

Hal Berghel

Value-Added Publishing

Without question, electronic publishing is one of the hottest topics in computing. Groups worldwide want to know how to do it well, how to advertise it effectively, how to enhance the capabilities of electronic publishing to include emerging multimedia technologies, and, most of all, how to make money at it. In the future it will be increasingly important for successful publishers to add value to publications over and above the original content.

This column outlines some of the fundamental issues connected with the addition of value to electronic publications. Some of these issues have already been translated into products and services, while others have not.

Information Delivery in the Gilded Age of Computing

While the term electronic publishing takes on a variety of different meanings in different settings, one core principle holds true across all domains: electronic publication involves the distribution of digital documents. In its simplest form, electronic publishing may amount to little more than a “porting” of printed information over to the digital networks via scanning, OCR technology, and so forth. Augmented with some very basic accounting software, many pub-

lishing sites are in the business of serving up static HTML versions of their publications via the Web. In its more complex forms, however, electronic publishing will redefine itself in the light of available computer and network technologies. Our present goal is to outline the ways in which this

redefinition may be achieved.

Although most of the critical technologies needed for electronic publishing have existed for decades, it has only been in the last few years that traditional, in-print publishers have taken it seriously. There were two basic reasons for this delay—one technological and one pragmatic. On the technology side, the primary intended venue for electronic publishing, the World-Wide

Web, lacked two essential capabilities. First, it lacked secure HTTP transactions until about 1995. Without secure transactions, selling via credit cards would entail excessive risk due to digital eavesdroppers, packet sniffers, and other network nematodes. At the same time, there were no widespread standards for, and implementations of, electronic billing systems.

Means had to be developed to charge in millisecond amounts and accumulate charges until they reached cost-effective invoicing limits. These two pieces of technology were in place (in a variety of different forms, in fact) around mid-decade, thereby making the simpler forms of electronic publishing possible.

On the pragmatic side, no one knew (in fact, it could be argued that it's still unknown) how to develop a sound business plan for electronic publishing. While it was widely assumed that adding electronic publishing products would irrevocably change the economics of publishing, few felt comfortable in speculating whether this would ultimately be good or bad for the industry. Many publishers jumped on the electronic publishing bandwagon for the worst of reasons: they were afraid of being left out of the future markets. In so doing, they packaged intellectual property basically the

same way as Gutenberg did except for the addition of digital delivery mechanisms.

That is the primary cause of the popularity of digitizing anything and everything in print; herd mentality dictates that if you don't have a good plan, do what everyone else does or thinks they should do.

In any case, all of the essential hooks for electronic publications are now in place. Advanced publishers can solicit, edit, produce and distribute electronic publica-

tions with not so much as a single piece of paper changing hands (not including signed copyright transfer forms and contracts, of course). The Web and the Internet has forever changed the face of publishing. But is this for good or ill?

THE BIGGEST MISCONCEPTION about electronic publishing is that its value lies in the ability to disseminate digital information over computer networks in a manner analogous to physical

distribution of hard copy. There seems to be a tacit faith in a twisted variation of Metcalfe's Law (the value of the Internet increases with the square of the number of nodes) to the effect that the value of electronic publishing increases with the square of the number of documents on the Internet. While this sounds good, it's likely to be false. It is more likely that the value of electronic publishing varies inversely with the square of the number of documents.

Potential Metalevel VAP Enhancements

Confidence Indicators

- Listing as citation classic by authoritative source
- Document status indicator (preprint, archived, old, not recently viewed)
- Awards received (weighted by importance, source)
- Reviews of document in the literature
- Referee reports from peer reviewers
- The perceived quality of the imprimatur
- Vetting by some community or constituency (praised by newsgroup, professional association, anthologized by reputable editors, and so forth.)

Recommending Systems

- Community review systems (Firefly [www.firefly.com])
- Helper agents
- Information "brokerage" to facilitate connection (by vendor/brokers, fulfillment agents, aggregators, and the like)
- Hyperlinked review chains (which interconnect all reviews of a document irrespective of source)
- Amalgamated or virtual reviews (which merge elements of individual reviews over related documents)
- Virtual editors ("personalized" variant of an electronic publication created by someone other than the author)
- Searching, indexing and database technologies

More Interactive and

Participatory Than Current Systems

- Provide dynamic, real-time document clustering

with innovative clustering topologies for display of results

- Preprint servers for preserving the ancestry of documents
- Postprint (archive) servers for maintaining definitive versions
- Data mining, including techniques based on association, sequence-based analysis, clustering, classification, estimation, fuzzy logic, genetic algorithms, and neural networks
- Data warehousing and data repositories (the ACM Computing Research Repository (www.acm.org/corr/) and ACM Digital Library (www.acm.org/dl))

Document Persistence Technology (www.sciam.com/0397issue/0397kahle.html)

