An analysis of the present system of scientific publishing: what's wrong and where to go from here

DOV GREENBAUM, JOANNA LIM and MARK GERSTEIN
Yale University, New Haven, CT, USA

The internet has produced an unprecedented opportunity to provide free and unhindered access to the wealth of scientific information, the volume of which continues to grow at a furious pace. The current balkanised system of individual journals limits possibilities for powerful search tools and for an integrated repository of the whole body of scientific literature. This paper reviews the current publishing environment, commenting on its strong and weak points (for example peer review, which is both a strength and a weakness). It attempts to find a viable solution to the current issues that plague STM (scientific, technical, and medical) publishing in the introduction of a centralised repository of scientific literature. Related issues such as the question of long term archiving and the justified fears of STM publishers of becoming obsolete are also discussed.

As recounted by Jean-Claude Guédon in *In Oldenburg's Long Shadow*, scholarly journals were initially founded in order to preclude intellectual property disputes. The *Philosophical Transactions of the Royal Society of London*, first published in 1665, was to be a register of scientific ideas, and *the* arbiter of what was science; as a secondary goal, it would also disseminate scientific ideas.[1] Henry Oldenburg, inspired by Francis Bacon's *Novum Organum*, was the pioneer behind the journal, and behind the idea of peer review; Oldenburg would have articles sent to experts for review before including them in the *Philosophical Transactions*.[2] A hundred years later the concept of peer review was cemented as a requirement for publication, when the editorial process of the journal was taken over by the Royal Society.[3] Subsequently, the notions of wide dissemination and peer review have become general hallmarks of scientific journal publishing. In addition, there are other objectives of scholarly journals, including the creation of archives for scientific data, of a system to prevent plagiarism of others' work, and of a sort of currency for scientists, demarcating their level of prestige as a function of the number and quality of articles published.[4] But journals as we know them are becoming less important for dissemination of scientific information; they are used more as a currency measuring scientific merit. Improved vehicles of communication, better able to conform to the diverse types of collaboration that are the norm in present day scientific research, are required.[5] In its present form the process of publishing scientific articles is in general slow, inefficient, costly, and sometimes even a hindrance to research and to the flow of information.[6] In addition, the traditional printed paper is 'difficult to produce, difficult to distribute, difficult to archive and difficult to duplicate'.[7]

As long ago as 1945, Vannevar Bush was of the opinion that 'methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purposes'.[8] Although there was then no practical alternative to print publication, the internet now presents an opportunity to reshape the whole scientific publication process. Still, the internet is only starting to make inroads into methods of transmitting research, and up to now much of the evolution of scientific information dissemination has resulted from a haphazard and undirected progression of research methodologies. The web gives researchers the ability to present much of their work in forums other than journals, for instance on private websites, or in the form of preprints, databases, newsletters, reports, working papers, theses, or online conference proceedings. While not peer reviewed, this 'grey' literature is gaining validity and importance in research as a source of scientific information.[9] For example, the US departments of Energy and Defense, as well as other government agencies, currently have well over a hundred thousand non-peer reviewed scientific and technical reports integrated into a central repository, the GrayLit Network.[10]

Nevertheless, to achieve a true paradigm shift in scientific publishing, we need a directed evolutionary event (cf. Ann Okerson's position[11]), a total and globally unified revamping of the system from the ground up. Although two thirds of all journals already publish online,[12] there are many issues with the present system of peer review in academic journals, problems that cannot be solved by simply making pdf copies of articles available online: 'An electronic document is not [simply] the electronic version of a traditional paper document … [Rather it is] a

 DOI 10.1179/030801803225005238
Published by Maney for the Institute of Materials, Minerals and Mining

document comprising a variety of different types of information presentations that are brought together by an author in order to present a comprehensive scientific argument.'[13]

This paper will examine some of the issues with the present system of scientific publication – such as rising costs, inadequate peer review, and slow dissemination of information – and present a possible alternative to this situation. The discussion itself is not novel: many groups have already attempted to tackle the issue and reform the world of scientific publishing and data dissemination.

Issues with the existing system

Formats

With the advent of high throughput experimental methodologies, molecular biology has become, like many other sciences, data intensive.[15] Consequently, experimental results more often than not do not fit the rigid guidelines of journal formats, and very often important data tables, if they are included at all, are relegated to online supplementary appendices or associated websites, often available only to paying subscribers. Moreover, in their present state, journal articles are not easily parsed for data mining because of the lack of any standardised formatting or ontology.[16] In addition the universal format presently used in STM journals (abstract, introduction, methods, results, discussion, conclusion) may not be appropriate for the presentation of web tools or databases and future research methods and results.

Grey information

Many laboratories choose to present their data on their own websites, providing access to raw, unverified experimental data, unconnected with any particular publication. This information is a rich source of cutting edge data, and its growing use as a research tool blurs the boundaries between formal and informal publication.[17] Such online databases are slowly encroaching on the journals' position as disseminators of information. Still, in contrast to journal articles that are centrally indexed, it becomes very difficult to keep track of and locate new results that are published in these 'grey' forums. While before this explosion in the volume of data, researchers could easily contact authors for additional individual datasets, with the advent of bioinformatics and the need to sift and analyse multiple huge datasets, all the data must be easily accessible in real time.[18]

Peer review

The peer review process, which is supposed to provide verification for the information found in scientific journals, and thus differentiate journal based information from grey information, is under attack. Both *Science* and *Nature* have recently taken flack for publishing questionable material.[19] For the most part, research scientists and their students make up the cadre of peer reviewers, and with increasing pressure for these scientists to produce, there is less time and incentive to review articles thoroughly, and a greater chance of bad science slipping through the cracks.

