



Archiving

Table of Contents

Why Archive?
Archiving Issues
Collaboration
Making Decisions
Technical Options
Business Considerations
Transitioning to the Future
References
Glossary
About the Authors

Archiving**From a Publisher's Point of View**

A White Paper Prepared for The Sheridan Press

A Restricted Document Distribution

The Sheridan Press has commissioned this paper as a service to its customers and to the publishing community as a whole. Reproduction without the prior written approval of The Sheridan Press is prohibited.

Publisher

The Sheridan Press

Contributing Editors

Barry Davis

Fred Fowler

Chris Gohn

Kevin Pirkey

Craig Rineman

Greg Suprock

Joan Weisman

Authors

Barbara Meyers

Linda Beebe

Copyright 1999, The Sheridan Press.

Second Printing 1999, The Sheridan Press.

ARCHIVING FROM THE PUBLISHER'S POINT OF VIEW

The best way to predict the future is to invent it.

—Peter Drucker

WHY ARCHIVE?

Publishers of scholarly information formerly gave little thought to the future access to their products, because libraries traditionally ensured the indefinite storage of published materials. This situation is evolving with the increased electronic distribution of information and the development of publisher site licenses and agreements that do not always provide libraries with ownership. Now both publishers and librarians face the problem of how access to information will be maintained into the future.

Publishers also have the opportunity to repurpose their information content as new technologies offer the potential for creating different products from original content. Publishers need to plan for multiple uses from the beginning and to produce data in a portable form. For existing products, publishers will need to select a method of archiving that is the most likely to result in multiple use. Yet, if publishers are going to maintain any data over time, they will have to migrate it to new media as technology changes.

Publisher Questions

This paper views the issues of archiving from the publisher's perspective. It considers the basic questions of why publishers should be concerned with archives, what materials publishers should archive, and what technologies publishers can use to archive the scholarly information they produce.

Brichford and Maher (1995, p. 701) set forth a concise definition of archiving: "Archives are retained information systems that are developed according to professional principles to meet anticipated demands of user clienteles in the context of changing conditions created by legal environments and electronic or digital technologies."

Archived materials are kept for their long-term value. Most scholarly publishers would be hesitant to suggest that the information their authors produce has only ephemeral or short-term value. However, the act of merely preserving information in its original form is no longer sufficient. With the challenges of sustaining scholarly information over the long-term, publishers must now make critical format decisions that will allow for the use of equipment yet to be developed.

Print books and journals need no additional instruction past purchase. After all, the school system taught us how to read and understand the printed word. But, as the Task Force on Archiving of Digital Information (1996, p. 2) reported, "reading and understanding information in digital form requires equipment and software, which is changing constantly and may not be available within a decade of its introduction." Therefore, publishers concerned about access face requirements for archiving that are fundamentally different from all that they have known in the 500-year history of print publishing.

Throughout the archiving literature, librarians have enumerated the myriad of external threats to the preservation of publications. Traditionally, libraries have coped with threats to preservation. Now, using electronic media, publishers must join librarians in the concern about ensuring continued access to information.

Publisher Questions

Why be concerned?

What should be archived?

How should it be done?

Gathering and Selection

It is important that scholarly publishers think about archiving when they create their products rather than considering it an activity that occurs after the fact. Publishers must augment their traditional gathering and selection of material with an increased understanding of the value of the intellectual content over time. They need to develop editorial and administrative policies to preserve their texts in the new digital environment.

Publishers will become capable of ensuring future access to an author's work or risk losing authors who are concerned with the long-term accessibility that implies permanent value. In addition to the publisher's imprint and dissemination capabilities, enduring accessibility will be an increasingly important author criterion. Hunt and Wegner (1996) observed the necessity to involve the end user in the design and production of search and retrieval systems. Publishers face the same necessity in developing digital information products for future use.

Publishers must recognize the challenges of enduring access to information in the digital age. Mandel (1996) outlined three new challenges: (1) the technology is impermanent; (2) licensing agreements do not give libraries the legal right to maintain information in perpetuity; and (3) the life cycle of digital information is undetermined.

Three New Challenges

Impermanent technology
Legal rights to maintain information in perpetuity
Uncertain information life cycle.

Refreshing Data

To assure future access to both paper and digital formats, publishers need to preserve the substance of their information, maintain sufficient context that future readers will understand the information, and provide the means to access and use the information in years to come. Mandel (1996, p. 456) noted the need to authenticate the “version that is selected for preservation.” For the short term, publishers must acquire the knowledge and staff to provide for the storage of data along with the capability to maintain and deliver it. For longer-term retention of files, Mandel counseled that additional techniques will be required: Publishers will need to refresh the data and migrate it to new platforms. Thus, in addition to preserving content, publishers must preserve functionality.

Responsibilities of Publishers

As Mandel (p. 456) noted, “archival preservation via migration requires a commitment to unknown future activities with unpredictable future costs.” The activities related to access and preservation cannot be determined, nor can the costs be borne, solely by the library community. Publishers need to take a more active role in determining whether and how well their materials will be preserved into the future.

