Electronic Publishing and Visions of Hypertext

Will editors of journals such as this in a few years' time be out looking for new jobs? Will a world, overrun with forests, only use paper for packing the potato crisps eaten by hungry hackers? Should you save this issue of *Physics World* as a possible collector's item?

Experience with the "World-Wide Web" (W^3) global information initiative suggests that the whole mechanism of academic research will change with new technology. However, when we try (dangerously) to see the shape of things to come, it seems that some old institutions may resurface but in a new form.

The change from paper to electronic form is essentially a change of timing. It will take the same amount of time to read a page of text, but to order a reference will take a few seconds rather than a few days. It will take the same amount of time to compose an article, but to search a library catalogue will take a few seconds rather than a few hours (including the trip to the library). The change in timing will affect the whole way we do work.

Let us assume a scenario in which any note I write on my computer I can "publish" just by giving it a name, and making it available to the outside world. In that note I can make references to any other article anywhere in the world in such a way that when you read my note you can click with your mouse, and bring the refernced article up on your screen. Suppose the mechanism for making that link is as simple as you can imagine – another couple of mouse clicks for example – and suppose moreover that everyone has this capability. These are the assumptions of "global hypertext" and it is generally supposed that this will lead to a tangled web of interconnected jottings representing the sum of human knowledge: Hofstadter's anthill at last. By selectively following links passed to me by friends, I can rapidly find anything I want to know. Modelling the real world with all its random associations, the "web" allows me to replace a day's worth of library visits, discussions over coffee and rummaging in filing cabinets with a dozen or so clicks with the mouse.

Before we jump in and take this to pieces, let me assure the cynical that to a certain extent, this exists, and where it exists, it works. The W^3 initiative at CERN and various other institutes has, along with a few similar projects, put together the infrastructure of network protocols and common software. It seems to be taking off, to judge from servers cropping up increasingly frequently, and readership of our own server doubling every other month. Even without global authorship, global readership of data provided by the few has been spectacularly successful. Almost all the data on the "web" is a window onto some other source, so it is not hand-crafted hypertext, but it still sells. (All the same, I can also assure you that as I write I am frustrated by my inability to embed into this article helpful links for you, gentle reader. Writing hypertext is more fun!).

The web is still seen by most people as a global information system, and so it will be until the next generation of compatible hypertext editors come along.

In this happy anarchy, two problems arise. One is that of collective schizophrenia.

The bulk of human understanding may well develop two independent pockets of knowledge about the same thing. This can happen on a small scale, when one writes a document with the sinking feeling that one has written it before but can't find it. It can happen on a global scale when researchers on different continents investigate the same phenomenon, unaware of each other.

To solve this, some global coordination is clearly required. However, centralised coordination is out of the question for an estimated 10^{14} documents.

A number of people have started to make lists of resources on the network and have generally been swamped by its growth. The most spectacular success is the "archie" project which keeps a mammoth index of the names of almost all the files available in the internet archives worldwide. Even Peter Deutch, its instigator, admits that network information is likely to grow faster than his disks, and he will have to specialise. My own attempt to edit a hypertext encyclopaedia, in which pointers to network information sources are classified by subject, leaves me overwhelmed even now. As I looked around for people to help, I realised that I was looking for specialists in particular fields to look after them – like specialised librarians.

The bringing together of the provider of information and the enquirer, the "resource discovery" problem, is up for grabs in the networking community. Solutions, however, always centre on some idea of "subject". The keyword list, or vocabulary profile, of a document is used to route it to some specialised index which will note it, and direct enquirers to it. Whether you take the Dewy decimal system or the English language as a basis, there need to be centers of knowledge on particular subjects. So, at CERN, we keep pointers to information at other HEP sites. I'd like to do more of this. I'd like to maintain an eminently readable hypertext overview of the field, with links to more detailed discussions of specific areas, and eventually to the work of particular groups and individuals. I am not sure whether I would call the result an encyclopaedia, or a journal, or a library. The job-title "cybrarian" has been suggested. However, I can tell you from experience, that it takes an incredible amount of time. The tasks of librarians and reviewers are not going to be usurped by academics in their spare time.

As a simple example, let me take the CERN "home" page. This is a page which the uninitiated user sees when he first types "www" or "telnet info.cern.ch". This page must be simple, clear and informative. It must provide direct links to the most frequently used information sources (such as the online phone book), but still provide a structure which will give the reader a basis for his browsing further afield. The home page is only 23 lines on a terminal, and only one of hundreds of hypertext documents we maintain, but it is constantly under review and its format is the source of interminable debate. If this page takes so much effort, it is evident how much effort is needed to keep the whole web in good shape.

