

Preserving Digital Periodicals

Dale Flecker
Harvard University Library

INTRODUCTION

Everyone has a vague, but few a very precise, idea of what constitutes a “periodical.” For the purpose of this paper, periodicals will be defined as a primarily text-oriented publication that regularly issues new content and intends to do so for the indefinite future. Digital periodicals come in many flavors: selections or versions of paper magazines, such as *Wired*; peer-reviewed scholarly journals; e-‘zines; on-line newspapers; boutique electronic updates or analyses for the business executive; trade, political, or special interest newsletters; etc. These may or may not exist in parallel print/paper form; the two formats may not constitute perfect substitutes for one another. The variety makes generalizations in this paper difficult; the analysis that follows will be accurate for the primary body of periodicals, but the wide variety of producers in this realm ensures that exceptions will be common.¹ Digital periodicals are sometimes based on e-mail delivery or occasionally on the use of specialized reader software, but most today are delivered over the World Wide Web, and that environment is our focus here.

In the paper era, libraries subscribed to and maintained collections of many periodicals (the Harvard libraries currently receive about 100,000 active titles), and collections were highly redundant. Libraries invested in a range of activities intended to maintain the usability of what they collected: binding materials in protective enclosures, repairing damage, housing collections securely and in environments designed to prolong the lifespan of paper, reformatting deteriorated materials through photocopying or microfilming. With the exception of microfilm masters, the copies of journals being saved for future generations were the same copies being read by today’s users. While in research libraries operations were always planned with one eye on the indefinite future, the actions that preserved materials for future generations also served to maintain them for current use.

The new world of Web-delivered periodicals is different. While libraries continue to subscribe to periodicals as they migrate to digital form (subscriptions to electronic journals number in the thousands today in most academic libraries), the service model has changed fundamentally. Libraries no longer receive and store materials locally, and subscriptions no longer provide copies but a license to access. This change has profound implications for the archiving and preservation of periodicals, as it removes two key attributes of the current system:

- copies of periodicals held by institutions that have as a primary role their maintenance for future generations;

¹ It is also worth noting that the analysis in this paper is informed above all by work in one specific domain, the scholarly journal.

- redundancy of copies, which ensures that accidents, theft, conscious destruction, or changes in policy or priority at any given institution do not result in the complete loss of the published record.²

Digital materials are surprisingly fragile. They depend for their continued viability upon technologies that undergo rapid and continual change. All digital materials require rendering software to be useful, and they are generally created in formats specific to a given rendering environment. In the world of paper, many valuable research resources have been saved passively: acquired by individuals or organizations, stored in little-visited recesses, and still viable decades later. That will not happen with the digital equivalents. There is no digital equivalent to that decades-old pile of *Life* or *National Geographic* magazines in the basement or attic. Changes in computing technology will ensure that over relatively short periods of time, both the media and the technical format of old digital materials will become unusable. Keeping digital resources for use by future generations will require conscious effort and continual investment.

In the new world of digital periodicals, copies of materials are often held by a single institution, and the investments required to maintain long-term viability must be made by that institution, which presumably owns it. Factors such as changes in the economic viability of materials, the high cost of a technical migration, a new market focus, company failure, or a reduction in available resources all cause worry about whether such continuing investments will be made. Without such investments, materials will be lost. Such concerns have led libraries to cling to paper copies, when available, even while they provide electronic versions of the same material for the daily use of their readers. This duplicate cost will obviously be problematic over time, and the issue of how to archive and preserve Web-based periodicals is widely felt to have reached a critical state.

TECHNICAL PROFILE OF DIGITAL PERIODICALS

Digital periodicals are surprisingly complex given the seeming simplicity of their paper antecedents, and the level of complexity is growing. The content of digital periodicals comes in a wide variety of technical formats, varying not just among publications, but within a single title or article. The following discussion is not exhaustive of the types of digital material that make up current periodicals, but is indicative of the scope of complexity involved.

