THE LONG ROAD TO INFORMATION INTEGRATION: SUGGESTIONS FOR THE WAY FORWARD

Suzie Alexander

Paper presented at the UKSG 21st Annual Conference, Exeter, April 1998

This paper provides a concise overview of the relative merits and disadvantages of the publisher direct, subscription agency gateway and aggregator models of electronic publishing, currently prevalent. Advantages and disadvantages of each approach are identified from a users perspective and the barriers to more extensive use of electronic resources are identified. The recommended way forward concludes that each player, publisher, subscription agent and software integrator should focus on their core skills and cooperate to provide the end user with a truly integrated electronic retrieval and delivery environment.

Suzie Alexander is Director of European Operations, Ovid Technologies, 107-109 Hammersmith Grove, London W6 0NQ
E-mail: suziea@ovid.com

As I visit teaching and research institutions across Europe, information integration is increasingly top of the agenda whenever I talk to anyone trying to juggle print, electronic and other information resources. The concept of integrated information, which is commonly understood to mean all information types, accessible from anywhere, through a standard interface, promises increased ease of access to information resources and increased usage. By contrast 'disintegrated information' conjures up an image which more fairly reflects how many end users view the current state of affairs in electronic information delivery: incomplete, badly organised, difficult to use, too slow, and print is easier to use, are amongst the politer comments I have heard.

In order to understand better why this might be, I will be looking at the electronic delivery models currently available on the market to assess how people use them and why none of them really respond to users' needs. Many of the examples I will use to highlight this point will relate to the experience of the scholarly community given that their information requirements are amongst the most stringent.

As we are all aware end users need to access complete content in the simplest, most time efficient way possible. Both Internet developments and fast evolving information retrieval technology are enabling delivery of content in electronic format, but users are encountering significant obstacles which are both time-consuming and which currently detract from the real benefits of electronic delivery. I perceive that it is the role of the information industry to take the lead in exploiting the opportunity presented by available technology to make integrated access in the electronic environment a reality.

In so doing we should not lose sight of the wider context in which we are all operating, i.e. that the only reason all of us in the information industry are here is to service the information needs of the end users. The case of the scientific community is particularly

acute in that those needs often relate to accessing information which they have originally created. Once it becomes apparent that the technology to integrate is available, but that the information industry itself is resistant to this objective, I sense that the scholarly community will become heavily involved in finding alternatives in order to achieve the integration they require, and pre-print publishing without publishers is one example of this.

There are three models prevalent today for access to electronic information, all three of which are Web-based. The three models are the Publisher model, the Distributed model and the Aggregated model. Let's look at common issues associated with end user access to electronic information and then how they fare in the context of our three models:

Delivery: how the information is delivered

Access: what are the points of access?; what kind of tool (generally a Web browser) is required to access and retain information

Login: how to login; password distribution and IP address authentication

Content: what exactly is being made available, how complete it is, and how reliable it is as an information source

Currency: how current is the information

Text format: what kind of format the information is in; what the implications are for viewing, storage and retrieval in the future

Browsing: this is an integral part of any end-user's approach to available information; how possible is this in the electronic environment?

Customisation; how customisable is both access to the information and the information itself?

Archive: how is the information going to be archived, and available for future use?

Cost and benefit: what is the cost and benefit of electronic information?

The Publisher model is currently the most prevalent and is where in particular the print publisher takes responsibility for the electronic publication of material. Internet delivery usually takes the form of access through a Web site URL, which the end user may find by searching on the Web and then bookmark, or through an institution's own listings of free or subscribed to publisher sites. This can prove satisfactory for the browsing of the latest issue of a particular journal. However, wishing to run a comprehensive subject search, herein lies the first important hurdle.

End users generally start an information search with a subject in mind, not a publisher or even a journal name. In Neil Jacobs' paper last year on the Pilot Site Licence Initiative he states, and I quote, 'Researchers generally do not know or care who publishes the journals in which they are interested...'[1] Can we seriously expect end users to search their subject repeatedly against each individual publisher site? End user searches often necessitate recovering everything ever written on any specific topic. Presumably the end user should not have to imagine, for example, in which journals articles appeared and then identify the publisher name in order to consult their web site, or at least not whilst there is still more than one publisher.

Let us take a step back. Access to the Web is via Web browsers such as Netscape or Internet Explorer which contain search engines. Your household Yahoo is not currently sophisticated enough, and was not designed for, searching on scientific information. This leaves the internet bibliographic database tools, each with their own interface, but nonetheless bringing the end user to an accurate, complete, and reliable set of article titles (which could number six or fifty or more) which he/she should print out, and then search title by title and publisher by publisher.

Assuming that your institution site has subscribed to the publisher site the login may either be passworded or direct through a site licence. Whilst technically possible to access from anywhere, at the user's convenience, whether this is allowed by the constraints of the publisher licence or the publisher's technology is again another matter.

