Access all areas?

The Internet has revolutionized how we live and work — with dynamic new technologies it is set to change the way in which science reaches its audience.

This week's Commentary pages offer three highly individual perspectives on future developments in access to the primary scientific literature. On this page, Tim Berners-Lee and James Hendler present their vision of a semantic web, where new technologies will allow computers, as well as people, to understand and communicate with each other and hence to revolutionize scientific publishing. On page 1024, Stevan Harnad proposes "freeing" the refereed research literature by implementing an authors' searchable, self-archiving system that would turn publishers into providers of a peer-reviewing service rather than producers of journals. And on page 1026, Ira Mellman argues that the "public library of science" initiative should focus on making journals' content free-access six months after publication, rather than on pressurizing them to permit authors to re-display published material on other sites.

These articles are also published, in slightly different form, on *Nature*'s website as part of its current web debate on electronic publishing initiatives in science (see *Nature* **410,** 613; 2001). Readers wishing to participate in the debate are invited to view http://www.nature.com/nature/debates/e-access. Contributions can also be submitted to Correspondence via corres@nature.com. In either case, publication will be offered according to the criteria described on page 613 of the 5 April issue.

Publishing on the semantic web

The coming Internet revolution will profoundly affect scientific information.

Tim Berners-Lee and James Hendler

To predict the future of scientific publishing on the World-Wide Web, it is important to understand how web technology is changing. We are in the early days of a new web revolution, one that will have profound implications on web publishing, and on the nature of the web itself. Just as current web technology is changing the world of publishing, the new semantic web technology (http://www.w3.org/2001/sw) may change the way scientific knowledge is produced and shared.

The web was designed as an information space, with the goal not only that it should be useful for human–human communication, but also that machines would be able to participate and help users communicate with each other. A major obstacle to this goal is the fact that most information on the web is designed solely for human consumption. Computers are better at handling carefully structured and well-designed data, yet even where information is derived from a database with well-defined meanings, the implications of those data are not evident to a robot browsing the web. More information needs to be in a form that the machine can 'understand' rather than simply display.

The concept of machine-understandable documents does not imply some magical artificial intelligence allowing machines to comprehend human mumblings. It relies solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data. So, instead of asking machines to understand people's language, the new technology, like the old, involves asking people to make some extra effort, in repayment for which they get major new functionality — just as the extra effort of producing HTML mark-up is outweighed by the benefit of having content searchable on the web.

A new set of languages is now being developed to make more web content accessible to machines. The Semantic Web Activity, run by the World Wide Web consortium, is defining new web technologies that will enable successively better tools that make it easier for people to create machine-readable content and make it widely available.

What impact might this have on scientific publishing? In the next few years, we expect that tools for publishing papers on the web will automatically help users to include more of this machine-readable mark-up in the papers they produce. Where a current tool using XML (see http://www.nature.com/nature/webmatters/xml/xml.html) can allow a user to assert that some part of a document is about an 'experiment', the new languages will let the scientist express that the experiment uses certain chemicals and reagents; that the system used involved some particular organic matter; that the experiment produced gels with certain DNA information on them (and that the images of these gels are located in particular places on the web); and so on.

Papers that include this new mark-up language will be found by new and better search engines, so users will be able to issue significantly more precise queries. More importantly, experimental results can themselves be published on the web, outside the context of a research paper. So a scientist can design and run an experiment, and create an emerging web page containing the information that he or she wants to share with trusted colleagues (see Fig. 1). Finding out about experiments and studies in progress will be easy, and work can be modified as a result of interaction with peers, with less need to wait for formal publication. Just as preprints challenge established journals' online versions, these new 'papers in progress' will be a significant challenge to online scientific publishers.

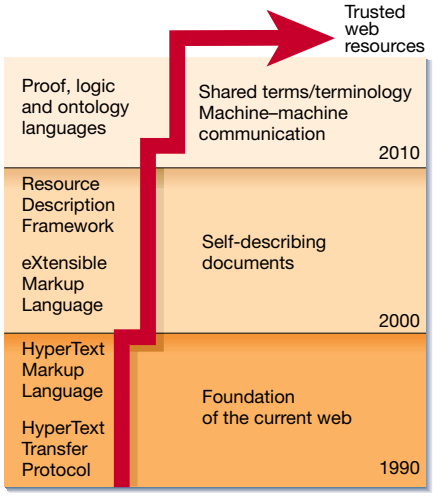

Figure 1 **Layers of the semantic web are built as new languages and tools anchored in XML. They build towards a world of trusted information shared among collaborating groups of users.**

In the long run, the effects on publishing may be far more profound. There is an eternal conflict between operating rapidly as a small group and taking the time to communicate more widely. The former is more efficient but produces a subculture whose concepts and results are not understood by others. The latter can be painfully slow. The world works as a spectrum between these extremes, with a tendency to start small — from the personal idea — and filter over time towards a wider commonality of concept. The joining together of subcultures when there is a need for a wider common language is an essential process in the

development of human communication.

The semantic web will facilitate the development of automated methods for helping users to understand the content produced by those in other scientific disciplines. On the semantic web, one will be able to produce machine-readable content that will provide, say, automated translation between the output of a scientific device and the input of a datamining package used in some other discipline, or a self-evolving translator that allows one group of scientists directly to interact with the technical data produced by another.

