The impact of preprint servers and electronic publishing on biomedical research

Gunther Eysenbach

Addresses
University Hospital Heidelberg, Department of Clinical Social Medicine and Public Health, Bergheimer Strasse 58, 69115 Heidelberg, Germany; e-mail: ey@yi.com

Current Opinion in Immunology 2000, **12**:499–503

0952-7915/00/$ – see front matter

Introduction

Scientific communication and scholarly publishing are in transition. The age of printed publications as primary means to communicate research results is ending, being replaced by the era of electronic publishing (also known as e-publishing). This form of publishing has far-reaching consequences not only for how scientists distribute, access, process and digest information but also for how research itself is done and will be evaluated.

The advantages of electronic publishing are immediately evident: research results can be disseminated faster and more cheaply, can be distributed to a wider audience more fairly (it offers equity of access, including the lay public and scientists in developing countries) and authors have virtually no space restrictions, and can therefore include huge datasets or even multimedia data. It is obvious that information is key to research and knowledge production. The famous phrase coined by American sociologist Robert Merton — "Standing on the shoulders of giants" — actually refers to scientists using past work in advancing knowledge. If information is so crucial, certainly faster and cheaper dissemination of and access to electronic information should lead to better research.

However, not all scientists share the enthusiasm of having yet more information at their fingertips, in particular if this seemingly comes at a cost of quality. The problem of 'excessive publication' and information overload in immunology was already decried 20 years ago [1] and since then the number of immunology journals has almost tripled; in addition, scientists now have access to an unprecedented amount of information on the Internet. Unfortunately, more information does not always mean better information. For example, information on the Internet is often reported as being of poor relevance and validity [2,3]. The recent outcry of many scientists, including The American Association of Immunologists (AAI), about having preprint servers for biomedicine (see Box 1) was partly driven by the fear of getting burdened by an avalanche of non-peer-reviewed electronic junk-science that is impossible to cope with. With this article I will gently oppose this view and argue that electronic publication in research actually refers to two different processes: firstly, sharing data and intermediate results for collaboration and discussion, where speed and relevance are more important than in-depth prepublication peer-review; secondly, communication to bring reasonably validated research results into practice. By making this distinction, the absurdity of the opposition to preprint servers, which contain non-peer-reviewed material, becomes clear.

What is electronic publication?

'Publication' literally means 'making public' and the word 'electronic' refers to information that is stored only in computers. Electronic publishing in the broadest sense can therefore mean many different things: I will give five examples.

The first example is papers that have already been published in print journals and that are in addition adapted into electronic form, published for example by electronic publishers such as HighWire Press at Stanford University (www.highwire.org). HighWire, which started in 1995 with the online production of the weekly *Journal of Biological Chemistry*, today offers more than 150,000 free full-text articles from more than 200 printed journals. Also in this category belongs electronically 'self-archived' material that has appeared elsewhere in print, for example authors of scholarly papers publishing their work on their homepages, or universities building up databases with theses and research reports. The second example is scientific papers published exclusively electronically (e.g. on the World Wide Web), either by the authors themselves (e.g. on their homepages, without peer-review) or by peer-reviewed electronic journals. The first biomedical journal that was published exclusively electronically was the *Online Journal of Current Clinical Trials*, which started in 1992. The third example is drafts of scientific papers submitted by authors and published in so-called preprint databases (also referred to as 'e-print servers'), such as Netprints ([4]; see Box 1). The fourth example is publication of data and information in databases, for example nucleotide sequences in the EMBL/Genbank databases. The fifth example is that, in a broader sense, even the publication of meta-information — such as bibliographic information stored in databases such as Medline — may be referred to as electronic publication.

It should be noted that a grey area exists between what constitutes electronic publication and what doesn't, depending on what is considered 'public'. For example, if a researcher circulates a manuscript among a few colleagues via e-mail, not many people would actually consider this as 'electronic publication' whereas posting a manuscript to an electronic mailing list with hundreds of subscribers or publishing it on a preprint server or a website may already be considered electronic publication. This would result in certain journals following the so-called

Box 1
Preprint servers

Early experiments of distributing preprints and other 'type-1' communications among scientists in written form were conducted in 1961 by the US National Institutes of Health (NIH) and called 'Information Exchange Groups'. In the pre-Internet era scientists received photocopied material, which was a very costly process and which led to an end to the experiment in 1966 [19].