Costs of acquiring journal articles

Journals are also becoming less available owing to high costs. Journal prices are rising, significantly faster than inflation, and many are no longer within the price range of the average university library. The US Association of Research Libraries claims that the price of journal subscriptions skyrocketed 207% from 1986 to 1999.[20] In conjunction with budgetary cutbacks, this is forcing many libraries to cancel subscriptions.[21] As a result, most refereed journals are not available to the average researcher.[22] The irony of this situation is that the universities are funding research, yet they often cannot afford to buy the results back from the journals. Even the electronic versions of journals, which were supposed to be cheaper than print subscriptions, are no more affordable.[23] (The high prices here have been attributed to the cost of customer support, as well as to the continued fixed costs of editing.[24]) However even with all the cutbacks and cancellations, STM publishing has been the fastest growing media subsector for the last fifteen years.[25]

Even with this incredible growth, journal publishing houses that maintain high prices may be pricing themselves out of the market, and as such should also be interested in reform. Recent research has shown that researchers preferentially read and cite articles that are made freely, or at least easily, available. Many are not willing to pay for expensive journals, nor are they willing to seek out printed copies of journals when they can access other journals online, effortlessly and for free.[26]

Journals ought to be free to the scientific community. Still, given that the PubMed/Medline database was only made freely available to the public in 1997,[27] the concept of providing totally free access to all information may be somewhat premature. Even so, there are many groups presently working towards providing free access to scientific journals. These include PubMed Central, BioOne, the Public Library of Science, and the Budapest Open Access Initiative.[28]

Too much information

The number of articles published annually has been doubling every decade or so for the last two hundred years;[29] there are, at present, approximately twenty thousand refereed STM journals producing in excess of two million articles each year.[30] Not only can researchers not possibly keep up with this deluge of data, surveys have shown that they do not even attempt to[31] – in fact it has been found that they do not *want* to read the seemingly inexhaustible literature.[32] With this growing number of articles, it is becoming increasingly difficult to sift the literature effectively for the required information. Even with the growing desire, and the computing ability, to mine the

literature for additional information,[33] the incredible lack of uniformity within the literature in terms of ontologies and formats makes this method of research difficult to conduct.

Speed and biases in information transmission

The process of getting an article from submission to publication, especially in competitive and fast moving fields, is much too slow. With the fear of getting scooped by their competitors, scientists often publish incomplete or partial research results so that they can stake their claim to potentially valuable research. Additionally, there is a general concern that too much power is held by the editors of journals and peer reviewers, so that their biases have the potential to prevent the publication of important, novel, or avant garde results.

An alternative

While only some of the concerns with the present system have been aired, it should be clear that Vannevar Bush's assessment (see above), voiced over half a century ago, is all the more pertinent today. What is needed is a total overhaul of the publishing system. Below, we present an outline of a possible future system of scientific dissemination. Following the presentation of a succinct framework, we flesh out some of the particulars and discuss some additional issues that need to be tackled.

Outline

We are not advocating a system similar to the present scheme where journals in print are also available online, but rather a total and unmitigated shift from print to online publication. We envisage the following multitiered system. After completing a project, the researcher submits a paper to a web based journal along with a standard (and reasonable) submission fee to cover the initial costs of editing. The journal's editorial board decides whether the project and the paper fit their basic criteria for publication, and if so the paper is uploaded to a limited access website. Other researchers in the field who have registered for access to this site, and have expressed interest in the subject matter, are notified automatically via email of the submission. Over the course of some flexible period of time, depending on the subject matter, other researchers can log in and evaluate the paper, posting their comments and suggestions; this online discussion is moderated by an editor assigned to the paper. Once this review period ends, the editor can decide, based on the comments, whether to accept the paper as is, request changes and send it back for another round of review, or reject it. Each draft of the article throughout the review process is saved and contains a unique identifier. On acceptance, the author is charged an additional fee to cover the costs of publication and archiving. The final paper, which should be immutable and authenticable,[34] may be uploaded to the journal's website, but must also be uploaded to a freely accessible archival website, providing unlimited access to anyone.

The journal

Historically, journals have played many important and essential roles in the dissemination of information. In their simplest form they are archives of information; one can dig up ancient copies of journals in any well stocked library to find data. In the pre-internet era they were the easiest way of distributing new information to the broadest possible audience; anyone who was interested in learning the most recent accomplishments in their field could flip through a copy of the appropriate journal and read a description of the research. Usually, the research was (and for the most part still is) presented in a common format including abstract, introduction, sections on methods and results, a discussion, conclusions, and references; readers knew where to look in the article for the information they needed.

Journals act as gatekeepers to the scientific archive, keeping out undeserving or plagiarised research. The fact that an article appears in a journal indicates that it has gone through some sort of peer review that has provided some kind of validation for the purpose, necessity, and results of the research. The fixed costs of publishing a journal are thought to be a barrier to entry for journals that have not reached a certain level of public acceptance or academic stature. Journals also provide some sort of qualitative comparative measure to the research. The more prestigious the journal, the more important and conclusive the research is thought to be.

With the prospect of creating a long term digital archive of all scientific data (as opposed to the present paper archive), it would not make economic sense for individual journals to maintain their own archives (see below for a discussion of the issues of maintaining a digital archive). Instead we envisage a much smaller yet important role for journals in our solution. As described, journals presently perform both a repository and an information service function.[35] In our proposal they would retain a portion of the service function, and spin off their repository functions. That is, they would retain only their most important and irreplaceable role as editors and facilitators of peer review (although some have claimed that the editorial process in fact diminishes the value of an article[36]). Rather than having each journal maintain copies of its articles, a system must be developed to maintain an easily accessible archive that would promote interoperability and so allow for large scale mining of the scientific literature. Journals should, though, maintain their own banners across the tops of their own articles in the archive, since journal names are somewhat indicative of the quality of an article.