The global explosion of information has galvanized libraries and archives to redefine what Marcum (1996, p. 452) called “the preservation imperative,” that is, preservation as a “fundamental responsibility of libraries and archives of record.” Marcum noted that this imperative, which has been insufficiently provided with print materials, is more complicated in the digital world.

ARCHIVING ISSUES

The expectation in the 1990s is that, along with libraries, publishers will also maintain archives in some fashion that will permit access in the future. Consequently, publishers now face the issues that librarians have been grappling with over time. At the same time, creating archives can enhance the potential for additional revenue from a product.

Paper Products

The fragility of print publications has received considerable attention since the U.S. Commission on Preservation and Access documented the “brittle book” syndrome in 1968 (Marcum, 1996). In general, the threats to preservation of paper can be classified as natural (such as light, moisture, excessive heat), organic (such as rodents and insects), and chemical (such as acid content, air pollution, and plastic). Careless storage with rubber bands, paper clips, or tape, as well as poorly designed receptacles, also contribute to deterioration. Mishandling can be due either to carelessness or maliciousness.

Electronic Products

In theory, we could learn from the exhaustive research on preservation of the print medium and apply the same principles to preserving electronic forms. However, in addition to all of the threats outlined above, there are other, perhaps even more complex, issues related to maintaining electronic products, such as obsolescence and data degradation.

Obsolescence

Brichford and Maher (1995) noted that, unlike paper and photographs, which remain accessible even in advanced states of decay, electronic products can be examined and used only if the hardware and software that deliver them remain available and operational. The rapid evolution of formats and equipment threatens a very short usable life span for many products, unless the producer plans for ongoing access. An electronic product can be preserved successfully only if the information is encoded in a format independent of the hardware and software that is used to produce it *and* if some software exists to manipulate it in its current form (Marcum, 1996).

Threats

Obsolete hardware
Obsolete software
Data degradation
Different versions

Dynamic Data

Corruption of data or data degradation is a concern in archiving. And the ease with which an electronic product can be altered, either by accident or intent, is a serious threat to preservation. Can two researchers be sure they are looking at the same document? Which version exists in this locale? Graham (1995) noted at least three kinds of potential change: (1) accidental, such as losing data or saving the wrong version; (2) well-meaning intentional change, such as new versions, updates, or the changes in an interactive document; and (3) fraudulent intentional change, such as changing evidence in one's own work or altering another person's work.

Access

Archivists have thought of access as an enemy of preservation. The more a product is used, the less likely it is to remain in good condition. This dichotomy between preservation and use is not generally the case for electronic products. Publishers will need to consider how information will be used and how it will be accessible. Simply retaining copies in storage will do nothing for knowledge development or for ongoing profits. Publishers need to consider what their readers will look for. Graham (1995) suggested that readers will continue to want what they expect today: Reliably locatable information that is easily accessible and immediately retrievable.

Copyright and Finances

Copyright and finances are two key business issues for publishers in planning their archives. For copyright, publishers must determine how they will meet the requirements of fair use while they protect the author's work and their own investment. Publishers must safeguard their rights, not just for the original work, but for any derivative work they might produce. And their responsibilities to protect the author's authorship (Gotze, 1995) remain in force.

Creating and maintaining archives requires a substantial financial investment. However, the cost of not archiving may be even greater if the publisher loses the opportunity to get maximum return from the original work. Perhaps, instead of looking only at the cost of archiving and refreshing the data, publishers should focus on developing a business plan to enhance revenue from original and derivative works, once they determine what should be archived.

What is past is past. The future, we hope, is going to be longer than the past; therefore, we need to be in a position to take advantage of it.

—Tonkery, 1995, p. 68

COLLABORATION

Information aggregators and technologists (such as computer programmers and systems designers) have joined the traditional players in the scholarly communication process—publishers, librarians, suppliers, authors, readers and users. The functions of all these players were distinct historically, but that may no longer be the case. It appears that players in the scholarly communication process will be taking on extensions to their traditional roles and will need to become more aware of each other's activities.

Brichford and Maher (1995, p. 709) outlined several of the changes each will need to address: “Authors will need to consider future accessibility of their work when they choose a publisher. Publishers’ responsibilities will have to go beyond the traditional role of simply distributing texts and instead extend to providing for the disposition of texts. Otherwise they will be delivering products of increasingly dubious utility because of their rapid obsolescence.”

“Librarians and archivists can play an important role in new relationships with publishers and information providers. Both are aware of the ways people use information especially for purposes well beyond those envisaged as the audience by the author or publisher. In a more structured relationship with publishers and authors, archives and libraries might serve as depositories for either hard copy ‘dumps’ of electronic texts or as predesignated repositories for access systems that will enable long-term research access to electronic texts. With careful up-front planning involving publishers, their boards, librarians, and archivists, the mechanisms to allow long-term access can be established.”

Marcum (1996) reinforced these concepts by describing an ideal environment in which all the players would work together. In her as yet unrealized scenario, publishers would design materials for long-term access in standard formats with documented software and metadata. Further, they would deposit the digital resources in appropriate repositories and accompany them with agreements for long-term preservation and access.

