowledge sources into a single network we just might be able to leverage the power of other technologies built for other networks. Like Google – but maybe more like the next Google, something as dramatically better and different and radical as Google was when we first saw it in the late 1990s.

There are two problems to be addressed here. One is the materials that underpin this knowledge, these databases and articles. Those materials are “dark” to the Web, invisible, not subject to the efficiency gains we take for granted in the consumer world. The second is the massive knowledge overload that the average scientist faces. I’ll outline two proofs of concept to demonstrate the value of investment in infrastructure for knowledge sharing that can address these problems.

“Web 1.0” Proof of Concept: E-commerce for biological materials

The Biological Materials Transfer Agreement Project (MTA) develops and deploys standard, modular contracts to lower the costs of transferring physical biological materials such as DNA, cell lines, model animals, antibodies and more. Materials represent
tacit knowledge – generating a DNA plasmid or an antibody can take months or years, and replicating the work is rarely feasible. Gaining access to those materials is subject to secrecy, competition, lack of resources to manufacture materials, lack of time, legal transaction costs and delays, and more.

There is significant evidence that the transfer of biological materials is subject to significant slowdowns. Campbell\(^1\) and Cohen\(^2\) have each demonstrated that materials are frequently denied. Legal barriers are part of the problem – more so than patents – but the greater problem is frequently the competition, secrecy, and incentive systems involved.

This is why we brought in funders of disease research and institutional hosts of research from the beginning – this is the part of infrastructure that is social engineering, not software. The secrecy and competition do not maximize the likelihood of meaningful discovery coming from limited funding, and thus funders (especially of rare or orphan diseases) have a particular incentive to maximize the easy movement of biological materials to maximize follow-on research.

The MTA project covers transfers among non-profit institutions as well as between non-profit and for-profit institutions. It integrates existing standard agreements into a Web-deployed suite alongside new Science Commons contracts and allows for the emergence of a transaction system along the lines of Amazon or eBay by using the contracts as a tagging and discovery mechanism for materials.

This metadata driven approach is based on the success of the Creative Commons licensing integration into search engines and further allows for the integration of materials licensing directly into the research literature and databases so that scientists can “one-click” inline as they perform typical research. And like Creative Commons licensing, we can leverage the existing Web technologies to track materials propagation and reuse, creating new data points for the impact of scientific research that are more dimensional than simple citation indices, tying specific materials to related peer-reviewed articles and data sets.

The MTA project was launched in collaboration with the Kauffman Foundation, the iBridge Network of university technology transfer offices, and neurodegenerative disease funders. It currently includes more than 5,000 DNA plasmids covered under standard contracts and is available through the Neurocommons project described in the next section.

**Proof of concept in knowledge sharing: a semantic web for neuroscience**

In collaboration with the W3C Semantic Web Health Care and Life Science interest group, we are integrating information from a variety of standard sources to establish core interoperable content that can be used as a basis for bioinformatics applications. The combined whole is greater than the sum of its parts, since queries can cut across combinations of sources in arbitrary ways.

We are also providing an operational knowledge base that has a standard, open query endpoint accessible by Internet. The knowledge base incorporates information marshaled from more than a dozen databases, ontologies, and literature sources.

Entities discussed in the text, such as proteins and diseases, need to be specifically identified for computational use, as do the entities' relationships to the text and the text's assertions about the entities (for example, a particular asserted relationship between a
protein and a disease). Manual annotation by an author, editor, or other "curator" may capture the text's meaning accurately in a formal notation. However, automated natural language processing (including entity extraction and text mining) is likely to be the only practical method for opening up the literature for computational use.

We were only able to process the abstracts of the literature as the vast majority of the scientific literature is locked behind firewalls and under contracts that explicitly prevent using software to automatically index the full text where it is accessible. Although most papers run more than five pages, the abstracts typically were limited to a paragraph.

For tractability, we limited the scope to the organisms of greatest interest to health care and life sciences research: human, mouse, and rat. We are also providing the opportunity for interested parties to "mirror" the knowledgebase and we encourage its wide reuse and distribution.

In combination with the data integration and text processing, we are also offering a set of analytic tools for use on experimental data. The application of prior knowledge to experimental data can lead to fresh insights. For example, a set of genes or proteins derived from high throughput experiments can be statistically scored against sets of related entities derived from the literature. Particular sets that score well may indicate what's going on in the experimental setting.