We assume that many journals will decide to continue publishing online; still, there should be a

universally accepted framework that would demand that articles be deposited in an archive soon, if not immediately, after publication. Some journals might also choose to continue to publish paper versions of online articles, possibly for a persistent Luddite population. Journals might also publish smaller, single page, abstractlike versions of their online content in print journals: for example the *FASEB Journal* publishes short summary versions in print but longer articles online.[37] Nevertheless, research articles ought to be provided to the scientific public for free.

Journals claim that providing free and unlimited access through an independent provider to online articles will deplete an economically important source of revenue, could lead to loss of quality control and abuse of content, and will put too much control within a centralised organisation, which they set against what they claim is the more stable system of hundreds of journals providing separate access.[38] In addition, it is said, the transfer and duplication of information from the journal into the archive could corrupt the data.[39] Journals claim that profits can be maintained if instead of providing information straightaway and for free to the public, they wait six months during which they can charge for access, after which time articles will be provided free of charge on the journals' own websites, where they can control and monitor access.

We propose a more research friendly profitmaking approach: To prevent loss of profits, journals will retool their revenue mechanisms. One possible solution is to charge authors for the costs of editing. Given the generally inelastic demand for publishing articles, journals should be able to charge enough to stay profitable. In any case, authors will pass these costs on to their funding agencies, and this should therefore not limit the ability of a researcher to publish. Moreover, given that the economic system of publishing tends to favour those who pay, a system whereby the author is paying is a system that will reflect the goals of the author, i.e. broad dissemination.[40] Additionally, by not maintaining any archival functions, journals will have no fear that copy submitted to the archive will be corrupted through reproduction; instead, journals should submit their copy immediately to the archive.

Peer review

The peer review process, existing in its present form only since the Second World War,[41] has been coming under fire for many of its failings[42] for quite some time. Some of the issues with the peer review process include: falsified data getting past reviewers;[43] reviewers holding up the review process, either out of spite or while they themselves publish similar results;[44] plagiarism;[45] sharing of confidential data;[46] slow or deficient work by researchers overwhelmed by their reviewing responsibilities; anonymity of the review process leading to unaccountability of reviewers[47] (but contrast this with Steven Harnad's comments elsewhere[48]); lack of credit given to the unpaid labour force of reviewers; reviewers having too much power over the dissemination of scientific information, which may be affected by their biases; cost – anywhere between five hundred and a thousand US dollars per article.[49] However, with all its faults, the peer review process is integral to scientific research. It provides assurance to authors, publisher, and the public that the work submitted is of acceptable quality. At the very least, it provides a process whereby work can be improved by the incorporation of outside ideas.

The translation of scientific data from paper onto the internet could help democratise the review process, making it more efficient and more discriminating. The present peer review process requires the editors of a journal to select reviewers on the basis of their perceived fields of expertise, contact these reviewers, and request them to review a paper. Often reviewers are slow to respond and may not have the time or desire to review. We propose a system whereby reviewers would be notified automatically via email when a new paper was submitted in their field. Moreover, in addition to the present incentives to review (for example the desire to keep bad science out of the field, or a feeling of academic responsibility), journals could provide financial inducements to review in the form of credit towards publication of the reviewer's next piece. In addition to providing an incentive, this method would also result in a situation whereby the more prestigious journals (in which more people would like to publish and whose credit would be more highly appreciated) would have more people reviewing submissions, in essence providing substantiation for work in better journals.

Addressing the issue of anonymity, reviewers would have to register to access presubmission pieces, and their access to the papers would be logged, thus allowing for a paper trail in a case where a reviewer was suspected of stealing information. Moreover, authors of papers would no longer be held up by procrastination of individual reviewers. The review process would take a finite period of time, after which the editor assigned to the paper would review the comments. Of course there would be cases where the editor might feel that a paper was not garnering enough attention for a comprehensive review. At this point an intervention could be made to assign reviewers for the piece or reject it outright. Still, as the success of sites such as eopinions.com shows, people are more than willing to give their opinion on anything at all. This system would also allow authors to collect a wide range of comments on their work from a significantly larger audience: reviewers would not be limited to a small cadre of researchers selected by the journal, but rather anyone in the field could register and offer their opinion.

Reviewers would also be able to increase their 'street cred', and the credit towards future publishing in the journal. Akin to the system already in place on amazon.com, readers of reviewers' comments would be able to evaluate the comments and note

whether or not they were helpful, helping to highlight the important comments and weed out the inane remarks often seen when the reviewer does not truly understand a paper. A reviewer who consistently presented strong comments would receive more credit for their review, and bad reviewers could be barred from the forum, providing an incentive for people to put in well thought out comments. The review process could also be simplified by requiring reviewers to stick to a specific syntax and format, answering a list of directed questions. Given the automation of the system there could be significant cost savings in this step of the publication process.

Finally, to prevent frivolous submissions from overwhelming reviewers, there could be some sort of automated check to determine an author's publication record, institutional affiliation, research grant status, and other background information that could act as an automatic first level of discrimination to at least determine whether a paper was of 'refereeable quality'. New authors could resort to alternative paths of entry, for example referrals from established scientists.[50]

Although it might be argued that such a peer reviewing system is faulty in that it relies on fellow authors volunteering to review articles instead of journals requesting experts in that field, this system rewards reviewers by giving them the opportunity to become known to the journal, whether or not they are already well known for their research accomplishments. This system of peer review allows for a greater breadth of response to each article, allowing feedback from all kinds of perspectives, and possibly even creating the basis for future collaborations.