Electronic preprint servers evolved in the field of physics from August 1991 onwards and are now in many research areas an established medium to communicate non-peer-reviewed results of ongoing research among researchers. Preprint servers are actually Internet-accessible databases; they allow scientists to deposit electronic draft articles in order to make them accessible to a wider academic audience, before they actually submit them to a peer-reviewed journal. Strictly speaking, 'preprint' is a grossly misleading term because it suggests that papers published on these servers will eventually be 'printed', which may not necessarily be the case: firstly, it is not certain whether papers published on preprint servers will ever be submitted or accepted for publication at all; and secondly, if they are accepted by a peer-reviewed journal, they may well end up in an electronic journal and not necessarily in a printed journal. The term 'e-print server' (which is somewhat oxymoronic in combining the terms 'electronic' and 'print') may be even more confusing. Thus, when using the term 'preprint' we actually mean 'pre-peer-review' or 'pre-submission' documents.

The preprint server in the field of physics – formerly known as the 'xxx preprint archive' (xxx.lanl.gov, now known as ArXiv.org) – today serves 25 research disciplines, such as high-energy physics, economics, and atmospheric and oceanic sciences.

On 22 April 2000, The NIH director, Harold Varmus, and colleagues David Lipman (Director of the National Center for Biotechnology Information) and Pat Brown (a geneticist at Stanford University in Palo Alto) circulated a proposal for the first preprint server in the field of biomedicine; the server was first named 'E-Biomed', later 'E-biosci' and is now known as 'PubMed Central' (http://pubmedcentral.nih.gov) [20]. "Taxpayers have paid for research already, so NIH should make the results widely available…" was one of the arguments for the Varmus proposal to establish PubMed Central [21], which originally was not only meant to become a electronic repository for already published research but also was supposed to contain a preprint section that allowed researchers to submit papers directly without peer-review. This latter part of the proposal soon came under severe fire. The *New England Journal of Medicine*, which has earlier already argued that "…publishing preprints electronically sidesteps peer-review and increases the risk that the data and interpretations of a study will be biased or even wrong." [6], published an editorial pointing out that "The best way to protect the public interest is through the existing system of carefully monitored peer-review, revision, and editorial commentary in journals." [22].

As a result of the fierce criticism from scientific publishers, the NIH later dropped the idea of an electronic preprint server containing unreviewed material [23] and currently PubMed Central seems to have become an electronic platform to distribute full-text papers that have already been published in traditional journals or that have gone through peer-review by an editorial board (in other words, a platform primarily for type-2 communications). Meanwhile, the European Molecular Biology Organization (EMBO) also decided to create a free website, named E-Biosci, as a portal site for life-science papers; this is the European counterpart to PubMed Central [24].

However, the idea of a preprint server to serve type-1 communication has already been taken up by the several commercial publishers: for example, the *British Medical Journal* Publishing Group together with the Stanford libraries launched www.netprints.com as a preprint server for the entire field of medicine [4,25].

Ingelfinger rule to reject the article due to prior publication [5]. But how 'public' does a document have to be to constitute 'publication'? The *New England Journal of Medicine* once made clear that "…posting a manuscript, including its figures and tables, on a host computer to which anyone on the Internet can gain access will constitute…publication. On the other hand, sending manuscripts by e-mail to a limited number of colleagues — a dozen or two, let us say — will not." [6]. If 24 readers are not sufficient to create a 'public', how many readers are needed to constitute 'publication'? In fact, different journals have different policies on what they consider prior publication; for example *Nature* sees publication of sequences in electronic databases or draft manuscripts on preprint servers as part of the scientific communication process: "Genomics databases, like preprint servers and conferences, represent a form of intra-community networking from which all researchers benefit. *Nature* does not count them as prior publications." [7].

