Current Full Text Models

Author
Bette Brunelle
Director, Database Technology
Ovid Technologies Inc.
betteb@ovid.com

Presenter
Mark Schregardus
Director, Asia /Pacific
Ovid Technologies Inc.
marksch@ovid.com

Abstract
This paper discusses the various full text models that have emerged in the last few years. Three basic models of full text, with variations, will be submitted to a reality check as concerns their advantages, disadvantages and challenges that lay ahead. We will be discussing Publisher-supplied full text; third-party, or Aggregator supplied full text; and Distributed, "linked" full text - in which a bibliographic database provider links to (usually) publisher-supplied full text.

Full Text Models

Introduction

In just a year or two full text has gone from a great amount of hype and promising press releases to enough of a reality that the various models postulated can now be submitted to a "reality check" as concerns their advantages and disadvantages. Although the buzz is still significant as new questions are posed – What will happen to print? Who will archive electronic materials? Does it cost more – or cost less – to create electronic versions of journals? Will publishers survive? – there is still a clamour on the part of potential users and purchasers of full text to access as much of the various products as possible. And there is nothing quite like a dose of reality to quiet the crowd.

The three basic models of full text, with variations, that seemed to be lining up a year ago are, in fact, the models that are with us today – Publisher-supplied full text; third-party, or Aggregator supplied full text; and Distributed, "linked" full text – in which a bibliographic database provider links to (usually) publisher-supplied full text.

Publisher-supplied full text is something entirely new under the sun with the advent of the World Wide Web. The WWW technology made it, for the first time, feasible from a technology and economic standpoint, for publishers to distribute their materials directly to users. From an economic standpoint, the long-term gain to elimination of the distributor is a "no-brainer" for publishers, particularly the society publishers, whose mandate has generally been to serve their members. From the technology standpoint, the Web put the accelerator on the long-anticipated but comfortably distant specter of the radical change from print to electronic publishing. Unfortunately, the publishers were for the most part really not ready for this speedup and have not, in fact, moved to electronic publishing – they have merely paid lip service to it by converting their print to *some* electronic format for distribution on the web. This is a far cry from *publishing* electronically, and has very definite effects on the product available, as we shall see.

Third-party, aggregator-supplied full text is also a "no-brainer" from the standpoint of the various traditional bibliographic aggregators. Full text is the next logical step in their service, and with or without the Web, these services have the technical know-how and production systems in place to process a large amount of data and the software development teams to provide highly-integrated bibliographic/full text products. They also have a large user-base, which is extremely receptive to the idea of full text. Unfortunately, since the aggregators do not own the full text content, they are at the mercy of the publishers to make agreements with them for full text, and in some cases to actually provide the full text as well. And some publishers, unsure what will happen to their print in this new world, are somewhat reluctant to provide their electronic journals to a vendor, especially if they have rushed online with their own service.

Some third parties have therefore taken another approach, which is to link from their bibliographic data to (usually) publisher full text, since the publisher will hardly object to a new avenue to bring users to their site at no cost to the publisher. This model, from the users standpoint, has the appeal of conforming to what the WWW has already in effect trained users to expect – linking from one resource to another. The linking in this case is a service the vendor can supply to its user-base, and the user-base is spared the idea of having to go to many different sites for access, while at the same time not being restricted to just the titles that an aggregator may put together. Unfortunately, there are no standards yet to make this model easy technically, and perhaps more unfortunately for the long-term, there is no economic model to sustain the vendor – other than the defensive one of preventing erosion of revenues from a lack of a full text offering.

Having looked briefly at the various models, and hinted at some of their benefits and disadvantages, let's look in more detail at how full text issues play out in the various scenarios.

Publisher Full Text

The Good News

Because publishers are the owners of their materials, they are the most perfectly positioned to add value to the print content through the many avenues that electronic distribution makes possible. There are plenty of examples of this already occurring (Academic Press IDEAL, Elsevier ScienceDirect, etc.), as publishers have begun to add electronic-only content to their web sites. Links to auxiliary or back-up materials to research, such as spreadsheets of data, motion pictures, etc. are already appearing at a number of sites. The concept of the journal as an individual, bound unit is falling away as publishers are able to rush certain high-value papers onto their web site way in advance of their "publication" with the other papers in the issue. It is also the fact that the interactive potential of the web is already changing the notion of a journal article, as readers are able to respond immediately to an article and have their response posted in real-time as a "letter to the editor." This means that the "journal" can be changing on an hourly basis as more commentary or review is added.