- Formal methods for post-hoc data utilization (which structure data differently or anticipate new data demands)
- Cyberspace snapshots that provide backups of documents whose links are fractured
- Version archiving strategies for citation permanence
- Variable-Link-Strength Technology Based upon Frequency of Use Statistics or User-Centered Evaluations
- Frequency of access and average visitor ratings of a site
- Detection of the number of inbound links to a particular site

Digital Village

This misconception has driven virtually every publisher into some form of electronic commerce. Nowhere is this more obvious than with academic and scholarly publications. Seen as a way of mitigating against the problem of slumping sales and an annual 5%–10% downturn in subscriptions, electronic offerings are thought to hold out the greatest promise of revenue growth—even a modest 5%–10% annual growth. But this reasoning ignores the fact that the decline of the academic

publishing industry is inextricably linked to the overall economy, the widespread perception that there is already too much information available for most personal bandwidths, and the perception that only a small percentage of the information in the typical publication is relevant. Readers are, therefore, “voting” with their pocketbooks by canceling subscriptions. Publishers worldwide are assuming electronic publishing is the silver bullet that will save the day. Some publishers

point to the capabilities of the networks to lower overhead and production costs, support a wider variety of advertising and marketing venues (for instance, broadcasting, narrowcasting, and personal-casting), and the ability to increase margins by dealing directly with the reader rather than distributors and middlemen as signs electronic publication will provide new opportunities for publishers seeking to turn their fortunes around. In other words, some

Document Persistence Systems which Help Ensure the Longevity of Linked Resources Especially with Respect to Mission Critical Environments (medical information systems, patents, copyrights, commerce)

- Revision control systems/version retention systems
- Web “snapshot archives”
- Version validation systems

Virtual Authoring

- Virtual documents (such as process-oriented document creation systems in which documents have no reality apart from current presentation)
- Dynamic contextual annotation added by authors and readers (like “pop-up” videos on cable television’s music channel, VH1)
- Trans-publishing in Ted Nelson’s sense (www.sfc.keio.ac.jp/~ted/)—where documents take on hyperstructure as they evolve in a structured way by inclusion of different authors and participants (projects Xanadu [www.xanadu.net] and ZigZag [www.xanadu.net/zigzag/]).
- Group authoring technologies (perhaps an outgrowth of computer-assisted, cooperative work, groupware)

Dynamic Document Creation (in which the documents are revised continuously)

- Author-revision systems (Stanford’s Encyclopedia of Philosophy [plato.stanford.edu])

- Author and reader revision systems (the good, the bad, and the ugly site www.acm.org/~h1b/email_gbu/)
- Thought swarms and “idea structuring”
- Online ACM Computing Reviews (in development)

II. Information Customization (see www.acm.org/~h1b/publications/cb5/cb5.html)

- Client-side document extraction
- Non-prescriptive, non-linear document traversal (not prescribed by document provider)
- Multi-document “collage” interface for multi-way look-ahead
- Related emerging technologies which will support value-added publishing
- Safe, open distributed archiving (Alexa [www.alexa.com])
- Ted Nelson’s transcopyright system (www.sfc.keio.ac.jp/~ted/)
- Security enhancements
- Watermarking and digital steganography (www.acm.org/~h1b/publications/dw_n/dw_n.html)
- Push technology (www.acm.org/~h1b/publications/push/push.html)
- Citation tree construction
- Agent-based citation locators (www.uark.edu/~iarg/)

publishers are working under the assumption that the decline in interest in scholarly and technical publications can be reversed if just those publications could be produced and marketed more cheaply electronically. It just won't work that way; the publications avoided in hardcopy will be avoided in electronic form as well.

If the digitization of things publishable won't get us far, what will? The payoff in electronic publishing in the future will be the deployment of new technologies for the integration of digital documents into the network fabric of associated ideas, texts, times, and people. Publishers will need to be more than just the providers of digital documents from their digital warehouses. They will also need to connect a document with its contexts. Thus, a digital document could be tightly integrated into the cybersphere of all related documents in a way that traditional publishing cannot permit. Such publishing could provide not just the documents, but their connections to other data sources, as well as other valuable information. This is the essence of "value-added" publishing.

VALUE-ADDED PUBLISHING (VAP) is a natural extension of traditional publishing with the additional feature that the publication vehicles and venues accept from and react to additional, previously integrated and assimilated networked media. The challenges of VAP are likely to lie in such areas as:

- Content enhancement
- The encouragement of synergy

between and among information providers, information consumers, and the resources they share

- The addition of interactivity and feedback loops to traditional delivery systems
- A reorientation of both the information provider and information consumer toward the "process" of publishing, rather than a focus on the individual products and services
- Metalevel analyses and intelligent restructuring of document collections
- Ad hoc document quality ranking and recommending systems

As one can see from this partial list of services, VAP must use a more advanced set of computational and network tools from that of its early electronic publishing ancestors. We illustrate these points with a selected enlargement of some of the aforementioned categories.