Type-1 and type-2 electronic publications

Much confusion and misunderstandings arise if people speak about electronic publishing and actually mean different things. Whereas traditional publication was a much better-defined dichotomous event, with a clear mission of transporting research results to the scientific community and the public, publication in the electronic age is much more a continuum [8] reflecting, and occurring during, the entire research process from hypotheses formulation to data gathering, raw data interpretation and the presentation and discussion of the final data. The more 'collaborative' research has to be, the earlier in this process electronic communication and 'publication' should occur. Electronic publishing in a broader sense includes the whole spectrum of electronic communication during the research process — for example, generating and sharing protocols and electronic draft data or draft manuscripts with research colleagues — whereas electronic publishing in a narrower sense refers only to the final, peer-reviewed release of data as a culmination of a research process.

It is important to discriminate between these two very different concepts of electronic publication. In the following I will refer to the former (electronic data released as part of the scientific collaboratory working process) as type-1 electronic publication and to the latter (carefully peer-reviewed electronic publication as a preliminary endpoint of a project) as type-2 electronic publication.

Type-1 electronic publications are characterized by opening work-in-progress to colleagues, thereby improving collaboration and quality. Typically 'published' information are draft data that need to be shared quickly among researchers, perhaps on a global scale, or preliminary

results that could benefit from the input of a broader research community. Genome databases containing draft nucleotide sequences are a typical example but so are preprint servers. Type-1 electronic publications have a similar validity to papers presented at conferences: they have not gone through a rigorous peer-review process but are primarily discussed during or after 'publication'. The very process of type-1 electronic publication is aimed at providing input from a broader research community. The emphasis of type-1 communications is not on validity (the reader is aware that he is dealing with draft data) but on openness and speed. Researchers look at these electronic publications because the results, despite being tentative, may be relevant to their own work. Researchers are expected to do their own 'downstream-filtering' of relevant information, which in the electronic world can be facilitated by providing meta-information [2].

Type-2 electronic publications are different. Their aim is to bring reasonably valid research results into practice. The results have important implications and they are expected to be acted upon on a wider scale. They may lead to changes in clinical practice or to policy changes. They are the preliminary endpoint of a long process of careful research, discussion and rigorous peer-review. The publication of a clinical trial in the *New England of Medicine* is a good example. Here the emphasis clearly lies on validity; 'upstream-filtering' in the form of peer-review prior to wide distribution is important.

Type-1 and type-2 communications are, in the electronic world, more difficult to discriminate from each other than in the traditional publishing world, where 'publication' was inevitably linked with the notion of peer-review and quality control and therefore immediately recognizable as type-2 communication. Unlike traditional publication, in type-1 and type-2 electronic publication the two processes of improving the quality and making the paper physically available are two distinct processes [9]. They may even occur in the opposite order as compared with traditional publishing — an ongoing peer-review process after publication is possible (e.g. by HighWire's 'rapid responses' or 'post publication peer-review [P3R]', as the journal *Pediatrics* calls it).

The confusion that arises if people fail to acknowledge that type-1 and type-2 communications are two different things can be best illustrated by the many responses to the PubMed Central proposal of having a preprint server for biomedicine (see Box 1). Among others, representatives of the American Association of Immunologists felt that the "…proposal compromises the cornerstone of scientific method: peer-review. The process described in your proposal…does not ensure a rigorous peer-review process. Without this we compromise our excellence (at best) and (at worst), pose potential harm to the scientific community as well as the public at large. Furthermore, scientists depend on the current peer-review process to give their work legitimacy and guidance; they do not want to be held to lesser standards." [10].

The concern here was that by having type-1 and type-2 communications on the same server, the non-peer-reviewed section would 'contaminate' and compromise the quality of type-2 communications. Other opponents of the proposal felt that readers could have trouble in distinguishing the different sections. *Proceedings of the National Academy of Sciences of the USA* felt that "…making non-peer-reviewed as well as peer-reviewed material available will confuse both scientists and the public…" [11]. However, this perhaps belittles the ability of scientists to recognize different levels of evidence and to be able to interpret labels that could make clear that certain material is non-peer-reviewed content, as used in Netprints — after all, "…this is the age of transparency rather than paternalism…" [8] as Richard Smith, editor of the *British Medical Journal*, put it.