In order to help illustrate the value of semantic Web practices, we are developing statistical applications that exploit information extracted from RDF data sources, including both conversions of structured information (such as Gene Ontology annotations) and relationships extracted from literature. The first tools we hope to roll out are activity center analysis for gene array data and set scoring for profiling of arbitrary gene sets, donated to Science Commons by Millennium Pharmaceuticals.

Taken together, we call these three efforts the Neurocommons – an open source, open access knowledge management platform, with an initial therapeutic focus on the neurosciences. And we hope to use the Neurocommons both as a platform to facilitate knowledge sharing and to secure empirical evidence as to the value of shared knowledge in sciences.

Conclusion: cyberinfrastructure for knowledge sharing

The Neurocommons project is a very good start. It shows the potential of shared knowledge systems built on open content. And it has the potential to explode through horizontal downloading, editing, and reposting, just as the Web exploded. The idea of connectivity via "viewing source" is an explicit part of our design methodology, and our tools have already been picked up and integrated into such systems as the Mouse BIRN Atlasing Toolkit (MBAT), which was built from the combined efforts of groups within the Mouse BIRN (Biomedical Informatics Research Network, a distributed network of researchers with more than $25,000,000 in U.S. Government funding).

But the Neurocommons is, at root, a proof of concept. And from it we are learning some basic lessons about the need for infrastructure for knowledge sharing. Science Commons is on a daily basis forced to create namespaces, persistent URLs, and line after line of "plumbing code" to wire together knowledge sources.

If we are going to get to the goal stated above, of dramatic increases in efficiency and radical transformation of outmoded discovery models, we are going to need a lot of infrastructure that doesn't yet exist.
We need publishers to look for business models that aren’t based on locking up the full text, because the contents of the journals – the knowledge – is itself part of the infrastructure, and closed infrastructure doesn’t yield network effects. We need open, stable namespaces for scientific entities that we can use in programming and integrating databases on the open Web, because stable names are part of the infrastructure. We need real solutions about long-term preservation of data (long-term meaning a hundred years or more). We need new browsers and better text processing. We need a sense of what it means to “publish” in a truly digital sense, in place of the digitization of the paper metaphor we have in the PDF format. We need infrastructure that makes it easy to share and integrate knowledge, not just publish it on the Web.

None of this is easy. Much of it is very, very hard. But the current system is simply not working. And the reward of pulling together what we already know into open view, in open formats, where geniuses can process and exploit it, could be a world in which it is faster, easier, and cheaper to find drugs and cure disease. This is possible. We just have to have the vision and courage to build the highways.
This article began with a simple attempt to identify trends that are changing scholarly communication. I expected to find trends that are supporting the progress of Open Access (OA) and trends that are opposing it or slowing it down. The resulting welter of conflicting trends might not give comfort to friends or foes of OA, or to anyone trying to forecast the future, but at least it will describe this period of dynamic flux. It might even explain why OA isn’t moving faster or slower than it is.

But with few exceptions, I have only found trends that favor OA. Maybe I have a large blind spot; I’ll leave that for you to decide. I’m certainly conscious of many obstacles and objections to OA, and I address them every day in my work. The question is which of them represent trends that are gaining ground.

While it’s clear that OA is here to stay, it’s just as clear that long-term success is a long-term project. The campaign consists of innumerable individual proposals, policies, projects, and people. If you’re reading this, you’re probably caught up in it, just as I am. If you’re caught up in it, you’re probably anxious about how individual initiatives or institutional deliberations will turn out. That’s good; anxiety fuels effort. But for a moment, stop making and answering arguments and look at the trends that will help or hurt us, and will continue to help or hurt us even if everyone stopped arguing. For a moment, step back from the foreground skirmishes and look at the larger background trends that are likely to continue and likely to change the landscape of scholarly communication.

I’ve found so many that I’ve had to be brief in describing them and limit the list to those that most affect OA.

1. First there are the many trends created by OA proponents themselves: the growing number of OA repositories, OA journals, OA policies at universities, OA policies at public and private funding agencies, and public endorsements of OA from notable researchers and university presidents and provosts. Each new OA repository, journal, policy, and endorsement contributes to a growing worldwide momentum and inspires kindred projects elsewhere. Funding agencies are now considering OA policies, in part because of their intrinsic advantages (for increasing return on investment by increasing the visibility, utility, and impact of research) and in part because other funding agencies have already adopted them. The laggards are being asked why the research they fund is less worth disseminating than the research funded elsewhere. The growing mass of OA literature is becoming critical in the sense that the growth is now a cause, and not just an effect, of progress. OA literature is the best advertisement for OA literature; the more we have, the more it educates new scholars about OA, demonstrates the benefits of OA, and stimulates others to provide or demand it.