This trend in electronic publishing has already caused a few publishers to declare that the electronic issue on their web site is the "official" journal of record – not the print. In one stroke, that changes the concept of the library as official repository of the journal literature, and has potentially profound issues for archiving and preservation of the literature long-term. The notion that a library is "buying" something tangible and will have access to it forever whether or not it continues to subscribe in the future, is simply disappearing.

It should be noted that most of the really innovative electronic publishing is being accomplished by the societies rather than by large commercial publishers. This is mostly a matter of scale and mandate – the societies have only a few journals to worry about, and they have a service mandate to their members, which is paramount. Commercial publishers more typically have far more journals and more economic constraints. Many commercial publishers have rushed onto web sites with PDF – a format that is cost-effective but very limited and which does not allow for publishing innovation. It should probably be noted that while PDF has huge popularity among users (mostly for its resemblance to the familiar world of print), it is widely considered by technologists to be an inferior, and possibly temporary, technology. It is widely felt by those technologists that PDF will fade as XML allows for more print-like displays and full publishing innovation. (1) It is also not often realized by “purchasers” of PDF that PDFs created as recently as four years ago often do not display well in the most recent generation of PDF readers (1). Although in some sense the entire issue of archiving is becoming muddled and less sure than ever, those who are comfortably thinking that they’ll be able to read their PDFs “forever” may be in for surprise.

Probably what the innovative, mostly society sites really demonstrate is the future to which all journals will eventually conform, at least in terms of content, format, and volubility. What remains to be seen is how the disadvantages of a completely distributed model will be overcome.

The Challenges

As mentioned earlier, publishers have not really embraced electronic publishing so much as they have scrambled to make their journals available in some electronic format on the web. This means that the publishing activity goes on as before, with all the focus of production, editing & quality-control focused on the print, and then when the typesetting tape is complete it is converted to one of the most common electronic formats – SGML, HTML or PDF. This conversion is actually most typically done not by the publisher at all, but by some third party, usually the organization that produced the typesetting tape. The publisher is charged for the conversion, and the publisher, whose profit margins are squeezed by the additional conversion costs, has no immediate motivation – or expertise – to take on any form of quality control on the conversion operation. Nor do they have any immediate motivation to ensure that the conversion is complete. As a result, the vast majority of publisher full text does not include very complete coverage (just as a result of minimal quality control) and purposely does not include complete coverage when it comes to such things as supplements, which are rarely available at many prominent publisher sites. If the publishers are really going to take on the role of “archive” by virtue of declaring that they will maintain the “official” version on their own site, they will have to rethink the idea that supplements aren’t valuable enough to convert to electronic format.

In addition, publishers are new to the online business, and typically do not have the infrastructure or expertise to run a 24 x 7 online business. With their squeezed profit margins, they have no immediate incentive to hire the expensive staff and software that would truly allow them to plan for the growth and capacity necessary to run an online business. More typically, they will acquire the minimum in the way of equipment and expertise that will allow them to operate on the WWW. This sort of operation will soon enough lead to performance problems both for the software and hardware and last-minute shifting of data on machines – and possibly, over time contribute to the astounding number of URLs that are no longer “live.”

Publishers with their materials on their own web site also unavoidably contribute to the proliferation of information overload that comes with having to remember where/how to access data. Although it may work fine for a few key “must-read” journals, it becomes amazingly difficult to do serious, subject-based research when journals are spread over the hundreds and then thousands of sites that are now available. Universities, who have embraced the full text revolution by scrambling to provide hundreds and even thousands of journals to their constituency, have found that the maintenance burden to locate all the sources; negotiate access under hundreds of different models (publishers are widely variant in how they define, for example, a “site”); establish and maintain links; and generally keep up with thousands of journals – is a huge undertaking. As any serials librarian could tell you, journals – with their constant name and ISSN changes – are not something you can deal with once and forget. And if that’s true in print, imagine how much more true it’s going to be over time with the quickly-changing electronic materials.