The benefits and problems of type-1 'open' electronic publication

One example of the benefits of open communication and data sharing comes from the 'open-source software' industry. This comprises computer programs, and developers freely distribute the source code and allow usage and modification. The Open Source Initiative explained the concept as follows: "The basic idea behind open source is very simple. When programmers on the Internet can read, redistribute, and modify the source for a piece of software, it evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing…We in the open-source community have learned that this rapid evolutionary process produces better software than the traditional closed model, in which only a very few programmers can see source and everybody else must blindly use an opaque block of bits." (http://www.opensource.org/). Replace 'software' with 'research' and 'programmers' with 'scientists', and you have a perfect justification for type-1 open-source publishing. It is also noteworthy that the initiative says that "The foundation of the business case for open-source is high reliability. Open-source software is peer-reviewed software; it is more reliable than closed, proprietary software. Mature open-source code is as bulletproof as software ever gets.".

Although, during the development process, open-source code may appear immature, preliminary, non-peer-reviewed and of lower quality than commercially available software, the software industry has learned that the end-product of open-source development is of superior quality. What is true for the software industry has strong parallels to the area of research: perhaps its strongest analogy in the field of genomics [12], where it is (according to the so-called Bermuda agreement) common practice for researchers to place sequence data on public and freely accessible databases as sequences are generated (non-peer-reviewed and in a draft status). The analogy may also be extended to preprint servers, which allow research protocols, draft papers and datasets to be published and reviewed by others, who could give valuable input.

In certain areas, such as in genomics research, type-1 electronic communication is a necessity to foster international collaboration. In other areas, for example clinical research, electronic publication can also help reviewers who attempt to synthesize research in an unbiased manner. A current problem for authors of reviews on the effectiveness of a clinical intervention is that the literature may be biased in favor of positive or promising results, which are more often published in paper journals than negative results (this is known as publication bias). This may affect the validity of systematic reviews [13]. Electronic registers of clinical trials (another kind of type-1 electronic publishing), where investigators publish their research protocols from the early stages onwards, can later help to identify negative trials that remained unpublished [14].

It is likely that, as in clinical research, many results in experimental research are only published if they are desired or significant. In experimental research, preprint servers could play a similar role as prospective trial registers in clinical research: scientists can deposit protocols of ongoing experiments and briefly report findings electronically that otherwise would not deserve publication, thereby providing a perhaps more genuine picture of reality.

Despite these considerations, it must however be acknowledged that openness also brings at least three problems concerning intellectual property issues: firstly, debates over priority, authorship and credit for analyzing draft data in depth that have been made entirely open may arise [7]; secondly, the danger of plagiarism from non-peer-reviewed electronic material that has been made public [15]; and, thirdly, the problem that European patent laws (contrary to US laws) do not allow the patenting of data that have been published [16]. Therefore, not all material is suitable for preprint servers.

Conclusions and outlook

Totally new concepts of 'publishing' and distributing data will evolve in the near future. Type-1 electronic publishing may be become a subtle process that will have nothing in common with what we traditionally know as publication. One example is software using so-called 'Napster technology' that allows searching for certain data across the hard disks of all scientists who are willing to share their data. This kind of software is already envisaged to help the annotating of genome sequences in a collaborative way. [17] The simple act of a scientist marking one of his files as publicly accessible may already constitute publication.

These developments also challenge the way that research currently is being evaluated. In their criticism of the PubMed Central proposal, The American Association of Immunologists wrote that presently "…scientists depend on the hierarchy of journals to help them select the most important studies in the plethora of information available to them. It is unclear how a single information source would assist this sorting process." [10]. Clearly, traditional methods to assess the value of research — such as journal impact factors — will become redundant [18]; however, new methods will evolve. The value of a manuscript will become more important than the impact factor of the journal in which it is published. Electronic publishing will provide alternative models, for example a 'paper auction' model: researchers could submit type-1 electronic papers to preprint servers for discussion and peer-review, and journal editors and publishers would pick and bid for the best papers they want to see as 'type-2 papers' in their journal. The best journals would be able to pay the highest prices for the best papers and the number of bidders or the sum that was bid for each paper would determine its value.