Once through this far to the first publisher site the end user must discover exactly which journals are included and period of time covered. The coverage is often inexplicably incomplete. Some titles are included others not. The time period varies according to a number of factors. Depending on which electronic format has been adopted (e.g. PDF, HTML, SGML) and when the publisher starting publishing in electronic format

will determine how far back the coverage goes. PDF format, whilst initially less costly, does not allow the publisher to publish electronically retroactively. Some publishers use a combination of formats to allow, for example, an SGML header field to be searchable, but the user is always searching on that particular publisher's content only. Whereas the currency of the electronic data is often linked to the publisher's concern for maintaining print revenue. This is being overcome somewhat through pricing models which encourage institutions to continue print subscriptions alongside the electronic.

Let us go back to how the user searches. Taking how the print is used by end users is probably not a bad start. A bibliographic search, retrieval of articles, perusal of references, browsing of special issues for further leads to articles of interest. In the electronic environment two issues arise here. In the first place once identified an article on the Web it is not possible to link on at the click of a button either within, or more likely outside the publisher site and therefore the user is constrained to return to the beginning and start again.

Having identified an article that the user wishes to view, viewing is again dependent on which format the electronic version is delivered in. PDF is currently the most widespread format in use. For examples of the practical difficulties encountered using this proprietary standard requiring Adobe Acrobat I would again refer you to the article written by Neil Jacobs last year[1]. I think we could forgive users at this stage from thinking that print was possibly quicker, certainly more user friendly, and in the cost-benefit equation although dealing with publishers directly is the cheapest option, probably little is gained compared to what is being spent.

Moving on to the second model, the Distributed model, which I would define as where an information intermediary, often a subscription agent, delivers access through a common software entrance point to a range of information sources which they do not modify. This approach sees some similarities and some advantages over the first model. Let's look at the main differences. It provides the user with a single point of access and a single login to a list of publishers and/or a list of journal titles. It oversees the updating of web links to reach the information subscribed to. But mainly it takes responsibility for the arduous task of administering electronic subscription licences with multiple publishers, consolidating invoicing for print and electronic subscriptions, allowing institutions to deal with one supplier rather than many.

However, I feel we have again lost sight of the end user here. The Distributed model is a reactive process, it contributes nothing in terms of how user-friendly the information it distributes is. Whilst it brings the broadest coverage in terms of amount of content, the end user is still faced with a different interface for each publisher or journal title, some requiring adobe acrobat, some not. Whilst a subject search can be performed, the search tools are relatively unsophisticated compared to those already available elsewhere on the market. The search result set obtained is often limited to the electronic resources made available by the distributor, and is therefore incomplete. And from the result set the end user is forced to go to one article at a time with no possibility to browse on seamlessly. The search process is artificial and currently less natural than using the print therefore cost and benefit are again debatable, but more acceptable than in the publisher model.

The third model, the Aggregated model, is where an intermediary takes a range of journal titles from different publishers and modifies the content on their behalf in order to facilitate consultation. Where SGML format is implemented this model allows the use of sophisticated bibliographic search tools to link directly to electronic articles and link on to further electronic articles under one interface. This model is where the available technology has so far been applied to best effect for the end user, which emulates their searching behaviour and which carries electronic information access visibly beyond the boundaries of print access and into new realms which the user welcomes.

Software and technology companies are also able to contribute in the area of practical implementation. The services they run are faster. They can better determine which format to put the information into to encourage users to read on the screen rather than print out, use thumbnail graphics to make the most efficient use of network bandwidth, allow user customisation of data in terms of how they want to go on and use that information in other software packages, know

how to best store the information for long term archive access, determine which are sustainable production processes to go forward.

Looking beyond the journals the next phase in information integration which is already underway will soon see the incorporation and integration of books, audio and video material, x-rays, in fact any information resource. They add value in the electronic environment, but at a higher cost than the other models, but most importantly they do not own the content and cannot obviously determine its inclusion.

Information integration will have to be a combined effort. There is a role for all existing players in the information industry, as well as some new ones. Electronic delivery needs to be viewed as a content issue, an economic issue and a technology issue, and not any of these in isolation. Without all three components electronic successful information integration is a long way off. Let us all do what we do best: publishers publish, subscription agents administrate, software houses develop technology.

Working groups should be created to facilitate the cooperation needed to develop standards which are acceptable to all parties. If, on the other hand, the outcome is to be determined solely by market forces, then I would appeal to librarians to consider carefully, spend wisely, and feel an obligation to keep both the end users and future generations very firmly at the forefront of their minds.

Reference

1 Jacobs, Neil . Pilot Site Initiative (PSLI): User Perceptions of Necessity vs. Novelty. *Serials,* Vol.10 No.2, July 1997.