The core content of most periodicals is text. The text of a periodical or periodical article, however, can be created and maintained in a number of ways. Some current periodicals are composed of digital pictures of printed pages (frequently, these are then embedded in PDF wrappers for delivery and viewing in the Web environment). More commonly, text is encoded in one of several ways. Some simple publications encode the output of word-processing programs in HTML for Web viewing. HTML provides a rather simplified level of content “mark-up,” primarily oriented toward good visual presentation in today’s Web browsers. More sophisticated publications, particularly

² The back-up and mirroring systems used for many large-scale publications represent only a partial form of redundancy. While good protection against accidents and hardware failure at a specific physical location, they still leave content vulnerable to institutional failure, changes in institutional policy, conscious “amendment” (think of the Stalinist removal from photographs of those who fell from grace), systematic software errors, and the like. Effective redundancy requires that independent players hold copies in separate political jurisdictions, and in differing technical environments, removing the sensitivity to destruction by any single element or agency.

those thought by their creators to be of lasting interest, are frequently encoded in SGML or XML, both of which support much more detailed labeling of components of a textual document. However, SGML and XML are enormously flexible, and different publishers use highly varied mark-up schemes (e.g., DTDs, schemas). Software to render text marked up in this way must be sensitive to the specific scheme used in the text being displayed.

A critical issue with computerized text is the character set used to represent the letters, ideographs, or other components. Standardization in the encoding of text components has progressed enormously in recent years, particularly with the development and adoption of Unicode³ by an increasing range of technology providers. Text for most contemporary languages can be fully encoded in Unicode. However, textual documents contain more than letters and words, and many of the specialized symbols used in periodicals do not have standard digital representations (or evolving standards are not yet widely implemented), including:

- mathematical symbols
- chemical formulae
- archaic scripts or ideographs such as Egyptian or Mayan hieroglyphs
- music notation

Publications containing such extended characters or notations today use a variety of conventions for storage, and rendering software must be sensitive to these conventions when preparing text for Web display.

Periodicals contain more than “simple text.” Visual materials such as photographs and drawings are extremely common and can be encoded in different technical formats. Increasingly, sound and video clips are found in periodical publications, again in a variety of technical formats.

Advertisements represent particular difficulties for archiving and preservation. In paper periodicals, advertisements were usually tied inextricably to specific issues. With Web publications, although most periodical content is relatively static once published, advertisements seen in a particular context can change minute-to-minute, day to day. Advertisements can be selectively displayed for specific audiences or national communities (varying in language or in response to legal restrictions, such as those for drug advertisements). Advertisements are often delivered from a different source than the periodical itself and in fast-changing, proprietary, and challenging technical formats that try to stay on the cutting edge to attract attention. Advertisements represent a rich source for historical research, and their preservation will be of interest. However, archiving and preserving advertisements will pose a significant challenge.

There are other new types of periodical content that raise technical issues. Increasingly, scholarly articles are accompanied by “supplementary materials,” files containing detailed research data, further explication of the article information, or demonstrations of points made in the article. These files contain many types of information (statistical data, instrumentation data, computer models, visualizations, spreadsheets, digital images, sound, or video), and come in a wide range of formats, usually dependent on whatever technical tools the author is using at a given moment. Journal editors and publishers frequently exercise no control over these formats, accepting whatever the author chooses to deposit. More than any other instance of periodical content, these

³ For information about Unicode, see: <http://www.unicode.org/>.

supplementary materials introduce a rapidly growing and essentially unbounded flow of new technical formats that will pose significant difficulties for long-term preservation.

Because digital periodicals are composed of many pieces, frequently in differing technical formats, some form of relationship information is required to map the pieces into a coherent form for delivery to a user. This relationship information can take many forms: “container” formats (such as PDF) that hold explicit or implicit relationships, XML documents, metadata databases, static HTML documents. Practices for what data are recorded and how they are structured vary enormously and are primarily based on the current rendering and delivery applications a publication uses.