2. Although knowledge of OA among working researchers is still dismally low, every new survey shows it increasing, and every new survey shows increasing rates of deposits in OA repositories and submissions to OA journals. The absolute numbers may still be low, but the trajectories are clearly up.

3. More scholars are posting their articles online even if they don’t have their publisher’s permission. As long ago as October 2005, Alma Swan found that seven
out of eight articles published in the inaugural issue of *Nature Physics*, which had a six month embargo on self-archiving, were free online somewhere on the day of publication. Regardless of what this shows about copyright, it shows a healthy desire for OA. We don’t know whether the volume of OA produced this way is large or small, but already (according to the April 2007 RIN study) readers routinely try Google and email requests to the author before interlibrary loan when they hit a pay-per-view screen at a journal web site. Both trends—posting OA copies, with or without permission, and searching for OA copies—are growing.

4. Subscription prices are still rising faster than inflation after more than three decades. A March 2006 study by the ALPSP found that high journal prices cause many more cancellations than OA archiving. Rapidly rising prices undermine the sustainability of the subscription model. They undermine publisher arguments that all who need access already have it. They undermine publisher arguments that we shouldn’t fix what isn’t broken. They undermine the credibility, and even the good faith, of publishers who argue that OA threatens peer review, by threatening their subscriptions, when their own hyperinflationary price increases are far more potent in the same cause. They strengthen the incentives for libraries, universities, funders, and governments to join the campaign for OA.

5. The cost of facilitating peer review is coming down as journal management software improves, especially the free and open source packages like DPubS, E-Journal, ePublishing Toolkit, GAPworks, HyperJournal, OpenACS, SOPS, TOPAZ, and the segment leader, Open Journal Systems. This reduces the cost of publishing a peer-reviewed journal, improves the financial stability of peer-reviewed OA journals, and multiplies the number of business models that can support them.

6. More publishers are launching hybrid OA journals, which will make any of their articles OA if an author or author-sponsor pays a publication fee. I’ve been critical of many of these programs, in part for high prices and needless restrictions that reduce author uptake. But even with low uptake they will (slowly) increase the volume of OA literature, (slowly) spread the OA meme to more authors and readers, and (slowly) give publishers first-hand experience with the economics of one kind of OA publishing.

7. More journals are willing to let authors retain key rights, especially the right of postprint archiving, and more are willing to negotiate the terms of their standard copyright transfer agreement. More authors are willing to ask to retain key rights and more institutions are willing to help them. More organizations are drafting “author addenda” (contract modifications to let authors retain key rights), and more universities are encouraging their faculty to use them. There are now major addenda from SPARC, Science Commons, OhioLINK, the Committee on Institutional Cooperation, and a handful of individual universities. The default in most journals and fields is still for authors to transfer nearly all rights to publishers, including the right to decide on OA for the peer-reviewed postprint. Corrections to this imbalance haven’t gone nearly far enough, but the slope of the curve is definitely up.

8. More and more toll-access (TA) journals are dropping their print editions and becoming online-only. Steven Hall of Wiley-Blackwell predicts that 50% of scholarly journals will become online-only within the next 10 years. As high-quality, high-prestige journals make this transition, scholars who still associate quality and prestige with print will (happily or unhappily) start to unlearn the association. At the same time, the rise of high-quality, high-prestige OA journals will confirm the new recognition that quality and medium are independent variables. TA publishers are joining OA advocates in creating an academic culture in which online publications earn full credit for promotion and tenure. Online publications need not be OA, of course, but changing the culture to accept online publications is more than half the battle for changing the culture to accept OA publications.
9. More journals, both OA and TA, encourage or require OA to the data underlying published articles. Major publisher associations like ALPSP and STM, which lobby against national OA policies for literature, encourage OA for data. Even when these policies don't cover peer-reviewed articles, they accelerate research, demonstrate the benefits of unrestricted sharing, and build expectations and momentum for OA in other categories.