A final, obvious but often overlooked problem with publisher full text is the simple fact that if you have to go from publisher site to publisher site to access full text, the body of literature is not searchable. Although individual journal titles may or may not be searchable at any one site, they are certainly not available for searching all together. Since the great majority of literature searches are subject-based, it is a shame to think that for research purposes the content of full text journal literature would be, in its electronic form, as isolated as it is in print.

The Aggregator

The Good News

The most obvious thing that an aggregator can do better than anyone else is to integrate full text both with other full text materials (journals, books, newsletters, directories) and with bibliographic databases. This integration can be at various levels – not just integration through linking, but also through indexing and searching integration. When you can search a large body of full text it is truly amazing the things, which would otherwise be missed, which can now be retrieved. It is often forgotten that 40 to 60% of the literature does not include an abstract – and a large amount of the un-abstracted materials are substantive articles or editorial comment, for which the only access is title searching, or indexing if it exists. It is often also forgotten that indexing often misses items of interest for complete research – including sub-topics, negative mentions, and hot topics for which there is no proper indexing. Although experienced searchers often view full text searching with some trepidation based on prior experience with false drops (particularly when what is wanted is “a few good articles”), the fact is that a wealth of information is being lost in bibliographic searching, even with indexing. Since there are many software techniques that could be brought to bear on the topic of false drops, it would be a real shame to lose the subject-based searching retrieval which an aggregator is in a position to supply.

Beyond simply unifying products through linking and indexing, the aggregator is also positioned to make the next great leap forward in electronic publishing by actually creating new content from old. Obviously there are licensing issues for an aggregator, but the fact is that keeping the wealth of content isolated to individual publisher sites or to mere linking is to miss a huge opportunity to use electronic journals and books and dictionaries and directories to make amazing new knowledge products. Right now the world of literature research looks like a collection of diverse materials which must be searched to find reading material in which there may be an answer to a question. But there is a great opportunity make a world in which questions can be asked directly and answers, rather than bibliographies or articles, can be directly provided. This sort of world requires massive integration of a variety of materials combined with superior software – something that aggregators are in the business of providing.

The Challenges

The biggest issue for aggregators is simply that they do not typically own data and have to fight a constant uphill battle to obtain rights to distribute it. As in the past, an aggregator's entrée into such distribution is superior software, production and execution, full stop. If an aggregator can prove that it can create a product that adds to, but doesn't subtract from, the publishers' profits, and can do it most effectively, or faster, or better than a single publisher, then the aggregator will usually get rights to the materials. Thus far, the aggregator has had a decent leg up on this problem by virtue of historical precedent and a good grasp of ready customers, as well as by, in fact, having expertise. Going forward, aggregators will be forced into ever more innovative software and products to keep ahead of this curve, as the publishers gain more customers and more expertise. One thing that will help is that the aggregators are in a very good position to make connections not only between publisher products, but also between customer's site and the aggregated materials, through various mechanisms such as APIs (Application Programming Interface) and hybrid online/local services. Such services would not scale well for individual publishers and end-users, but work very well for vendors selling to large institutions and consortia.

From the customer standpoint, if an aggregator can aggregate "enough" materials, its product can be extremely appealing, since it can solve all the licensing, access, distribution and maintenance problems associated with the publisher model, as well as provide a seamless interface for the user, the added value of aggregated subject searching, and unique new products. The customer's main problem with this model will be defining what is "enough". Although the customer will probably wish to use both aggregator *and* publisher products, issues will arise when a title is available both from the aggregator and a publisher. The customer – often an institution -- has compelling reasons to want access to the title in both places under different circumstances, but little interest in paying for the access twice.

For this reason, and because publisher products will increasingly diverge from what is made available to aggregators, the smart aggregator will link to publishers for some items – even sometimes to titles they also have on-site, in addition to leveraging the materials they can aggregate into unique new products. Some bibliographic vendors have, in fact, chosen simply to link to publisher sites and avoid the difficulties of product integration altogether.