One other type of periodical content warrants note. A particular strength of the Web is its ability to link distributed pieces of content, a power as frequently used in digital periodicals as in other types of Web objects. Such linkages come in many forms: some links are to other content in the publisher’s delivery system, where both the link and its target are under the control of the same organization; others are to independent sources. The latter can be of the casual reference sort (“if you are interested in this, that site over there also has relevant material”); other links to separate systems, however, can be integral to the publication (e.g., Web bibliographies or pointers to data in knowledge-bases such as genetics or astrophysical databases). Some links are standard URLs, providing static addresses for specific objects on specific computers. Other links point instead to intermediary systems, capable of finding the current location(s) of the pointed-to object (the Digital Object Identifier, for example⁴). In archiving digital periodicals, it will be important to determine the best way to handle links and the level of responsibility an archive has for maintaining the ability to find independent linked-to objects referenced in archived periodicals.

ORGANIZATIONAL ISSUES

The Open Archival Information System (OAIS) reference model⁵ is a powerful abstract model for digital archiving that has informed much contemporary thinking and practice. OAIS defines roles for three players in archiving: creators, archive operators, and end users (see figure 1).

⁴ For information about the Digital Object Identifier, see: <http://www.doi.org/>.

⁵ For a general introduction to the Open Archival Information System model, see <http://www.oclc.org/research/publications/newsletter/repubs/lavoie243/>. For a detailed description of the model, see: <http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf>.

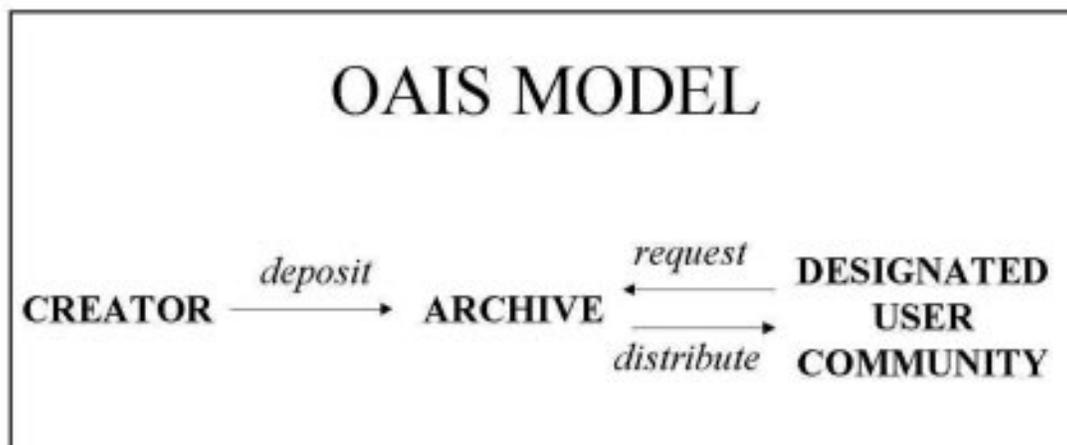


Fig. 1. OAIS model of players and roles

Creators/Depositors

In the case of digital periodicals, “creator” is not a sufficient term, as there are many players involved in digital content creation, formatting, distribution, and ownership. A scholarly journal, for instance, can involve any or all of the following:

- Author(s)
- Copyright owner(s) of included material (e. g., photographs, drawings)
- Scholarly society that owns the journal
- Publisher responsible for peer review, editing, layout, etc.
- Distributor(s) providing online access to the title
- Aggregator(s) that includes an article in a larger online compilation

At least some of these players have a role in “deposit.” It may be useful here to distinguish among players who have the rights, the motivation, and the appropriate technical manifestation to deposit materials and to cooperate in archiving.

Rights. The deposit of materials into an archive necessarily involves questions of ownership and rights: who is legally positioned to provide content to an archive and to negotiate appropriate licenses, if required, for archiving? Because digital periodicals are composed of many separately created pieces, the issue of ownership can be complex. Authors can vary from scholars (who generally, but not always, turn over all copyrights to the periodical owner) to publisher’s employees (whose work is automatically owned by the employer) to freelance writers and illustrators (whose rights may vary with the nature of their contracts). Individual articles can contain separately owned objects, whose owner’s rights also vary (the same picture used under the fair use right of criticism in one periodical requires permission when used in an advertisement in another). The same article can be included in different compilations, for example in the periodical in which it originally appeared and as an aggregated database, such as LexisNexis or ProQuest. Periodical aggregates as well as individual titles could be subject to archiving.