Format

One of the main strengths of our framework is the possibility of creating a homogeneous body of scientific literature that will allow for thorough searching and data mining.[51] To this end it is imperative that a set of universal standards for formatting scientific articles be established. In addition, it is also important to create a standardised language to describe the information contained within the articles.[52]

With all of the text of each article available online, large scale literature searches, similar to database searches, will allow users to integrate and incorporate disparate information for analysis. Large scale global searches will allow users to pick out keywords from the entire body of scientific literature. To facilitate more powerful searches, we envisage a standardisation of formats and keywords – similar to the MESH terms in the NCBI's PubMed system.[53]

Within the potentially unlimited extent of cyberspace, articles will expand and provide not only more information, but more information in a more efficient manner. One potential way of setting an internet journal format would be to have the data presented in multiple 'layers' (the concept of layering has been proposed by Paul Ginsparg, founder of the arXiv physics preprint archive[54]). Articles are accessed by a wide variety of readers (experts, non-experts, casual readers), all of whom have different information requirements. These different requirements could be satisfied by layering: for example, the first layer might include the primary data, the information on which the article is based, with little or no textual elaboration, allowing experts to quickly scan and retrieve data; a second layer could provide more information on materials and methodology; a third layer might resemble a short article providing in succinct form the data, methods, and some discussion and conclusions; finally, a fourth layer could include background information helpful to the uninitiated reader, including an extended introduction, methods section, discussion, conclusions, and supplementary materials. While currently space limitations force authors either to leave out information or to publish it as supplementary material, a wholly online format would allow researchers to incorporate all their data and textual information into the article.

In addition to the extra space, an online format would allow authors and editors to integrate hyperlinks into papers, providing readers with access to further information on the subject at hand within the article itself, but also to other sites, grey information, articles, and, importantly, errata.[55] Furthermore, a list of citations as well as links to derivative works could be continuously and dynamically updated.[56] Readers should have the opportunity to post comments on individual articles, organically developing what on paper would have been an inert document. Present paper based articles have static tables and figures. An online literature allows for interactive vibrant and informative figures, where users can zoom in on parts of particular interest or rotate three-dimensional structures. Additionally, the internet allows for dynamic updating of tables that could also be available for bulk download.[57]

All new ideas take time to be accepted, and some scientists might baulk at the idea of 'layering' their articles, but in the end such formats will be to their own benefit when they access others' work. Such formatting also requires an integrity of work, laying bare all research and results for scrutiny, allowing for no ambiguity. Some authors might also be averse to careful structuring of their articles to conform to some seemingly arbitrary standards. However, computers are much better able to parse and handle structured and well designed information; an author's minor efforts will go a long way in providing significantly more functionality. In the long run, it is in the interests of authors when their work can be communicated more widely.[58]

Archives

With journals retaining only the editing and peer review aspects of their original functions, the issue of presenting and archiving data needs to be addressed.

Will there be one central archive, i.e. a 'megacentre' for the whole body of scientific knowledge akin to the PubMed abstract archive, or a system of federated archival libraries, like the BioMed Archives Consortium, Project Muse, Highwire Press, or CrossRef?[59] Will it be controlled privately (as is the case now with journals) or publicly? Should the archive include only peer reviewed information, or grey literature as well?

One commonly used example of a central archive that has done exceptionally well is the physics preprint archive. In 1991 Paul Ginsparg launched arXiv.org, a groundbreaking archive of physics preprints (formerly operating out of the US Department of Energy's Los Alamos National Laboratory, and now hosted by Cornell University). The archive, which receives tens of thousands of papers annually, functions to provide rapid and efficient dissemination of articles as soon as they are ready, even before they are published.[60]

The international nature of scientific research would seem to make the concept of a centralised database politically unlikely,[61] however central archives have their proponents. Matt Cockerill of BioMed Central claims that it is imperative that data be stored within a central location for efficient searches to be possible. Additionally, a central repository could provide for a simple and operator friendly interface; fears of lost data could be limited by using multiple mirror sites.[62] Additionally, the costs of maintaining any long term digital archive favour a centralised archive over some balkanised system of small independent and non-interoperable systems.

CrossRef, which aims to include not only journals but also grey information such as books, reference works, and databases, claims that the degree of interoperability that a central archive could achieve might just as well be attained through the use of consensus standards, at the same time avoiding many of the limitations inherent in a centralised system.[63] SPARC (the Scholarly Publishing and Academic Resources Coalition) is another example of a decentralised group, composed of universities that publish and archive an aggregate of leading research journals at prices that are 'sensitive to the interests' of publishers and subscribers accessing journals.[64]

A digital archive, in whatever final form it might take, would have many advantages over the paper archives in our libraries. For example, in contrast to present day libraries that cannot feasibly curate their physical stacks to remove wrong, misleading, or outdated information, the dynamic nature of an online archive would allow for the sequestering and possible removal of bad data. Moreover, similar to present online databases, the archive would be organic, growing and evolving on the basis of the present and future needs of the research community. The role of present day libraries would change from being physical repositories of information, to being gateways of information, i.e. advanced search systems and centres of expertise on how best to access the different levels of the chain of information in the archives.[65]

Future issues

Aside from the question of *who* should do the archiving, is the potentially more important issue of *how* to archive data. Given the rate of technological change, it is highly unlikely that any system implemented today will be anything like whatever system is used to archive data in a couple of decades; media decays, standards change, software and the machines that run it become obsolete and lost. The US Census information from 1960, originally stored on digital tapes, in addition to hundreds of other reels of tape from multiple government departments, has already become obsolete.[66] Any long term archive will need significant recurring investment to keep it operational.