10. More journals (OA and TA) are integrating text and data with links between text and data files, tools to go beyond viewing to querying data, dynamic charts and tables to support user-driven what-if analyses, and multimedia displays. Some types of integration can be done in-house at the journal and kept behind a price wall. But other types, especially those added retroactively by third parties, require OA for both the text and data. OA invites motivated developers to use their cleverness and creativity, and the existence of motivated developers invites authors and publishers to make their work OA.

11. Thomson Scientific is selecting more OA journals for Impact Factors, and more OA journals are rising to the top cohort of citation impact in their fields. For scholars and institutions using Impact Factors as crude metrics of quality, this trend legitimates OA journals by showing that they can be as good as any others. There are other gains as well. Because OA increases citation impact (studies put the differential at 40-250%, depending on the field), high-quality OA journals can use citation impact to shorten the time needed to generate prestige and submissions commensurate with their quality. OA journals with rising impact are successfully breaking a vicious circle that plagues all new journals: needing prestige to attract high-quality submissions and needing high-quality submissions to generate prestige. OA journals high in impact, quality, and prestige are improving their own fortunes and changing expectations for other OA journals. But they are also making two little-known truths better known: first, that OA can help journals, not just readers and authors, and second, that a journal's impact, quality, and prestige do not depend on its medium or business model - except insofar as OA models actually amplify impact.

12. New impact measurements are emerging that are more accurate and nuanced, more inclusive, more timely, and less expensive than Impact Factors. These include Eigenfactor, h-Index, Journal Influence and Paper Influence Index, Mesur, Usage Factor, Web Impact Factor, and Y Factor. What most of them have in common is the harnessing of new data on downloads, usage, and citations made possible by OA. In this sense, OA is improving the metrics and the metrics are improving the visibility and evaluation of the literature, especially the OA literature.

13. Download counts are becoming almost as interesting as citation counts. Not only are they being incorporated into impact metrics, but a CIBER study (September 2005) discovered that senior researchers found them more credible than citations as signs of the usefulness of research. A Brody-Carr-Harnad study (March 2005) found that early download counts predict later citation counts. No one thinks download counts mean the same thing as citation counts, but they’re easier to collect, they correlate with citation counts, and they’re boosted by OA. In turn they boost OA; repository managers have learned that showing authors their download tallies will encourage other authors to deposit their work.

14. The unit of search has long since shifted from the journal to the article. Now the unit of impact measurement is undergoing the same transition. Authors and readers still care about which journal published a given article, but they care less and less about which other articles appeared in the same issue. More and more, finding a relevant article in an OA repository, separated from its litter mates, gives searching scholars all they want.

15. Big publishers are still getting bigger, merging with one another (most recently, Wiley and Blackwell), acquiring smaller publishers, and acquiring journals. Market consolidation is growing, monopoly power is growing, and bargaining power by subscribers is declining. This interests government anti-trust officials - who in the UK, for example, would already have acted
if the OA movement hadn't give them a reason to watch and wait. It gives the players representing research rather than for-profit publishing (universities, libraries, funders, and governments) additional incentives to work for OA. It also gives the smaller, non-profit publishers, excluded from big deals and competing for limited subscription funds against the market titans, reasons to consider the big publishers more threatening than OA and reasons to consider OA a survival strategy.

16. More for-profit companies are offering services that provide OA or add value to OA literature such as repository services, search engines, archiving software, journal management software, indexing or citation tracking services, publishing platforms, print preservation. These services create or enhance OA literature, fill the cracks left by other services, create a market for OA add-ons, and show another set of business judgments that OA is on the rise. If you think, as I do, that one promising future for non-OA publishers is to shift from priced access to priced services for adding value to OA literature, then these projects can help opponents become proponents.

17. More mainstream, non-academic search engines like Google, Yahoo, and Microsoft are indexing OA repositories and journals. This makes OA content easy to find for users unacquainted with more specialized tools. This in turn helps persuade even more publishing scholars that OA really does increase visibility and retrievability. Hence, it helps correct a particularly common and harmful misunderstanding among authors: that work on deposit in an institutional repository is only visible to people who visit that particular repository and run a local search.