Partnerships and Alliances

Publishers will begin to see the value of creating new alliances with their suppliers and their consumers. They are likely to forge alliances to create specific products rather than negotiating mergers of organizations. Publishers will draw upon the expertise of other new players in the scholarly communication process, such as information aggregators, to augment their own skills in developing products that can be sustained and used in the future.

Now more than ever, publishers must keep all users, institutional and individual, in mind; otherwise, their products will have no permanent value. Thus, in addition to staff with traditional publishing expertise, publishers will hire members of the new technology communities, either directly or through new alliances with aggregators, librarians, and suppliers.

Librarians, as Grycz (1995) observed, have always provided archival services, but will develop new approaches in their archival role. One approach is expected to be increased collaboration and cross-industry communication. Librarians will provide input about archival issues early in the publishing process.

Recently, suppliers of traditional book and journal manufacturing services have also expanded their skills and expertise to support publishers in creating products with the greatest potential for sustainable access in the future. Future archives may not reside only in libraries. For example, as publishers assess their positions for the future, we are already seeing suppliers providing archives of digital publications.

New Alliances

Publishers & librarians

Publishers & suppliers

Publishers & aggregators

MAKING DECISIONS

Not everything warrants archiving. For example, a work may be a transitory one that is being replaced by a more complete version. If so, archiving the first version may have little value and providing access to it may actually be confusing. (Witness the multiple versions of articles that authors leave on the Web.) Lievesley's rejection criteria for archivists (1995) may serve publishers as well. Those criteria include "low likelihood of usage," "poor quality," "inadequate documentation," and "confidentiality problems." The contents of many electronic bulletin boards, email conversations, and informal publications may not justify an investment in archiving.

What to Archive

Substantive bodies of work such as the contents of journals, reference works, and books are the most likely candidates for archiving. For these types of works, the publisher has already made selection decisions and enhanced the value of the original work by editing, designing, producing, and marketing it. More ephemeral communications may warrant archiving if they are well documented or can be combined with others to develop a new product with added value.

In determining what to archive, publishers may want to consider a series of questions:

- Does the work contribute to knowledge development?
- Will it continue to have intellectual value in the future?
- How has the audience responded to the totality of the work?
- Can archiving in electronic form add value to the existing work?
- Is there likely to be an audience for the work in another format? Are secondary or parallel products likely to meet audience needs?
- Is there potential for combining this work with another to create a more expansive and useful product?

If the answers to most of these questions are positive, the work probably should be archived.

Reasons to Archive

The most obvious reason to archive is to preserve valuable information. Primary journals, in particular, have a responsibility to maintain their entire body of literature so that all the research published in any given discipline is accessible to future researchers. However, other factors reinforce this need for archiving. Once information is captured in digital format, it is available to be revised, reformatted, and delivered in different media to multiple markets with different needs.

Publishers struggling to meet the needs of an increasingly diversified market can take advantage of their original investment in digitizing data by producing specialized subsets of the first work. Regardless of the product or its intended derivative work, updating is significantly easier and less expensive working from a digital version.

How to Archive Data

The publisher will consider several factors in determining which container or carrier to use for archiving. How large is the work? How frequently will it be updated? Will the material be kept internally and used only to develop derivative works or will it be delivered in its archival form to customers? How will the audience use the data? Publishers must also consider their total product, not just the text, but graphics and page images as well. They have their choice of several options for formats.

Carrier Options

Diskette/Tape

CD-ROM

Online

Offline

Diskette/Tape

The advantage of diskettes is that everyone who has a computer is a potential user. The disadvantage is that a diskette holds relatively little data, only up to 1.4 megabytes of information. Furthermore, diskettes are platform-dependent; publishers must produce them in Windows, Macintosh, DOS, and UNIX formats. In mainframe environments, the medium used is magnetic tape.

CD-ROM

CD-ROMs provide an answer to the two drawbacks of diskettes. They can handle all platforms on a single disk, and they hold much more data—640 megabytes. In fact, few products fill a CD-ROM to capacity. The market potential is also large and growing daily, as nearly all desktop computers in 1997 are shipped with built-in CD-ROM drives. Since the mid-1980s, pundits have described the CD-ROM as a transitory technology and have predicted its demise. Yet the technology continues to be the most widely used because of the CD's large capacity, multi-platform capability, and significant market penetration. CDs present several disadvantages to publishers relying heavily on the library market. Libraries lack the storage and player space to handle all of the CD-ROMs they have collected. Further, librarians are concerned about multi-user access. Unless CDs are offered in a network environment, librarians will prefer online access.

Online/Offline

Delivering content on the World Wide Web (WWW) eliminates all concerns about platforms or storage for the library and the user. Although fewer people have access to the WWW than own CD-ROM drives, the market is growing rapidly. However, there are few successful business models; the ability to charge sufficient fees and collect them is still problematic in 1997. Publishers may choose to store more current materials online (client server) and keep the older ones in offline (disk or tape) storage, as the MIT Press has done with *The Chicago Journal of Theoretical Computer Science* (Fisher, 1995).