Long term archiving requires that data should be well maintained, and easily accessible, displayed, and recreated. Moreover, one cannot simply print out hard copies of the archive, as this would defeat the purpose of going digital, and in any case much of the information could anyhow not be meaningfully displayed on paper (hyperlinks, for example).[67] The issue of data archiving is complex and mostly beyond the scope of this paper, but we will now present, succinctly, some of the options.

It is imperative that whatever system is used should allow for easy migration of data from one system to another, bearing in mind the exponential growth in archived data. The ability to transfer data, dynamically recreating the entire archive using the new technology, is critical in light of the fact that many of the media used to preserve digital data are unstable and degrade without active preservation, in contrast to paper archives. Even within the lifetimes of current technologies, the storage media on which the digital information is stored have finite lives; data will inevitably degrade or be corrupted.[68] Additionally, as the archive grows and technology changes, newer, cheaper, and better storage media will become available.

What is needed is a long term solution, one that does not call for heroic efforts or continual interventions to maintain it.[69] One idea would be to use some sort of semistructured representation of the data, which with each digital object would include basic information such as the attributes of the data (structure and physical context, information on the organisation and display of the information).[70] Platform independent technologies such as XML[71] could be used to describe the data and to provide a simple and flexible format, and in consequence to give the data a longer lifetime.[72]

A similar idea, since digital archives are inherently software dependent, would be to keep the original software and, as technology changes, to run it under emulation on future systems; present systems also have a short physical life and as such cannot be maintained to run the software.[73] Alternatively, instead of creating emulators of outdated software,

software could be designed to run on some 'universal virtual computer' that would be standardised and maintained.[74]

In addition to the issues of data storage, there is a more basic issue of what deserves to be stored. As stated above, there are already archives focused on informal publications, the so called grey literature. But how much of the grey literature deserves to be archived? Is all scientific data pertinent to the future and worth the cost of storage; for example, will the data play an important role in deciding who is deserving of scientific accolades and/or intellectual property rights for results? And even within the so called formal literature, of peer reviewed articles, how many states of an article deserve to be preserved (prereviewed versions, drafts in progress, etc.), and should they, like the definitive form, be preserved indefinitely?

Finally, another issue that has to be dealt with before the establishment of an archive is that of ownership of published articles, and of the underlying results. Although we assume that results of scientific research, especially of work funded by governmental grants, are intended for the public domain, this is often not the case. As a result of the 1980 Bayh-Dole Act,[75] US universities have been encouraged to protect and profit from their research by exercising intellectual property rights. One current area where the idea of ownership of scientific fact is hotly debated is in relation to databases.[76] With regard to the archive in particular, the issue of who should own the copyright of articles continues to be debated.

The copyrighting of scientific articles, like the patenting of scientific results funded by government funds, has been termed 'public taxation for private privilege'.[77] It goes against the spirit of the law 'to promote the progress of Science and the Useful Arts' by limiting the dissemination of research results. The United States Supreme Court ruled some time ago in the case Universal *v*. Miller that research results cannot be copyrighted. Still, a trend has developed over time for journal publishers to require that authors sign over all copyrights to the journal. Authors acquiesced to this Faustian bargain, in which by handing over copyrights they in return received affirmation that their work would be disseminated and protected in perpetuity.[78]

In 1996, the US Congress, in the National Information Infrastructure Copyright Protection Act, considered expanding the rights of owners of copyrighted articles at the expense of the academic community.[79] More recently it has been proposed that authors should maintain their copyright, perhaps through new legislation requiring that authors of government funded research do so;[80] there has also been a grassroots campaign to encourage authors not to sign over copyright,[81] and in cases where they are forced to, to boycott journals.[82] Alternatively, it has been suggested that journals maintain copyright only for a very limited time, after which rights are transferred over to a central journal repository.[83] With the growing trend towards more collaborative work in scientific research, it has in practice become significantly harder even to determine who owns copyrights to what.[84]

Notes and literature cited

1. J.-C. GUÉDON: *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*; 2001, Annapolis Junction, MD, Association of Research Libraries (www.arl.org/arl/proceedings/138/guedon.html).
2. E. R. WERTMAN: 'Electronic preprint distribution: a case study of physicists and chemists at the University of Maryland', MSc thesis, Virginia Polytechnic Institute and State University, VI, USA, 1999 (scholar.lib.vt.edu/theses/available/etd-042499-103003).
3. R. SPIER: 'The history of the peer-review process', *Trends in Biotechnology*, 2002, **20**, 357–358.
4. C. TENOPIR and D. W. KING: 'Lessons for the future of journals', *Nature*, 2001, **413**, 672–674.
5. A. M. ODLYZKO: in *Access to Publicly Financed Research: The Global Research Village III*, (ed. S. Wouters), 273–278; 2000, Amsterdam, NIWI.
6. A. DE KEMP: in *The Impact of Electronic Publishing on the Academic Community*, (ed. I. Butterworth), 4–9; 1998, London, Portland Press.
7. P. GINSPARG: 'Creating a global knowledge network', in Proc. Symp. on Electronic Scientific, Technical, and Medical Journal Publishing and its Implications, Washington, DC, 2003, National Academies Committee on Science, Engineering, and Public Policy.
8. V. BUSH: 'As we may think', *Atlantic Monthly*, 1945, **176**, 101–108.
9. See 'Grey literature: an annotated bibliography', prepared by the US Association of College and Research Libraries' STS Subject & Bibliographic Access Committee and available online at personal.ecu.edu/cooninb/greyliterature.htm.
10. W. WARNICK: 'Tailoring access to the source: preprints, grey literature and journal articles', 3 May 2001, www.nature.com/nature/debates/e-access/articles/warnick.html.
11. A. OKERSON: 'What price ''free''?', 5 April 2001, www.nature.com/nature/debates/e-access/articles/okerson.html
12. 'Great expectations', *Nature Neuroscience*, 2001, **4**, 1151.
13. J. KIRCZ: 'New practices for electronic publishing: how to maintain quality and guarantee integrity', in Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/~icsuinfo/kirczppr.htm).
14. For information on the various initiatives, see A. M. F. BUCK, R. C. COLES and B. COLES: 'Scholars' Forum: a new model for scholarly communication', library.caltech.edu/publications/ScholarsForum/default.htm; www.arl.org/sparc; and S. K. BAKER *et al.*: 'Principles for emerging systems of scholarly publishing', May 2000, www.arl.org/scomm/tempe.html.
15. See J. RUMBLE: 'Publication and use of large data sets', in Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/~icsuinfo/rumbleppr.htm).