18. While OA journals and repositories continue to multiply, other vehicles for delivering OA are finding more and more serious scholarly applications: blogs, wikis, ebooks, podcasts, RSS feeds, and P2P networks. This is more than a geeky desire to play with cool new tools. It's a desire to find ways to bypass barriers to communication, collaboration, and sharing. Serious researchers are discovering that these tools are actually useful, not just cool, and are taking advantage of them. Like cell phones, wifi, and the Internet itself before them, these tools are overcoming the stigma of being trendy and moving from the periphery to the mainstream.

The direct benefit is that all these tools presuppose OA. Their widespread use enlarges the volume of OA research communication and spreads the conviction that research benefits from both the speed and the reach of OA. The indirect benefit is that they foster disintermediation (and hence reduce costs and delays) without sacrificing peer-mediated forms of quality control. Since the rise of peer-reviewed journals in the 17th century, most publicly disseminated works of scholarship have been vetted and distributed by publishers. Letters and lectures were exceptions. Today, the categories of exceptions, the volume of research-reporting they represent, and their integration into the workflow of ordinary research, are all growing.

19. New and effective tools for collaboration are triggering adoption and excitement, e.g., social tagging, searching by tags, open peer commentary, searching by comments, social networking, community building, recruiting collaborators, facilitating work with collaborators you already have, following citation trails backwards and forwards, following usage-based "similar to" and "recommended" trails, open APIs, open standards, and mash-ups. Collaboration barriers are becoming almost as irritating and inimical to research as price and permission barriers. A new generation of digital scholars is deeply excited by the new collaboration services that presuppose and build on OA.

To focus on social tagging and folksonomy tools for a moment; when applied to research literature (Connotea, CiteULike) and combined with OA and search engines, they do more than OA alone, or OA plus search engines, to enhance the discoverability of OA literature. Because tags are added retroactively to open content by uncoordinated users, they stimulate the
imagination of creative developers; once we have OA to literature and data, we can add layers of utility indefinitely.

20. Interest in OA and projects to deliver on that interest are growing fast in the humanities. Humanists are exploring OA for books and journals, and exploring the universe of useful services that can be built on an OA foundation, from searching and annotation to text-mining, co-writing, and mash-ups. We already knew that OA was useful to scholars, as authors and readers, in every field. But the humanities are now showing that OA is not limited to fields with high journal prices (to serve as a goad) or high levels of research funding (to pay for it).

21. Huge book-scanning projects, particularly those from Google, the Open Content Alliance, The European Library, the Kirtas-Amazon partnership, and Project Gutenberg, steadily increase the number of print books available in some free-to-read digital form. We'll soon reach a crossover point when more full-text, public-domain books are freely available online than on the shelves of the largest university library. Apart from lowering the access barriers to a large and uniquely valuable body of literature, the book-scanning projects together create one more large real-time demonstration that useful literature is even more useful when it's OA.

22. At the same time, the price of book scanning is dropping quickly. The usefulness of book literature and the absence of legal shackles on public domain texts are attracting large corporations, whose investments and competition are driving down the costs of digitization. The repercussions will be felt in every category of print literature, including the back runs of print journals.

23. Steady decreases in the size and cost of hardware and memory are making possible steady increases in the volume of data people can carry in their palm or pocket. Free offline access to usefully large digital libraries is on the horizon. Every effort for free online access will free up content for portable offline access as well.

24. Evidence is mounting that OA editions increase the net sales of print editions for some kinds of books, including scholarly monographs. This not only enlarges the corpus of OA literature, but chips away at the simplistic, reflexive fear that OA is incompatible with revenue and profit. Every month another university press explores this space by creating an imprint or division dedicated to dual-edition monographs (OA editions and priced/printed editions), OA-plus-POD monographs, or OA-only monographs.

On the other hand, OA books may only stimulate sales of print editions as long as most people dislike reading whole books on screen. This trend may be reversed by a counter-trend to improve ebook readers.

25. A textbook pricing crisis is stimulating OA solutions just as the journal pricing crisis before it stimulated OA solutions. There are now major projects to produce OA textbooks from the following:

- Amedeo
- Atomic Dog
- BookPower
- the California Open Source Textbook Project
- CommonText
- the Free Curricula Center
- Free High School Science Texts
- Freeload Press
- FreeTechBooks
- Global Text Project
- Libertas Academica
- Liberty Textbooks
- Medical Approaches
- MedRounds Publications
- next
text
- the Open Textbook Project
- the Potto Project, Science Classics
- Textbook Revolution
- Wikibooks.
26. More universities and independent non-profits are creating open courseware, OA teaching and learning materials, and other open educational resources (OERs). These help all teachers and students, even those at affluent schools. Like the kindred movement for OA to research literature, this one is demonstrating to a growing audience that useful content is more useful when OA. It also helps generalize a fact highlighted by researchers: that content creators who depend on salaries rather than royalties, or who write for impact and not for money, have everything to gain and nothing to lose by consenting to OA.