TECHNICAL OPTIONS

Text Formats

Although technology has taken us a long way from when retyping was the only option if authors did not use the same word processing program as their publishers, text formats still remain an important consideration. Today most text formats are platform independent, but they offer different levels of robustness (strength of construction) and flexibility.

Text Formats

ASCII
PDF
SGML
HTML
XML

ASCII

Files may be produced using either ASCII format or binary. ASCII (American Standard Code for Information Interchange) encodes plain text or data with none of the formats added in word processing, graphics, or data files. The format uses seven of the eight bits in a byte to define codes for 128 characters. Originally developed for use with Teletype machines, ASCII is frequently used for transferring files across the Internet. However, the files must then be converted to a binary format for display or use. ASCII, therefore, works well for transmitting data, but does not on its own supply the visual requirements nor does it provide linkages. Because it is a universal language, ASCII is used in other formats, such as SGML.

PDF

PDF stands for Portable Document Format, which enables the creation of electronic documents that retain all of the original formats, fonts, and layouts of the print product. Produced by Adobe Systems as a derivative of their PostScript page description language, the platform-independent format is marketed as Adobe Acrobat. The flexible format allows printing in full PostScript and searching by keyword. Some observers have noted that PDF is on the rise with PostScript files sent to composition going into PDF. High resolution files are needed for printing and archiving, but low resolution files are needed for Web publishing.

The format also enables cross-document links. For example, a publisher could deliver a single set of hyperlinked PDF files that work on local drives, network drives, a CD-ROM, and the WWW. Users on the WWW must download the Adobe Acrobat Reader to print the PDF files; however, Adobe has created plugins to Netscape Navigator and Microsoft's Internet Explorer that allow users to view a PDF document in the browser window. A disadvantage is that PDF files take longer to download.

Markup Languages

In all electronically-produced documents, codes are embedded to store the information needed to process the document, that is, to establish the font, the style, or even the structure (Marchal, 1996). Embedding those codes constitutes the “markup,” just as editors used to hand-write instructions to the compositor. Typically, electronic codes are proprietary, subject to frequent change, and inadequate for multiple linkages. With proprietary coding, there is the danger of obsolete programs that may result in expensive conversions (Donovan, 1997). Consequently, there was a need for a “generalized” markup language that would be process and platform-independent.

SGML. SGML is the abbreviation for Standard Generalized Markup Language. The International Standards Organization accepted it as a standard (ISO 8879) in 1986. A meta-language, SGML is a parent to other languages such as HTML.

The ISO version of SGML is a generic international format in ASCII used to define methods of representing electronic texts (University of Virginia, No Date)¹. The format allows producers of various kinds of documents—books, journals, dictionaries, even letters—to describe their structure and mark them up appropriately (Marchal, 1996).

A key component in the system is the DTD or Document Type Definition, which is the blueprint for the structure of a particular type of document. DTD is the set of rules that a publisher uses to define the logical organization of the document and to describe such things as element or tag names, special characters, and element content and order. Elements are specific parts of the document such as a paragraph, title, or list, which are the building blocks of the document. Tags are characters that identify the start and end of an element. Maintained in a separate file from the document itself, DTD provides a checking step not found in other systems (Comstock, 1997).

An advantage of SGML is that it allows publishers to move their data among systems, platforms, and software programs. This portability greatly enhances the value of the data. Publishers can easily produce various formats from one set of data; therefore, they can produce derivative or repurposed works more cost-effectively. SGML eliminates document interchange problems, and it reduces production costs and time. Once they have revamped their systems for SGML, compositors generally have found that it improves the accuracy of their files.

A disadvantage of SGML is that producing a DTD can be extremely labor intensive, which translates to expensive. Therefore, many experts recommend that publishers avoid producing their own if they can and follow the ISO 12083 standard for book, article, and serial DTDs. Success in delivering and archiving information will require that publishers do not view DTDs as a matter of house style as they do the design for a book or a journal. If every publisher uses a different DTD, the hopes of ubiquitous access in the electronic environment will not be realized.

SGML Characteristics

Platform-independent

Portable

Based on a DTD

Complex & powerful

Robust

Potential derivative works

HTML. HTML (HyperText Markup Language) was designed to send tagged information over the Internet to different computers with different software (Horrocks, 1996). It is the primary format used for documents on the WWW. Constituting a relatively simple set of rules, HTML helps define parts of a document such as headers and paragraph boundaries so that a user's Web browser can display them appropriately. HTML has several advantages: It is simple to learn, it transmits quickly, it is relatively powerful, and there are many software tools available to help beginners set up WWW pages.

HTML also suffers from disadvantages. Critics have noted that, because it was created without a DTD (Document Type Definition), the original and subsequent versions have suffered from a lack of the structure a DTD imposes. Its weak formatting capabilities have frustrated many producers, and Greenspun (No Date) noted that it is the "worst of two worlds" in that it produces "ugly documents without formatting or structural information." The user's browser will display the content in the way it interprets the tags; consequently, there is no uniform delivery of design. The publisher cannot control the fonts, either the format or the size, and the line length is not fixed.

Publishers need to be certain that the appropriate graphic callouts are embedded so that the correct image appears on the screen. HTML permits publishers to insert a thumbnail sketch that loads quickly and when clicked expands to full size.