16. M. J. GERSTEIN and J. JUNKER: 'Blurring the boundaries between scientific ''papers'' and biological databases', 7 May 2001, www.nature.com/nature/debates/e-access/articles/gernstein.html.
17. A. M. CORREIA and M. C. NETO: 'The role of eprint archives in the access to and dissemination of scientific gray literature: LIZA – a case study by National Library of Portugal', in Proc. Int. Workshop on Electronic Media in Mathematics, Coimbra, Portugal, September 2001, Departamento de Matemática da Universidade de Coimbra (www.isegi.unl.pt/ensino/docentes/acorreia/preprint/EMM.pdf).
18. N. M. LUSCOMBE, D. GREENBAUM and M. GERSTEIN: 'What is bioinformatics? A proposed definition and overview of the field', *Methods of Information in Medicine*, 2001, **40**, 346–358; D. GREENBAUM, N. M. LUSCOMBE, R. JANSEN, J. QIAN and M. GERSTEIN: 'Interrelating different types of genomic data, from proteome to secretome: 'oming in on function', *Genome Research*, 2001, **11**, 1463–1468.
19. D. ADAM and J. KNIGHT: 'Journals under pressure: publish, and be damned …', *Nature*, 2002, **419**, 772–776.
20. D. SHULENBURGER: 'Principles for a new system of publishing for science', in Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/∼icsuinfo/shulenbergerppr.htm); R. SMITH: 'Electronic publishing in science', *British Medical Journal*, 2001, **322**, 627–629.
21. Proc. Symp. on Electronic Scientific, Technical, and Medical Journal Publishing and its Implications, Washington, DC, 2003, National Academies Committee on Science, Engineering, and Public Policy.
22. S. HARNAD: 'The self-archiving initiative', *Nature*, 2001, **410**, 1024–1025.
23. A. M. CETTO: 'The contribution of electronic communication to science – has it lived up to its promise?', in Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/∼icsuinfo/cettoppr.htm).
24. 'Costs of publication', in Proc. Symp. on Electronic Scientific, Technical, and Medical Journal Publishing and its Implications, Washington, DC, 2003, National Academies Committee on Science, Engineering, and Public Policy.
25. P. GOODEN, M. OWEN, S. SIMON and L. SINGLEHURST: 'Scientific publishing: knowledge is power', Morgan Stanley, London, UK, 2002 (www.alpsp.org/morgstan300902.pdf).
26. B.-C. BJÖRK and Z. TURK: 'How scientists retrieve publications: an empirical study of how the internet is overtaking paper media', *Journal of Electronic Publishing*, 2000, **6**, www.press.umich.edu/jep/06-02/bjork.html; B.-C. BJÖRK and Z. TURK: 'A survey on the impact of the internet on scientific publishing in construction IT and construction management', *Electronic Journal of Information Technology in Construction*, 2000, **5**, 73–88, www.itcon.org/2000/5.
27. See the 26 June 1997 press release, 'Free MEDLINE', available online at www.nlm.nih.gov/news/press_releases/free_medline.html.
28. See www.pubmedcentral.nih.gov; R. J. ROBERTS: 'PubMed Central: the GenBank of the published literature', *Proceedings of the National Academy of Sciences*, 2001, **98**, 381–382; www.bioone.org; www.publiclibraryofscience.org; and J. E. TILL: 'Success factors for open access publishing', *Journal of Medical Internet Research*, 2003, **5**, e1, www.jmir.org/2003/1/e1/index.htm.
29. A. M. ODLYZKO: 'Tragic loss or good riddance? The impending demise of traditional scholarly journals', *International Journal of Human–Computer Studies*, 1995, **42**, 71–122.
30. S. HARNAD: 'The self-archiving initiative' (see Note 22).
31. C. TENOPIR and D. W. KING: 'Designing electronic journals with 30 years of lessons from print', *Journal of Electronic Publishing*, 1998, **4**, www.press.umich.edu/jep/04-02/king.html.
32. H. E. ROOSENDAAL, P. A. T. M. GEURTS and P. VEN DER VET: 'Higher education needs may determine the future of scientific e-publishing', 18 September 2001, www.nature.com/nature/debates/e-access/articles/roosendaal.html.
33. H. YU, V. HATZIVASSILOGLOU, C. FRIEDMAN, A. RZHETSKY and W. J. WILBUR: 'Automatic extraction of gene and protein synonyms from MEDLINE and journal articles', in Proc. AMIA 2002 Symp., San Antonio, TX, USA, November 2002, 919–923; M. KRAUTHAMMER, P. KRA, I. IOSSIFOV, S. M. GOMEZ, G. HRIPCSAK, V. HATZIVASSILOGLOU, C. FRIEDMAN and A. RZHETSKY: 'Of truth and pathways: chasing bits of information through myriads of articles', *Bioinformatics*, 2002, **18**, S249–S257; V. HATZIVASSILOGLOU, P. A. DUBOUE and A. RZHETSKY: 'Disambiguating proteins, genes, and RNA in text: a machine learning approach', *Bioinformatics*, 2001, **17**, S97–S106.
34. M. FRANKEL, R. ELLIOTT, M. BLUME, J.-M. BOURGOIS, B. HUGENHOLTZ, M. G. LINDQUIST, S. MORRIS and E. SANDEWALL: 'Defining and certifying electronic publication in science: a proposal to the international association of STM publishers', July 2000, www.aaas.org/spp/sfrl/projects/epub/define.shtml.
35. R. K. JOHNSON: 'Whither competition?', 15 June 2001, www.nature.com/nature/debates/e-access/articles/johnson.html.
36. P. BROWN: 'What must scientists do to exploit the new environment', in Proc. Symp. on Electronic Scientific, Technical, and Medical Journal Publishing and its Implications, Washington, DC, 2003, National Academies Committee on Science, Engineering, and Public Policy.
37. M. A. KELLER: 'The changing role and form of scientific journals', Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/∼icsuinfo/kellerppr.htm).
38. 'Is a government archive the best answer?', *Science*, 2001, **291**, 2318–2319.
39. I. MELLMAN: 'Setting logical priorities', 5 April 2001, www.nature.com/nature/debates/e-access/articles/mellman.html.
40. P. BOLMAN: 'The effects of open access on commercial publishers', in Proc. Symp. on Electronic Scientific, Technical, and Medical Journal Publishing and its Implications, Washington, DC, 2003, National Academies Committee on Scientific Engineering and Public Policy.
41. F. GODLEE: 'Making reviewers visible', *Journal of the American Medical Association*, 2002, **287**, 2762–2765.
42. F. J. GODLEE and T. JEFFERSON (ed.): *Peer Review in Health Sciences*; 1999, London, BMJ Publishing.
43. E. J. LERNER: 'Fraud shows peer review flaws', *Industrial Physicist*, 2002, **8**, (6), 12–17.