27. There is a rising awareness of copyright issues in the general public, rising frustration with unbalanced copyright laws, and rising support for remedies by governments (legislation) and individuals (Creative Commons licenses and their equivalents). Copyright laws are still grotesquely unbalanced, and powerful corporations who benefit from the imbalance are fighting to ensure that they are not revised in the right direction any time soon. But in more and more countries an aroused public is ready to fight to ensure that they are not revised in the wrong direction either, something we haven't seen in the entire history of copyright law.

However, this only guarantees that the content industry will have a fight, not that users and consumers will win. Nearly every week the content industry scores another victory, either in court or in arm-twisting another developing country into harmonizing its copyright laws with the unbalanced North. There are victories for balance as well, but less often in courts and legislatures than in think tanks, conference declarations, government commissions, and newspaper editorials. The trend is only for awareness and opposition, but it's sharply up.

28. The shock of the new is wearing off. OA is gradually emerging from the fog of misunderstanding. For this one, I won't be brief.

Scholars who grew up with the Internet are steadily replacing those who grew up without it. Scholars who expect to put everything they write online, who expect to find everything they need online, and who expect unlocked content they may read, search, link, copy, cut/paste, crawl, print, and redistribute, are replacing those who never expected these boons and got used to them, if at all, looking over their shoulder for the copyright police. Scholars who expect to find the very best literature online, harmlessly cohabiting with crap, are replacing scholars who, despite themselves perhaps, still associate everything online with crap.

Some lazy scholars believe that if something is not free online, then it's not worth reading. This has never been true. However, it's gradually becoming true, and those who want it to become true can accelerate the process. Those who want to live in a world where all peer-reviewed journal literature is free online are themselves growing in numbers and will soon have the power in universities, libraries, learned societies, publishers, funding agencies, and governments to bring it about.

Moreover, as the OA percentage of research literature continues to grow, more users will start to act (with or without justification) as if all research literature worth reading is already OA. As this practice spreads, it will function as one more incentive for authors and publishers to make their work OA.

In short, generational change is on the side of OA.

But even the passage of time without generational change is on the side of OA. Time itself has reduced panic and panic-induced misunderstandings of OA. Everyone is getting used to the idea that OA literature can be copyrighted, the idea that OA literature can be peer-reviewed, the idea that the expenses for producing OA literature can be recovered, and the idea that OA and TA literature can co-exist (even for the same work). Surprisingly, many of the early obstacles to OA can be traced to the fact that many seasoned academics just couldn't grasp these ideas. The problem was not incoherent ideas or stupid people - though both hypotheses circulated
widely - but panic, unfamiliarity, and the violation of unquestioned assumptions. For some stakeholders, clear explanations, repetition, or experience with working examples solved the problem. But for others it just took time.

When *Nature* broke the story in January 2007 that the Association of American Publishers, American Chemical Society, Elsevier, and Wiley, had hired Eric Dezenhall, “the pit-bull of PR,” to help their lobbying campaign against OA, the resulting controversy brought OA to the attention of many academics for the first time. Unlike earlier waves of newcomers, for example, after Congress asked the NIH to develop an OA policy in July 2004, this wave typically got it right the first time. “Of course OA is compatible with peer review.” “Of course there are no copyright problems if the copyright-holder consents.” “Of course the public deserves OA to publicly-funded research.” “Of course the argument that OA is a kind of censorship is Orwellian doublespeak.”

When newcomers got OA wrong in the past, sometimes they had been misled by an explicit error published somewhere, perhaps by another newcomer. But most of the time they just made unconscious assumptions based on incomplete information and old models. This is the shock of the new at work. If OA uses the Internet, then it must bypass peer review. (Right?) If OA articles can be copied ad lib, then there must be copyright problems. (Right?) If OA is free of charge for end-users, then its proponents must be claiming that it costs nothing to produce and it must be impossible to recover the costs. (Right?) These conclusions, of course, were uninformed leaps. Many who understood the conventional model (priced, printed, peer-reviewed, copyright-protected) saw a proposal for something different and didn't know how many parameters of the old paradigm the new proposal wanted to tweak. Their hasty and incorrect surmise: all of them. It was a classic case of seeing black and white before seeing shades of gray.