A disadvantage of HTML is that it is not viable for long-term storage because it has so few tags and lacks the robustness needed to create secondary and parallel materials.

In July 1997, the World Wide Web Consortium (W3C) announced the first working public draft of HTML 4.0. Their press release (Weber Group, 1997) quoted Dave Raggett, lead architect for the activity, announcing that "You get much greater control over forms, frames, and tables, and all the benefits of scripts, style sheets, and objects." At this writing no comments had been circulated.

HTML Characteristics

Platform-independent
Relatively powerful
Simple to learn
Weak format capacity
Not robust

XML. W3C members released a working draft version 1.0 of XML, the Extensible Markup Language, in June 1997 (Bray & Sperberg-McQueen). XML was developed to permit SGML to be served, received, and processed on the WWW just as HTML is.

As in SGML, the layout and logical structure of a document (DTD) are specified separately from the content or markup. The language is fully described in Bray and Sperberg-McQueen. Essentially, XML is HTML with user-defined tags.

Graphic Formats

Since publishers began their foray into the world of electronic publishing, one of the major challenges has been to capture, display, transmit, and retrieve graphic images. The need for graphics, and the quality of graphics that are needed, varies greatly across disciplines. Humanities texts with few graphics do not pose the same challenges as the scientific, technical, and medical publications with their need for visual depictions. Yet any information product demanding high-resolution graphics presents new challenges for the publisher in the environment of digital access and archiving.

In observing problems of access for existing electronic archives, Law (No Date) observed that "a picture may be worth a thousand words, but a GIF file takes longer to transmit."

Publishers creating graphics-intensive products need to work closely with their suppliers to learn the benefits and limitations of available graphic formats for the unique characteristics of their publications. As publishers seek to produce information that can be migrated to new media to preserve access, they must be aware that many current systems are platform dependant and therefore will be more difficult to carry forward.

For electronic archives, the importance of output is equal to that of input. This is especially true for graphic images. The following are the most commonly used graphic formats in mid-1997.

Graphic Formats

GIF
TIFF
JPEG
PNG

GIF

GIF stands for Graphic Interchange Format, which the CompuServe online service developed specifically for the WWW. The purpose was to make image files as small as possible to increase transmission speeds over phone lines and via slow modems. With low resolution, GIFs contain only 256 colors. Usually sufficient for the WWW, this is generally undesirable for printing needs, but if there is a need for rapid dissemination such as via the WWW, then relatively small GIF files may be the format of choice (Gentry, 1996). GIF files are suitable for line art and flat tone images, such as logos, but not for halftones.

TIFF

TIFF, a format used for storing graphics, is the abbreviation for Tagged Interchange File Format. These file formats contain bitmapped information and use compression factors that allow the publisher to embed the file within a PostScript file. TIFF files contain the most color information, but the resulting files can become huge (multiple megabytes), so they are not suitable for viewing directly on the Web (Gentry, 1996). Because TIFF has become an industry standard over the years, a publisher can presume that most suppliers can exchange, read, and use this format as well as convert it to other formats, such as GIF for use on the Web.

JPEG

JPEG, which stands for Joint Photographic Experts Group is a relatively new format that is popular on the WWW because it uses a compression method to squeeze a file containing 1.7 million colors into a file smaller than a GIF of 256. Best for photo images because it achieves small file sizes by "losing" information that would not normally be detected by the human eye, JPEG would not be appropriate for schematics or detailed maps (Lane, 1996). For the same reason, it is not applicable for print products. One application would be indexing image archives.

PNG

PNG, Portable Network Graphics, is a new format that is on par with JPEG compression over GIF so it works well in the digital environment. In contrast to JPEG, PNG compresses data with no loss of visual information. On the other hand, it can provide palette-mapped images to only 256 colors. Both GIF and PNG support transparent backgrounds, whereas JPEG cannot. This issue is important only if a publisher is designing a Web-based information product. In mid-1997, the popular Web browsers do not yet recognize PNG, although changes in this area happen quite rapidly.

Graphics Considerations

There are other graphic (image file) formats, but the ones described here are the best known. As we noted earlier, nearly every image-creation program uses its own proprietary format.

Several document services encourage publishers to archive their documents in a digital format. They claim two advantages. The first is that a digital format is edited easily. The second is that a digital format allows for on-demand printing. Storing the information is the challenge—depending on the amount of color and graphics in a publication, storage can take a considerable amount of space in the electronic medium of choice. Current archiving will protect against expensive conversion in the future. What is important to remember is that if the information is not saved somewhere, it will be lost forever.

Standards

Standards specify how a procedure is to be accomplished. In information services and publishing, standards assure a level of interoperability so that information can be made universally available (NISO, No Date). For publishers, adhering to technical standards broadens their potential markets and enables them to achieve economies of scale.

ISO 12083

Titled “Electronic Manuscript Preparation and Markup,” ISO 12083 provides standard DTDs for books, articles, serials, and mathematics. The DTDs can be used with most SGML software without modification. The American Association of Publishers and the European Physical Society proposed a standard for marking up scientific documents, which became ISO 12083 approved in 1996. EPSIG (Electronic Publishing Special Interest Group) was formed as an association to promote the standard. Nonetheless, there is little conformity in the industry, and standardization would help all parties.