These new products will help users communicate with each other even before a common vocabulary has developed to express the concepts they have in common. The semantic web will provide unifying underlying technologies to allow these concepts to be progressively linked into a universal web of knowledge. Thus it will aid in breaking down the walls of miscommunication, allowing researchers to find and understand products from other scientific disciplines. The very notion of a journal of medicine separate from a journal of bioinformatics, separate from the writings of physicists, chemists, psychologists and even kindergarten teachers, will some day seem as out of date as the print journal is becoming to our graduate students.

Talking point: Tim Berners-Lee backs more participation from computers in web publishing.

Does this sound like a crazy science-fiction dream? A decade ago, who would have believed a web of text, conveyed by computer, would challenge a 200-year-old tradition of academic publishing? ■

Tim Berners-Lee, inventor of the World-Wide Web, is at the World Wide Web Consortium, MIT, 545 Technology Square, Room NE43-356, Cambridge, Massachusetts 02139, USA (http://www.w3.org/People/Berners-Lee). James Hendler is in the Computer Science Department, University of Maryland, College Park, Maryland 20853, USA.

The self-archiving initiative

Freeing the refereed research literature online.

Stevan Harnad

Unlike the authors of books and magazine articles, who write for royalty or fees, the authors of refereed journal articles write only for 'research impact'. To be cited and built on in the research of others, their findings have to be accessible to their potential users. From the authors' viewpoint, toll-gating access to their findings is as counterproductive as toll-gating access to commercial advertisements.

With the online age, it has at last become possible to free the literature from this unwelcome impediment. Authors need only deposit their refereed articles in 'eprint' archives at their own institutions; these interoperable archives can then all be harvested into a global virtual archive, its full contents freely searchable and accessible online by everyone (see Box).

Unlike the royalty/fee-based literature, which constitutes the vast majority of the printed word, the special, tiny literature of refereed journal articles is, and always has been, an 'author giveaway'. Researchers never benefited from the fact that people had to pay access tolls to read their papers (as subscriptions, and for the online version, site-licences or pay-per-view). On the contrary, those access barriers represent impact barriers for researchers, whose careers and standing depend largely on the visibility and uptake of their research.

There are currently at least 20,000 refereed journals across all fields of scholarship, publishing more than 2 million refereed articles each year. The amount collectively paid by those of the world's institutions which can afford the tolls for just one of those refereed papers averages $2,000 per paper[1]. In exchange for that fee, that particular paper is accessible to readers at those, and only those, paying institutions.

The research libraries of the world can be divided into the (minority) Harvards and the (majority) Have-nots — the last by no means limited to the developing world. It is obvious how the Have-nots would benefit from free access to the entire refereed literature, for without it their meagre serials budgets can afford only a pitifully small portion. But not even Harvard can afford access to anywhere near all of the literature (see http://fisher.lib.virginia.edu/newarl/index.html). Hence, most refereed articles are inaccessible to most researchers. For the authors, this means that much of their potential impact is lost. And it is solely this curtailed research impact and access that is being purchased by the collective $2,000 outlay per article mentioned above.

This is the way things had to be in the past, when publishing as print-on-paper was the only medium, and the sizeable costs of printing and distribution had to be recovered somehow. The new online era may be threatening the majority, royalty/fee-based literature (books, magazine articles) in the form of digital piracy; but for the 'giveaway' research literature, it has at last made it possible to eliminate all those counterproductive access/impact barriers.

Not all costs have vanished, of course. Although the costs of printing and distribution (and their online successors, such as publishers' PDF page-images) are no longer essential ones, the cost of the quality-control and certification that differentiates the refereed literature from an unfiltered, anarchic vanity press still needs to be paid. Paper and PDF files have become mere options, purchasable by those who want and can afford them. Refereeing, however, is essential.

Essential costs of refereeing

Refereeing (peer review) is the system of evaluation and feedback by which expert researchers assure the quality of each others' research findings. Referees' services are donated free to virtually all scientific journals, but there is a real cost to implementing the refereeing procedures, which include archiving submitted papers on a website; selecting appropriate referees; tracking submissions through rounds of review and author revision; making editorial judgments, and so on.

The minimum cost of refereeing has been estimated as $500 per accepted article by the American Institute of Physics (see http://documents.cern.ch/archive/electronic/other/agenda/a01193/a01193s4t8/transparencies/Doyle.ppt), but that figure almost certainly has inessential costs wrapped into it (for example, the creation of the publisher's PDF). I think that the true figure for peer-review implementation alone across all refereed journals probably averages closer to $200 per article, or even lower. Hence, quality-control costs account for only 10% of the collective tolls actually being paid per article.

Can this situation, in which the authors' and referees' giveaways are needlessly being held hostage to obsolete printing costs and cost-recovery methods, be remedied? Note that it is not simply a matter of lowering the financial access barriers: even if those were slashed by 90%, most researchers would still be unable to access most research papers. There is an optimal solution, and it is inevitable: the refereed research literature must be freed online for everyone, everywhere, for ever. The irreducible 10% or so quality-control cost need no longer be paid for by readers' institutions; it can be paid in the form of quality-control service costs, per paper published, by authors' institutions, out of their savings on subscription costs.

Journal publishers certainly will not scale