Suddenly, everything good about the present system had to be defended, as if it were under attack. A lot of energy was wasted defending peer review, when it was never under attack. A lot of energy was wasted defending copyright - or celebrating its demise - when it was never under attack. (More precisely, copyright was under attack from other directions, but OA was compatible with unrevised, unbalanced, unreconstructed copyright.) The debate about OA often drifted toward the larger debate about what was good and bad, or functional and dysfunctional, in the present system of scholarly communication overall. This was valuable, but mixing narrow OA issues with broader ones created false impressions about what OA really was, how compatible it was with good features of the present system, and how easy it was to implement.

The OA debates still waste a lot of energy talking about peer review and copyright. The shock of the new hasn't fully worn off; it's wearing off gradually. OA advocates, growing in numbers and effectiveness, can't keep the idea from being distorted or misunderstood. But they have kept it from being distorted or misunderstood as much as it would have been otherwise.

As time passes, we see a steady rise in the proportion of correct to incorrect formulations of OA in the widely-read discussions. When people encounter a fragmentary version of the idea for the first time today, their guesswork to flesh it out is guided by a much more reliable range of clues than just a few years ago. If they take the time to run an online search, the chances that they'll find good information before someone else's guesswork are nearly 100%.

It's tempting to focus on the elegance of OA as a solution to serious problems and overlook the need for the sheer passage of time to overcome the shock of the new. Even if we acknowledge the need for cultural change in the transition to OA - far more critical than technological change - it's easy to underestimate the cultural barriers and the time required to work through them. Yes, OA is compatible with peer review, copyright, profit, print, prestige, and preservation. But that doesn't quiet resistance when those facts about it are precisely the ones hidden by false assumptions caused by the shock of the new.
I’m not saying that all resistance to OA is, or was, based on a misunderstanding of the idea itself. But much past resistance was based on misunderstanding; that portion is in decline; and that decline is largely due to the passage of time and the rise in mere familiarity with a new idea.

The changes wrought by time point up a sad irony in the 15 year history of OA. Nobody is surprised when cultural inertia slows the adoption of radical ideas. But cultural inertia slowed the adoption of OA by leading many people to mistake it for a more radical idea than it actually is.

I know that this account of trends would not be complete without those that work against OA. But there aren’t many. I’ve mentioned the improvement in ebook readers, which may interfere with the ways that OA books increase sales for print editions. Here are two more.

29. Researchers themselves control the rate of progress toward OA, but after all these years most of them are still oblivious to its existence and benefits. As I’ve noted above, there is a trend toward greater familiarity and understanding. But there is also a longstanding counter-trend of impatience with anything that distracts attention from research. This preoccupation is generally admirable and makes researchers good at what they do. But even from the narrow perspective of what advances research, it is having perverse consequences and limiting the audience, impact, and utility of the research on which scholars are so single-mindedly focused.

30. Some publishers have opposed OA from the beginning and sometimes their opposition has been fierce. But some who opposed it apparently saw it as a utopian fantasy of naïve academics that would never be embraced by serious researchers, let alone by serious institutions like universities, libraries, foundations, and government agencies. Publishers in the second camp, who thought OA would be alarming if it caught on, but then hit the snooze button, are now hearing the alarm. While some publishers actively support OA, or experiment with it in good faith, those that oppose it are getting their act together and spending serious money to lobby against government OA policies. In money and person-power, their lobbying forces in Washington and Brussels vastly exceed our own. All we have going for us are good arguments and good trends.

Acknowledgment
An earlier version of this article appeared in the SPARC Open Access Newsletter, May 2, 2007. http://www.earlham.edu/~peters/fox/newsletter/05-02-07.htm#trends
CTWatch Quarterly is a publication of
Sponsored by

© 2007 NCSA/University of Illinois Board of Trustees
© 2007 The Regents of the University of California

Any opinions expressed in this publication belong to their respective authors and are not necessarily shared by the sponsoring institutions or the National Science Foundation (NSF). Any trademarks or trade names, registered or otherwise, that appear in this publication are the property of their respective owners and do not represent endorsement by the editors, publishers, sponsoring institutions or agencies of CTWatch.