The second problem which the information web faces really started with the laser printer. Before laser printers came around, you could tell something about the reliability of an article by its feel: handwritten scrawl torn off a spiral notebook never carried quite the authority of glossy typography. Nowadays, it all looks the same, elegantly Optima 10 point. The same will be true of networked information. It is true that individuals slip into new conventions for conveying formality or lack thereof. A lower case i for the first person gives electronic mail a "hastilly scribbled" impression, not to mention the conventional faces on their side (:-) of internet news. As well as conventions, new ethics for electronic publishing are developing. However, one needs the equivalent of a refereed journal to convey authority. In the hypertext world, the actual physical distribution of data is not the issue here: it is the organisation. The feel of the paper of a document will be replaced by its registered name. A document registered under my personal authorship will not carry the same weight as one registered in the International Standards Orgnisation's catalogue of standards.

I see the need for two organs: the newsletter and the encyclopaedia. An encyclopaedia will be an overall attempt by the knowledgeable, the learned societies or anyone else, to represent the state of the art in their field. An encyclopaedia will be a living document, as up to date as it can be at any time, instantly accessible. It will contain carefully authored explanations and summaries of the subject, as well as computer-generated indexes of literature. A reference to a paper from the encyclopaedia conveys authority and acceptance by academic society. A measure of a paper's standing maybe conveyed by the number of links it is is away from an encyclopaedia.

The newsletter is a commentary on the changes in the field. A personalised newsletter can be generated automatically by looking for changes in the encyclopaedia and linked works. A user may effectively ask his computer each day, "Tell me anything new which has been linked to one of my favorite subjects". It may be possible to generate a newsletter largely automatically, but a human being does so much better. Especially this is true of the job of summarising in a review or contents page.

I deliberately say "an encyclopaedia" rather than "the encyclopaedia". Another fundamental change will be the low start-up cost of publishing. Anyone can start a new encyclopaedia, and, if enough people refer to it, will be widely read, and quoted by society's established authorities. This allows for many encyclopaedias, even many parallel societies. Conventional science will have no hold on pockets of alternative ideas and, so long as innocent individuals are not misled, I see great worth in this freedom. Here we hope to see a market economy in information. The quality of an article is judged by its own contents, and by the quality of the articles to which it refers. There is therefore an incentive to refer to good articles, so the better articles will be most referred to, and most read. Authoritative sources will take care only to refer to reliable work, but deserving small journals can start and rapidly gain prestige. The many medium-size discussion groups of the internet news system provide a vehicle for bringing new sources quickly to light and establishing acceptance.

These arguments convince me that publishing houses, far from being unnecessary, will be in for very exciting times. Their jobs and those of librarians seem to have merged into one as classifiers and reviewers of the world's knowledge. The practical issue arises of how to pay these good people.

There is something distasteful about charging by the byte. The idea of freedom of information is sullied by a price tag on an icon, a taxi-meter ticking away on the corner of the screen. Also, it is difficult to find a general method for previewing a document so that one can see what one will get for one's money. How can one ascertain which documents were really read, or even then were actually useful to the reader? The clear advantage of this technique, however, is that the information "market" becomes more real and more direct when real money is tied in at a low level.

Payment by subscription, then, has its appeal. Just as one subscribes to a journal, or l gains the right to use a library, so one would subscribe to an information service. I pay not for the information I read but for that which I have available just in case. However, the power of global hypertext to represent knowledge lies in the unconstrained way links can cross boundaries between organizations, subjects, and continents. Following a link should ideally take under a quater a second not to disturb the train of thought. It should not be accompanied by questions about account numbers and credit ratings.

Alternatively, the charging and the paying is done between organisations, over negotiating tables, behind the back of the poor researcher. Let a consortium of physics institutes commission an electronic journal, give it a budget and review it from time to time. Cross-licensing between societies so that, for example, members of the Institute of Physics will be granted access to chemical journals. Perhaps we can imagine an association of publishers which, operating like the Performing Rights Society, attempts to redistribute the money in a fair fashion.

A mixture of such schemes may exist, and the market may decide which one works best. The market will be fierce, and enthusiastic amateurs will always be willing to compete where they feel a professional service falls below a an certain standard.

The existing web gives a good feel for what is possible, using free information. Within high-energy physics, the web contains mainly user manuals, online help, phone books, discussion lists, announcements, news, the minutes of meetings and preprint lists. In other subjects, data range from catalogues of DNA sequences and chemical formulae, through poetry, prose and religious books to the weather forecast. Thanks to the spread of the Internet, this is available in most academic institutes (even, nowadays, in the UK). Already it shows us a more efficient way to pool our knowledge, while keeping up standards of freedom of information which academic world, and the Internet, have always promoted.