PostScript

Created in 1985 by Adobe Systems, PostScript is a page description language that has become the standard for printing and imaging technology. But not all PostScript is the same. Depending on the composition system used, such as Miles, PageMaker, or Xyvision, there can be differences in the graphics, fonts, and nomenclature between files.

PostScript describes to a printer what a page should look like, including the text, graphics, and any scanned images. Because it provides a high level of consistency and clarity, PostScript is used for black and white or color printers, slide recorders, imagesetters, and screen displays. One of the advantages of PostScript is that it is platform-independent; consequently, it works with any DOS, Macintosh, Windows, UNIX, or mainframe computers, and it will communicate with any printer that has Adobe PostScript installed.

Encapsulated PostScript (EPS) is an alternative to some of the graphic formats, such as TIFF. EPS can produce crisp images when treating textual material, such as tables and math, as graphics.

Z39.50

As the database and full-text technology blossomed in the 1980s, the need for standardized methods of retrieving information became more and more apparent. Publishers were producing a variety of CD-ROMS and online products, many with entirely different search and retrieval protocols. Learning so many disparate systems and teaching them to users became an almost insurmountable task for librarians. Iltis (1995) noted that learning the systems was not the only problem: Getting systems to talk to each other and establishing licensing agreements were also difficult.

Clearly, there was a need for information retrieval standards beyond the exiting MARC record format (ANSI/NISO Z39.2). Z39.50, which is an applications-layer protocol, provides a common language to select information and retrieve it (NISO, No Date). For users, Z39.50 means that they no longer need to learn the unique menus, command language, and search procedures of each database system (Turner, 1995).

Ward, Wood, Finigan, and Iannella (1996) pointed out that tools for indexing documents using Z39.50 were available; however, there were no tools for moving existing networked databases to the standard. Consequently, publishers are in danger of losing legacy data. Z39.50 offers so many options for sophisticated operations that in some instances there may be only a base level of interoperability.

BUSINESS CONSIDERATIONS

In developing plans for archiving, publishers must work several fronts. They must determine what existing products they will archive and how they will accomplish the task. At the same time, they must plan for how they will produce, deliver, and maintain products in the future. There is a payoff, however, in the ability to exploit what they have already produced (Grycz, 1997).

Author Relationships

Publishers must assure that they have the rights to publish material in any of several formats over the life of the product. This may mean negotiating changes to existing contracts and changing how they write contracts in the future.

Finances

Developing data in different formats, migrating to new media, creating multiple products all will require substantial outlays of cash. The notion that electronic publication will result in free or inexpensive scholarly communication is probably a pipe dream. As Gotze (1995) noted, the costs of information exchange are not likely to decrease although there are likely to be some shifts in expenditures and in work. Publishers will need to be certain they have the ability to monitor use, control use, and be compensated for their products, or they will go out of business.

They also need to take into account the changing financial pictures of their various market segments. Grycz (1995, p. 50) outlined some critical changes in libraries, commenting that libraries are “rapidly turning from a collections-oriented financial model, to an access-based budgeting model.” Among the forces he saw behind this change were

- Costs of maintaining collections
- Prices for print-based publications
- Electronic infrastructures
- An abbreviated shelf life for some information
- Changes in patron demands
- Operational issues such as real estate values and space usage.

Marketing Strategies

New technology eliminates the need to publish in a “one size fits all” mode. Although the capability of slicing and dicing data to serve various market segments is abhorrent to some, it offers great marketing opportunities to the publisher who makes the investment in learning what each market segment wants. In addition to traditional market research, publishers are likely to engage audiences in testing and developing new products. Further, the task of informing the audience about products has moved beyond traditional print and telemarketing to interactive communications.

What we can do immediately is work for backwards compatibility and to work for forwards convergence.
–Lindquist, 1996

TRANSITIONING TO THE FUTURE

Any discussion of archiving in the digital environment would be incomplete without acknowledging the impact of the Internet and WWW. As scholarly publishers have learned, these new distribution channels make it possible for every person with a computer to be an author with no publisher involved in the process. Most scholarly publishers probably resonate with Holoiviak’s (1995) concerns that this self-publishing will not assure access for future scholars. As she succinctly noted (p. 111), “Archiving functions are utterly lost in a world of self-publication.”

As we discussed earlier, the major technical issue related to archiving and preserving material in the digital environment is the migration of data across changing technology. After centuries of using a single technology (ink on paper), publishers must now become proficient with a multitude of formats based on the newly available technologies. While publishers wrestle with financial viability as they attempt to provide readers with information in all media available today, the challenges to future access of that information become all the more daunting.

Thus, all the players—publishers and authors, publishers and librarians, publishers and suppliers, publishers and aggregators—must collaborate on the best approaches to delivering and maintaining scholarly communications. They will have to deal with the formidable task of refreshing data, copying data to new formats or media, and assuring platform independence. During the past decade, archivists have adopted the technique of “refreshing” digital information by copying it onto the most recently developed media. However, they have found that, despite these activities, their efforts to preserve digital information have limitations.