Content enhancement. One convenient way of viewing electronic publishing is as the exchange of information between an information provider and an information consumer via an intervening computing network infrastructure. While the content of a document is central to this exchange, it is not necessarily paramount since its value is utilitarian rather than intrinsic. That is, the value of the content is not independent of the ability of people to read it, view it, use it, reference it, and so forth. From the point of information retrieval, information which cannot be found or used is worthless.

Content enhancement involves the study of enrichment of the semantic and syntactic content of a document. The

enhancement of semantic (alt., conceptual, deep) content can be thought of as an attempt to extract more meaning from the documents. A report, summary, extract, abstract, translation, or "gist" by an intelligent agent would be considered a semantic enhancement in this sense, as would results reported by natural language understanding and translation systems, and the automated inclusion of new hyperlinks.

The enhancement of syntactic (alt., grammatical, tag-based) content, on the other hand, would affect the way documents are structured, indexed, taxonomized and linked within the intervening network and computer resources. An example of enhancing syntactic content would be adding structure to documents for the benefit of helper agents, search engines, indexing tools, data mining, and warehousing applications.

Value-added metadata.

While content enrichment of electronic publications is the holy grail of VAP, it is at the same time the most difficult strategy to implement. Some problems (complete natural language understanding, for one) are intractable given the current state of the computationalists' art. Adding value through metadata, while less ambitious, holds out much greater promise in the short term.

Metadata is information "about" an electronic document, resource, or the operation of a computer system. For example, "confidence indicators" might provide useful information about a document or resource. We would expect that knowing an

Digital Village

electronic publication produced a Pulitzer Prize would increase the credibility of the author and the value of the document (at least as an object of study), as would favorable reviews by the leading authorities on the subject, for example. The imprimatur of a publication might also be relevant, as some electronic publishers might be known to have higher standards than others.

Similarly, recommender systems assign assessments or recommendations to documents and resources that are as reliable as the confidence one has in the recommender system. Helper agents, brokerage systems, flash lists, and so forth also provide metalevel value in their evaluation and recommendation of documents.

Revision control systems, which collect metalevel information about various versions of a document, add value by helping create stability and continuity in network documents. In these systems, versions of documents are indexed in such a way that any particular version may be retrieved, with or without its predecessors or ancestors.

The sidebar illustrates the types of enhancements that might result from the judicious collection and use of metalevel information about electronic offerings.

Feedback, Interactivity, and Support

Content-based and metadata-based value-adding are two of the four strategies for building value in electronic publications. We add to the list two more components: (1) feedback-based value-adding and (2) interactive value-adding. Services of this type collect data from users that

reflects their perceptions of their experience. Out of that collective experience might come useful comments, identifications of “hot” documents by some measure of use, average rankings of sites, group interactions, and so forth that will speak to the issue of the perceived value of content.

To this we must also add support-based value-adding—technologies that may not directly add value to a document, but that support the addition of value by other means. In other words, they are necessary conditions for the deployment of a VAP system. This might include database technologies, statistical and clustering tools, revision control system software, editing tools, information customization clients, and so forth.

Electronic publishing in the next century will be fundamentally different than it is today. I predict the most successful, early applications of VAP will be such things as:

- Publications with limited commercial appeal
- Publications with narrow audience appeal
- Digital digests (i.e., personalized magazines assembled from many sources)
- Focused retrieval publications (personalized encyclopedias)
- Home-grown, personal publications
- Interactive publications
- Public interest/public awareness publications
- Reference materials

Electronic publishing will evolve as developers and researchers are inspired to take

more extensive advantage of computing and network technology, and slowly but inexorably move away from the notion that the paramount value of a document is its content. Additional enhancements such as those outlined in this column will establish the importance of the role of the digital or cyberspace context of information.

Many of these thoughts have evolved as a result of my serving nearly six years on the ACM Publications Board. By continually revisiting the questions of what we were doing and why we were doing it, this conceptual overview of the future of electronic publications began to take shape. The launch point was my belief (controversial, as it turned out) that ACM should move away from the policy of holding copyrights for its publications

(www.acm.org/pubs/copyright_policy/). I remain convinced that trying to fix one version of an electronic publication as definitive and copyrightable will prove as difficult as trying to paint falling leaves. In my view, electronic publications of the future will resemble filmstrips—each frame will incorporate some improvement, alteration, or reference which (in the ideal case) will have more value than its predecessor. In this sense, Ted Nelson’s notion of transpublishing is much like many layers of intersecting filmstrips, each of which has one cell that aligns with the cells of others. **■**

HAL BERGHEL is a professor of computer science at the University of Arkansas and a frequent contributor to the literature on cyberspace. His virtual home is www.acm.org/~hbl.