The Clearinghouse or Distributed Model

If there was ever a “no brainer” service in a world in which “traditional” services are large bibliographic vendors, new full text services are provided by publishers, and the milieu for distribution is the World Wide Web, then that service is linking from bibliographic data to publisher full text. The bibliographic databases are the vehicle by which an entire generation of professionals and researchers have “searched” the journal literature, and the users are perfectly thrilled when these services extend to providing, not just the bibliographic citation but also a link to full text. As any product developer knows, the most eagerly embraced features from an existing customer base are not dramatic new developments, but improvements to comfortable, existing products. And a link from a bibliographic database to the full text is a definite improvement. It is also a nice fit – the publisher model lacks a way to make the body of literature searchable – and the distributed model provides it. A large user-base is already comfortable with using bibliographic databases as the gateway to the literature. The demand for anything more than “more of the same, but better” isn’t really there, so in many cases the distributed full text model supplies exactly what is wanted.

The Challenges

One of the biggest problems for this model, at least in the short term, is that there are no existing standards that make it easy to link from bibliographic Vendor A to Full text Publisher 1, 2, 3 … or 7,000. This surprises many people, who either think that the WWW *is* a standard for linking, or that there is “something” out there like the DOI (Digital Object Identifier) that makes it “automatic.” While it’s beyond the scope of this paper to explain the technicalities, the fact is that linking from Vendor A to Publisher 3 is a one-off task, some aspects of which have to be changed when Vendor A subsequently wants to link to Publisher 22. We are also early enough in the WWW revolution that URLs, the basic mechanism by which the final link is made, are notoriously unstable. So even once a connection is made, the quality and stability of the links can be variable. Initiatives such as DOI or XLink (an XML initiative to make a linking standard) are in such preliminary stages as to make them long-term solutions at best. In the short term, the amount of work to make the links is considerable and the stability is not always great, even though it’s wonderful when it works.

The more interesting challenge for this model is to come up with a long-term, viable economic model that doesn’t depend on “toll-taking” the user has to be aware of. Right now, the vendors that link to publisher full text generally do so at their expense, for no direct monetary reward other than to retain customers who might otherwise drift away to services that could provide full text. Licensing and access “permission” is generally left to the customer and publisher to work out completely apart from the vendor. This aspect, as with the publisher model, can soon leave the poor customer with an administrative hassle of some magnitude.

Although the idea of a vendor providing a service at their own expense may sound lovely to the users who are currently benefiting, it is likely that a product or service that provides no economic benefits to the provider will eventually change into something else. What that something else might be is a very interesting question, and one that won't necessarily have a wonderful answer, at least from the customer's perspective.

As with the publisher model, where searchability of the overall journal literature is sacrificed, in the distributed model, searchability of the journal literature is artificially frozen into a model that is over 25 years old. Not to belabor the point, but there is real, provable value in the ability to search aggregated full text rather than just bibliographic citations – even indexed bibliographic citations. The distributed model also puts off, for some far-away time, with much faster bandwidth, much better commerce software, more robust standards, and more advanced text software, the time when the various kinds of literature – dictionaries, journals, textbooks, directories, etc. can be combined in creative and amazing ways into new products. Things that could otherwise occur today could be put off for decades.

Summary

It's no wonder there is such a clamour for full text! What was once a dead-end process, searching a bibliographic database electronically only to have to trudge to the library to access the full text, is suddenly, tantalizingly, within reach – but comes with a plethora of issues and tradeoffs. At this point in the game, it would behove no one to declare victory in creating the perfect model for full text. Many of the ramifications and issues associated with the various models are not widely understood, and for those issues that are understood there is often no clear resolution – only uncertainties. Since the current purchasers and users of full text are going to help determine the future course of electronic publishing just by the purchase decisions they make today, as clear an understanding of the issues and tradeoffs as possible may help minimize future disappointments. And in the meanwhile, the many voices raised in the bid and clamour for full text will be making informed, practical decisions which may not ultimately be perfect, but which are enriching and exciting in ways only dreamed of as recently as a few years ago.

[1] Trafford, Ben. "XML, Open eBook and Other Standards", *Formats for Electronic Publishing, Descisions & Trends,* LITA Electronic Journals Publishing Interest Group, ALA 1999, June 28, 1999, New Orleans.