The issue of archiving is one of many that are challenging publishers near the end of the 20th century. Many authors and some librarians have suggested that publishers do not provide added value in today’s intellectual communications or that their legitimate roles are more limited than they were in the past. One can find dozens of papers on the WWW (see Ginsparg, 1996 for an example) calling on authors to break loose of the “tyranny” of publishers. Ironically, many of the papers opposing publishers contain typographical, factual, and grammatical errors that would not escape most publishers. Furthermore, many are undated, and multiple undated versions exist side by side.

Arnold (No Date), in an essay on rethinking scholarly communication, noted that the “ability to target and tailor information is at the heart of the electronic revolution. As information becomes more widely and rapidly available, people are going to want to control the flow to suit their own purposes, not the purposes of the supplier (that is, the publisher.)” Although the prospect may sound frightening to publishers, it really boils down to what we’ve always known: We have to listen to our audience and deliver information in the format and marketplace they desire.

FOOTNOTES

¹ References cited with “No Date” are sources from the WWW. The lack of a date is a frequent problem on the WWW.

REFERENCES

Arnold, K. (No Date). The body in the virtual library: Rethinking scholarly communication. [Online]. (9 pp.). Available: <http://www.press.umich.edu/jep/works/arnold.body.html>

Bray, T. & Sperberg-McQueen. (1997). Extensible Markup Language (SML): Part I. [Online]. (45 pp.). Available: <http://www.textuality.com/sgml-erb/WD-xml-lang.html>

Brichford, M. & Maher, W. (1995). Archival issues in network electronic publications. *Library Trends*. 43. 701-712.

Comstock, J. (1997). *Brief introduction to SGML*. Presentation to Washington Women’s Information Network. January 1997. Washington, DC.

Donovan, T. (1997). SGML: The chameleon of publishing technology. In *The Editorial Eye*. [Online]. (4 pp.). Available: <http://www.eecom.com/eye/sgml.html>

Fisher, J.H. (1995). Electronic journal update: CJTCS. *Serials Librarian*. 28. 135-138.

Gentry, L. (1996). What is a GIF, JPEG, BMP, etc? Thunderway Publishing. [Online]. (1 p.). Available: <http://thunderway.com/faq/m0003.htm>

Ginsparg, P. (1996, February). Winners and losers in the global research village. Paper presented at UNESCO HG, Paris. [Online]. (8 pp.) Available: <http://xxx.lanl.gov/blurb/pg96unesco.html>

Gotze, D. (1995). Electronic Journals—Market and Technology. *Publishing Research Quarterly*. Spring 1995. 3-19.

Graham, P. (1995, November). Preserving the digital library. In *Long-Term Preservation of Electronic Materials*. Part of the Electronic Libraries Programme, University of Warwick. [Online]. 4-10. Available: <http://ukoln.bath.ac.uk/fresko/warwick.txt>

Greenspun, P. (No Date). We have chosen shame and will get war. [Online]. (6 pp.) Available: <http://www.press.umich.edu/jep/works/greenspun.shame.html>

Grycz, C., Ed. (1997). Preservation. (pp. 24-26). In *Professional and Scholarly Publishing in the Digital Age*. New York: Association of American Publishers.

Grycz, C. (1995). Technological change and its influence on the practice and role of information management. *Serials Librarian*. 28. 43-53.

Holoviak, J.C. (1996). The mixed blessings of society publishing. *Logos*. 7. 106-112.

Horrocks, T. (1996). Design issues on the World Wide Web. *Learned Publishing*. 9. 67-71.

Hunt, L. & Wegner, L. (1996). Project ISLA: A space/time/full-text/format search and retrieval system designed by end users. *The Journal of Academic Librarianship*. 6. 440-449.