Motivation. The interests of different possible deposit agents vary with the nature of the content, intended audience, and business model associated with specific materials. Some

players' concerns will be purely short term. The economic value of some products falls quickly following publication, and the audience served has little interest in anything but today's information. Such players are unlikely to want to invest in archiving or preservation of their content, but they may also have little concern if others want to do so. Other players may believe that their publications have enduring economic value and may therefore be enormously concerned about independent archives' holding copies of their content and, if archiving is permitted, about the terms and conditions of access to archived content. Still others, such as scholarly societies and original authors, may want to have their materials preserved and may be willing to invest in that preservation.

Technical manifestation. A number of "middlemen" are often involved between the owner and the user of periodical content. In the scholarly journal example, the publisher, distributors, and aggregators all play the role of middleman. Each middleman has its own systems, and copies of periodical content contained in each system can vary, based on the particular nature and function of those systems. A key consideration in archiving periodical content will be the location of an appropriate archival copy. In many cases, the most appropriate copy for archiving may be held by someone other than the owner.

Archive

There is an increasing belief that archiving needs to be the responsibility of institutions for which it is a core mission, and not simply an ancillary operation of an organization whose central interest lies elsewhere. Digital archiving will be a technically and organizationally challenging task, and it is unlikely that a large number of institutions will have the motivation, skill, or resources to undertake the long-term archiving of digital periodicals. The great majority of periodical subscribers and readers will, over time, probably rely on a few institutions to provide storage and preservation of periodical content.

Archives are likely to differ in focus. The organization of archiving activity across institutions involves several important issues:

Collection policy. Each archive will need to consciously and clearly delineate the bounds of its archiving activity. Different institutions may define their responsibilities in different terms: topically, by source of publication (publisher, distributor), selecting specifically important titles, sampling across various specific literatures. As discussed, some level of redundancy is desirable, particularly for titles of potential historical importance. Equally important is the issue of coverage: is an adequate portion of the periodical literature being archived for the use of future generations?

Targeted user community. Both the selection of content for archiving and the specifics of archiving and preservation practice are sensitive to the particular user community for which archiving is being done. Different user communities have different requirements as to what is saved, how it is organized and accessed, the technical formats available from the archive (the writer of popular history needs materials in a form immediately accessible in current technology, the statistical researcher may want data unaltered from the original format), and the technical and support services available from an archive. A key observation of the OAI model is that archiving activity needs to be designed with an understanding of the specified community being served.

Relationship to depositors. An archive does not automatically have the right to copy and store the publications of any given owner. In some cases, archiving activity may fall under the blanket of copyright deposit. But even then, unless the conditions of archiving

are clearly specified in copyright legislation, the owner of archived material may legitimately require a specific license covering the terms of archiving. Given the very large number of publishers and owners of digital periodical content, the transactional cost of negotiating archiving agreements will have to be minimized for archiving to scale appropriately over time. Among the elements that will help here are community agreement on archiving parameters and conventionalized licenses for archiving.

Archiving will come at a noticeable cost. A key issue in the relationship among archives, owners, and users will be the distribution of costs. Some of the major cost elements involved (arranged roughly in order of occurrence) are:

- Notification/discovery of content to be archived
- Creation of an archival version of content
- Creation of archiving metadata
- Storage, monitoring, and management of the archival collection
- Preservation of archived content
- Service to users

These costs can be distributed to the parties in various patterns. One might wonder whether the arrangement above suggests a model of costs distributed to owners, archives, and users as one moves down the list.

Users

As mentioned, the OAIS model suggests that archiving is done to meet the needs of a “specified user community.” User communities vary not only with the nature of publications, but also with the passage of time. While some periodical content will continue to be used primarily as originally intended (e.g., “how to” literature, works describing events or scientific observation, literary or critical works), over time other kinds of use become common. The historian of science or the analyst of trends uses material in ways that are different from the original audience of a publication.