44. 'Bad peer reviewers', *Nature*, 2001, **413**, 93.
45. T. GURA: 'Scientific publishing: peer review, unmasked', *Nature*, 2002, **416**, 258–260.
46. R. DALTON: 'Peers under pressure', *Nature*, 2001, **413**, 102–104.
47. F. GODLEE: 'Making reviewers visible' (see Note 41).
48. S. HARNAD: *Science*, 1980, **208**, 974, 976.
49. W. Y. ARMS: 'Quality control in scholarly publishing on the web', *Journal of Electronic Publishing*, 2001, **8**, www.press.umich.edu/jep/08-01/arms.html.
50. P. GINSPARG: 'Can peer review be better focused?', March 2003, arxiv.org/blurb/pg02pr.html.
51. 'Is a government archive the best answer?' (see Note 38).
52. See 'Task Group on Access to Biological Collection Data (ABCD)', www.bgbm.org/TDWG/CODATA/default.htm and M. J. GERSTEIN and J. JUNKER: 'Blurring the boundaries' (Note 16).
53. J. MCENTYRE and D. J. LIPMAN: 'GenBank – a model community resource?', 5 April 2001, www.nature.com/nature/debates/e-access/articles/lipman.html.
54. P. GINSPARG: 'Creating a global knowledge network' (see Note 7).
55. S. HITCHCOCK, L. CARR, W. HALL, W. HARRIS, S. PROBETS, D. EVANS and D. BRAILSFORD: 'Linking electronic journals: lessons from the Open Journal project', *D-Lib Magazine*, 1998, **4**, (12), www.dlib.org/dlib/december98/12hitchcock.html.
56. D. M. EAGLEMAN and A. O. HOLCOMBE: 'Improving science through online commentary', *Nature*, 2003, **423**, 15.
57. M. S. FRANKEL: 'Seizing the moment: scientists' authorship rights in the digital age', American Association for the Advancement of Science, New York, NY, USA, 2002, www.aaas.org/spp/sfrl/projects/epub/finalreport.pdf.
58. T. BERNERS-LEE and J. HENDLER: 'Scientific publishing on the ''semantic web''', 12 April 2001, www.nature.com/nature/debates/e-access/articles/bernerslee.htm.
59. See http://140.234.1.105 (BioMed Archives Consortium), muse.jhu.edu (Project Muse), highwire.stanford.edu (also M. A. KELLER: 'The changing role and form of scientific journals' (Note 37)), and (on CrossRef) E. PENTZ: 'Evolution and revolution: pragmatism versus dogmatism', 28 August 2001, www.nature.com/nature/debates/e-access/articles/pentz.html.
60. M. SINCELL: 'A man and his archive seek greener pastures', *Science*, 2001, **293**, 419–421.
61. R. LUCE: 'Evolution and scientific literature: towards a decentralized adaptive web', 10 May 2001, www.nature.com/nature/debates/e-access/articles/luce.html.
62. M. COCKERILL: 'Distributed and centralized technologies: complementary tools to build a permanent digital archive', 28 August 2001, www.nature.com/nature/debates/e-access/articles/cockerill.html.
63. See E. PENTZ: 'Evolution and revolution' (Note 59).
64. *Information Today*, 1999, **16**, 32; S. C. MICHALAK: 'The evolution of SPARC', *Serials Review*, 2000, **26**, (1), 10–21.
65. A. KLUGKIST: 'The changing role of the librarian – a virtual library and a real archive?', Proc. Second Joint ICSU Press–UNESCO Expert Conf. on Electronic Publishing in Science, Paris, France, February 2001, ICSU/UNESCO International (users.ox.ac.uk/~icsuinfo/klugkistppr.htm).
66. J. ROTHENBERG: 'Ensuring the longevity of digital documents', *Scientific American*, 1995, **272**, 42–47.
67. J. ROTHENBERG: *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*; 1999, Washington, DC, Council on Library and Information Resources (www.clir.org/pubs/reports/rothenberg/contents.html).
68. M. A. E. DEMENTI: 'Access and archiving as a new paradigm', *Journal of Electronic Publishing*, 1998, **3**, www.press.umich.edu/jep/03-03/dementi.html; K. GUTHRIE: 'Archiving in the digital age: there's a will, but is there a way?', *Educause Review*, 2001, **36**, (6), 56–65 (www.educause.edu/ir/library/pdf/erm0164.pdf).
69. See J. ROTHENBERG: *Avoiding Technological Quicksand* (Note 67).
70. R. MOORE, C. BARU, A. RAJASEKAR, B. LUDASCHER, R. MARCIANO, M. WAN, W. SCHROEDER and A. GUPTA: 'Collection-based persistent digital archives', *D-Lib Magazine*, 2000, **6**, (3), www.dlib.org/dlib/march00/moore/03moore-pt1.html.
71. B. RAGON: 'Castles made of sand: building sustainable digitized collections using XML', *Computers in Libraries*, 2003, **23**, 10–14.
72. F. ACHARD, G. VAYSSEIX and E. BARILLOT: 'XML, bioinformatics and data integration', *Bioinformatics*, 2001, **17**, 115–125.
73. See J. ROTHENBERG: *Avoiding Technological Quicksand* (Note 67).
74. R. LORIE: 'A project on preservation of digital data', *RLG DigiNews*, 2001, **5**, (3), www.rlg.org/preserv/diginews/diginews5-3.html#feature2.
75. PL 96-517; Bayh-Dole Act, 35 USC 200-212, 1980.
76. D. GREENBAUM: 'The database debate: in support of an inequitable solution', *Albany Law Journal of Science & Technology*, 2003, **13**, 431–515.
77. D. KREEGER: *Law and Contemporary Problems*, 1947, **12**, 7414–7445.
78. S. HARNAD and M. HEMUS: 'All or none: no stable hybrid or half-way solutions for launching the learned periodical literature into the post-Gutenberg galaxy', in *The Impact of Electronic Publishing on the Academic Community*, (ed. I. Butterworth), 18–27; 1998, London, Portland Press.
79. S. I. COLBERT and O. R. GRIFFIN: 'The impact of ''fair use'' in the higher education community: a necessary exception?', *Albany Law Review*, 1998, **62**, 437–465.
80. S. BACHRACH, R. S. BERRY, M. BLUME, T. VON FOERSTER, A. FOWLER, P. GINSPARG, S. HELLER, N. KESTNER, A. ODLYZKO, A. OKERSON, R. WIGINGTON and A. MOFFAT: 'Who should own scientific papers?', *Science*, 1998, **281**, 1459–1460; C. MCSHERRY: *Who Owns Academic Work? Battling for Control of Intellectual Property*; 2001, Cambridge, MA, Harvard University Press.
81. L. GUERNSEY: 'Research libraries' newsletter examines profits of journal publishers', *Chronicle of Higher Education*, 1998, 30 October, A29.
82. M. WADMAN: 'Publishers challenged over access to papers', *Nature*, 2001, **410**, 502.
83. See D. SHULENBURGER: 'Principles for a new system' (Note 20).
84. R. DREYFUSS: 'Collaborative research: conflicts on authorship, ownership, and accountability', *Vanderbilt Law Review*, 2000, **53**, 1162–1232.