- Iltis, S. (1995). Z39.50: An overview of development and the future. [Online]. (8 pp.). Available: <http://www.cqs.washington.edu/-camelz/z.html>
- Lane, T. (1996). *JPEG image compression: Frequently asked questions*. [Online]. (8 pp.). Available: <http://www.cis.ohio-state.edu/hypertext/faq/usenet/jpeg-faq/faq.html>
- Law, D.G. (No Date). *Electronic data archiving and access*. [Online]. (1 p.). Available: <http://astro.fys.ruu.nl:8000/iau/unesco/node5.html>
- Lievesley, D. (1995, November). Strategies for managing electronic archives. In *Long-Term Preservation of Electronic Materials*. Part of the Electronic Libraries Programme, University of Warwick. [Online]. 10-13. Available: <http://ukoln.bath.ac.uk/fresco/warwick.txt>
- Lindquist, D. (1996). *Report from the [Lund University] discussion group: Archiving Issues*. [Online]. (2 pp.). Available: <http://www.ub2.lu.se/NNC/workshop/arkiv.html>
- Mandel, C.A. (1996). Enduring access to digital information: Understanding the challenge. *European Research Libraries Cooperation: The LIBER Quarterly*. 6. 453-464.
- Marchal, B. (1996). *An introduction to SGML*. [Online]. (18 pp.). Available: <http://www.brainlink.com/~ben/sgml>
- Marcum, D.B. (1996). The Preservation of digital information. *Journal of Academic Librarianship*. 22. 451-454.
- NISO. (No Date). *More about standards*. [Online]. (3 pp.). Available: <http://www.ni.org/pub/NISO>
- NISO. (No Date). *The ANSI/NISO Z39.50 protocol: Information retrieval in the information infrastructure. What does Z39.50 do?* [Online]. (2 pp.). Available: <http://www.cni.org/pub/NISO/docs/Z39.50-1992/www/50.brochure.part02.html>
- Task Force on Archiving of Digital Information. (1996). *Preserving digital information*. Commissioned by The Commission on Preservation and Access and The Research Libraries Group, Inc. Washington, DC.
- Tonkery, D. (1995). Reshaping the serials industry. *Serials Librarian*. 28. 65-72.
- Turner, F. (1995). *An overview of the Z49.50 information retrieval standard*. [Online]. (4 pp.). Available: <http://www.nic-bric.ca/ifla/VI/5/op/udtop3.htm>
- University of Virginia. (No Date). SGML. [Online]. (2 pp.) Available: <http://jefferson.village.virginia.edu/iath/treport/sgml.html>
- Ward, N., Wood, A., Finigan, S., & Iannella, R. (1996) *Discussion paper: Networked information retrieval standards*. [Online]. (5 pp.). Available: <http://www.dstc.edu.au/RDU/reports/webir.html>
- Weber Group. (1997). *World Wide Web consortium publishes public draft of HTML 4.0*. [Online]. (3 pp.). Available: <http://www.w3.org/Press/HTML4>

GLOSSARY

- ANSI*** ANSI is the American National Standards Institute.
- ASCII*** ASCII (American Standard Code for Information Interchange) encodes plain text or data with none of the formats added in word processing, graphics, or data files.
- BICI*** BICI stands for the Book International Contribution Identifier, which is used to identify book chapters.
- DOI*** DOI is the abbreviation for the Digital Object Identifier proposed by the Association of American Publishers. Similar to the BICI and the SICI, the DOI is an identifying number to be composed of a publisher prefix and a publisher-chosen suffix and used for unique digital objects.
- DTD*** DTD is the abbreviation for Document Type Definition, which is a blueprint for the structure of a specific type of document.
- EPS*** EPS is the abbreviation for Encapsulated PostScript, a format that is particularly useful for providing textual information, such as tables and math, as a graphic.
- GIF*** GIF is the Graphic Interchange Format used for image files on the World Wide Web.
- HTML*** HTML (HyperText Markup Language) was developed to transmit text across the Internet to any computer.
- HTTP*** HTTP is the abbreviation for the HyperText Transfer Protocol, which is a generic, object-oriented protocol that can be used for many tasks, such as data representation, on the Internet.
- ISO*** ISO is the International Standards Organization.
- ISO 12083*** This standard for “Electronic Manuscript Preparation and Markup” provides standard DTDs for books, articles, serials, and mathematics.
- ISBN*** An ISBN, which stands for International Standard Book Number, is the unique number used to identify each book and its publisher; the ISBN is widely used for ordering and cataloging books.
- ISSN*** An ISSN, which stands for International Standard Serials Number, is the unique number used to identify each journal, magazine, or other serial; the ISSN is widely used for ordering and cataloging serials.
- JPEG*** JPEG (pronounced “jay-peg”) stands for Joint Photographic Experts Group. This graphic file format, named for its developers, uses compression factors to create files smaller than GIF files, but with more color.
- NISO*** NISO is the National Information Standards Organization.
- PDF.*** PDF stands for Portable Document Format, a format developed by Adobe Systems for publishing on the World Wide Web. It delivers searchable files that can be printed in Postscript.
- PNG*** PNG is the abbreviation for Portable Network Graphics, a new format that delivers lossless compression.

<i>PostScript</i>	PostScript is a page description language that has become the standard for printing and imaging technology.
<i>SGML</i>	SGML (Standard Generalized Markup Language) is a platform-independent coding format used to define the structure of a publication. It is a standard (ISO 8879).
<i>TIFF</i>	TIFF, Tagged Interchange File Format, is used to store graphics in a bitmapped image.
<i>XML</i>	The Extensible Markup Language is a new format announced in June 1997. Essentially it is HTML with user-defined tags, so that it combines the flexibility of SGML with the Internet operability of HTML.
<i>WWW</i>	The World Wide Web (WWW) contains linked resources (text and graphic files) in a client/server model. Using the http prefix, users can find documents produced in HTML.
<i>Z39.50</i>	Z39.50 is an ANSI/NISO Standard for information selection and retrieval widely used in systems for libraries.

ABOUT THE AUTHORS

Barbara Meyers is President, Meyers Consulting Services (MCS), which provides expertise in management, marketing, research, and planning to professional societies, scholarly publishers, and commercial firms.

Linda Beebe is President, Parachute Publishing Services, which provides project and program management and support for all publication phases from content development through promotion, distribution, and evaluation.