As the number of projects that all share the common goal — to improve electronic scholarly communication — is increasing, co-operation and interoperability between these developments are becoming key challenges. Although Internet technology provides the basic protocols to link different services physically, higher-level standards are needed to ensure interoperability. The Open Archives initiative (www.openarchives.org) has recently taken a first step in proposing a convention that provides a technical and organizational framework to support basic interoperability among e-print archives.

The costs shift away from publishing and distributing information, and towards finding and managing relevant and valid information. Accessibility and connectivity of information need to be improved: in type-2 publishing, data are filtered upstream (by means of peer-review) whereas in type-1 publishing, scientists need to be able to select and filter relevant information downstream, which requires labeling with computer-readable meta-information [2].

In an ideal future, researchers should be able to browse through a global knowledgebase in order to search across different literature databases, full-text archives and digital libraries, and to navigate seamlessly from one publisher's server to another and from database producers and preprint servers.

References

1. Waksman BH: **Information overload in immunology: possible solutions to the problem of excessive publication.** *J Immunol* 1980, **124**:1009-1015.
2. Eysenbach G, Diepgen TL: **Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information.** *Brit Med J* 1998, **317**:1496-1500.
3. Allen ES, Burke JM, Welch ME, Rieseberg LH: **How reliable is science information on the web?** *Nature* 1999, **402**:722.
4. Delamothe T, Smith R, Keller MA, Sack J, Witscher B: **Netprints: the next phase in the evolution of biomedical publishing.** *Brit Med J* 1999, **319**:1515-1516.
5. Angell M, Kassirer JP: **The Ingelfinger Rule revisited.** *N Engl J Med* 1991, **325**:1371-1373.
6. Kassirer JP, Angell M: **The Internet and the journal.** *N Engl J Med* 1995, **332**:1709-1710.
7. Anonymous: **Debates over credit for the annotation of genomes.** *Nature* 2000, **405**:719.

8. Smith R: **What is publication?** *Brit Med J* 1999, **318**:142.
9. Eysenbach G: **Challenges and changing roles for medical journals in the cyberspace age: electronic pre-prints and e-papers.** *J Med Internet Res* 1999, **2**:E9.
10. E-biomed – A proposal for Electronic Publications in the Biomedical Sciences on World Wide Web URL: http://www.nih.gov/about/director/ebiomed/com0627.htm#aaoi185
11. Macilwain C: **PNAS joins peer-reviewed PubMed Central. Proceedings of the National Academy of Sciences.** *Nature* 1999, **401**:733.
12. Russ AP, Aparicio SA, Carlton MB: **Open-source work even more vital to genome project than to software.** *Nature* 2000, **404**:809.
13. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR: **Empirical assessment of effect of publication bias on meta-analyses.** *Brit Med J* 2000, **320**:1574-1577.
14. Tonks A: **Registering clinical trials.** *Brit Med J* 1999, **319**:1565-1568.
15. Eysenbach G: **Report of a case of cyberplagiarism — and reflections on detecting and preventing academic misconduct using the Internet.** *J Med Internet Res* 2000, **2**:E4.
16. Abbott A: **Germany rejects genome data 'isolation'.** *Nature* 1997, **387**:536.
17. Butler D: **Music software to come to genome aid?** *Nature* 2000, **404**:694.
18. Brunstein J: **End of impact factors?** *Nature* 2000, **403**:478.
19. Confrey EA: **Information exchange groups to be discontinued.** *Science* 1966, **154**:843.
20. Marshall E: **Varmus circulates proposal for NIH-backed online venture.** *Science* 1999, **284**:718.
21. Marshall E: **Varmus defends E-biomed proposal, prepares to push ahead.** *Science* 1999, **284**:2062-2063.
22. Relman AS: **The NIH 'E-biomed' proposal — a potential threat to the evaluation and orderly dissemination of new clinical studies.** *N Engl J Med* 1999, **340**:1828-1829.
23. Marshall E: **E-biomed morphs to E-biosci, focus shifts to reviewed papers.** *Science* 1999, **285**:810-811.
24. Butler D: **Europe strengthens its hand in bioscience website talks…** *Nature* 1999, **401**:413.
25. Delamothe T, Smith R: **Moving beyond journals: the future arrives with a crash.** *Brit Med J* 1999, **318**:1637-1639.