Dov Greenbaum
Department of Genetics
Yale University
PO Box 208114
New Haven
CT 06520–8114
USA
dov.greenbaum@yale.edu

Dov Greenbaum is a sixth year genetics graduate student in Mark Gerstein's lab at Yale University. He received his bachelor's degree, in biology and economics, from Yeshiva University (New York) in 1998. He has written a number of papers on the subjects of genomics and proteomics. In addition, he has published on the legal issues surrounding databases and copyright.

Joanna Lim
Department of Molecular
Biophysics and Biochemistry
Yale University
PO Box 208114
New Haven
CT 06520–8114
USA
joanna.lim@yale.edu

Joanna Lim is a junior at Yale University, majoring in history and molecular biophysics and biochemistry. She is currently doing bioinformatics and proteomics research in Mark Gerstein's laboratory.

Mark Gerstein
Department of Molecular Biophysics
and Biochemistry
Yale University
PO Box 208114
New Haven
CT 06520–8114
USA
mark.gerstein@yale.edu

Mark Gerstein is codirector of the computational biology and bioinformatics programme and A. L. Williams Associate Professor of Biomedical Informatics at Yale University. He is also an associate professor in the Department of Molecular Biophysics and Biochemistry and holds a joint appointment in the Department of Computer Science. Dr Gerstein studied physics as an undergraduate at Harvard and received his doctorate for work with Cyrus Chothia and Ruth Lynden-Bell at the MRC in Cambridge. Prior to his employment at Yale, he worked in Michael Levitt's lab at Stanford. Dr Gerstein, who has received numerous young investigator awards, has published over a hundred and forty papers thus far on many subjects including structural biology, bioinformatics, genomics, and proteomics. This paper grew out of a contribution to a *Nature* web debate on 'Future e-access to the primary literature'.