In general, the owners of archived content can be expected to be quite sensitive to two primary questions about users:

Who can access archived content? At least while content is not in the public domain and continues to have economic value, many owners will want to limit the population that can access the archive. Restrictions may vary in scope:

- auditors of the archive
- users with subscriptions to the archived content
- users within the walls of the archive
- users within the institutional bounds of the archive
- users making specific types of use (e.g., the archived objects could be made available to the historian of science, but not the researcher in a pharmaceutical company).

When can content be accessed? Many archiving discussions revolve around the idea of “trigger events,” that is, conditions under which archived content becomes more widely available. The following are examples of trigger events discussed in various venues:

- when a given periodical is no longer accessible on-line;

- a specified elapsed time after initial publication (this is the current policy of PubMed Central, an archiving initiative of the National Library of Medicine, which calls for deposited content to be openly available no more than 1 year after publication⁶);
- when a title changes hands.

Trigger events will probably vary from owner to owner and with the type of publication involved. It is interesting to note the contrasting business models in today's periodical environment that are likely to influence a time-based trigger event. Some publishers charge significant subscription fees for current issues, while providing free access to backfiles⁷. Others, including some newspapers and magazines, provide free access to current issues but charge for access to backfiles. And other business models may yet emerge.

TECHNICAL ISSUES

There are naturally many technical issues involved in periodical archiving that will have to be faced by the various players (owners, archives, and users).

Preserve look, feel, and function? Digital periodicals as perceived by users are composed of a complex of elements: the digital content itself, the display software used to render that content, and a variety of system functions provided by the Web site delivering the periodical. What parts of this complex should be archived? There are a number of questions raised if one were to consider archiving more than the "raw content" (e.g., the words, pictures, or sounds, of the publication):

- Archive display formats, or underlying data? Formats used for ready rendering on the Web frequently differ from the format of content in the underlying publishing system. A publisher may have text marked up in SGML or XML in its asset management system, but deliver HTML or PDF formats, or both, to users today. HTML or PDF may well be easier formats to use if one wants to faithfully recreate the original look of a publication, but many believe they will present archiving problems because the rendering software will certainly be superseded over time. The SGML or XML marked-up text will be less sensitive to technological change, but ensuring the ability to re-render it as it was originally displayed will be technically complex.⁸
- Archive periodical sites? Digital periodicals are delivered through Web sites that frequently offer a wide variety of functions, such as specific organization of content, search facilities, order forms, and communication facilities (to e-mail the editor or participate in a threaded discussion, for example). Archiving entire Web sites with all associated functionality will introduce a significant additional level of complexity beyond archiving periodical content.

⁶ For information about the PubMed Central policy, see: <http://www.pubmedcentral.nih.gov/about/newoption.html>. There is a great deal of discussion in the scientific community about whether all scientific research literature should become freely available after a defined interval. The intent is to provide the publisher with a period of exclusive use for revenue generation. After this period, the literature is open for use by the entire scientific community. A leading initiative in this area is the Public Library of Science proposal, described at: <http://www.publiclibraryofscience.org/>.

⁷ For example, see: <http://www.highwire.org/lists/freeart.dtl>.

⁸ Note that the "original" rendering may in fact be fleeting, as the original publisher may in any case choose to alter and improve display of publications over time.

- Use emulation as a preservation strategy? Emulation has been proposed by some as a means of preserving the original look and feel of digital objects. In this strategy, an archive stores not just the digital objects, but also the software originally used for rendering. Because the software will depend on a specific technical environment (hardware, other software), the archive must build or acquire software capable of emulating that original technical environment, thus permitting obsolete software to run in new environments. Emulation as a preservation technique is highly controversial, with opinions about its practicality differing widely.⁹

What content is archived? At first hearing, most people assume that periodical archiving is simply concerned with the content of articles. While, indeed, articles are the intellectual core of periodicals, in fact digital periodicals contain many other kinds of information. Examples of content commonly found in scholarly journals include the following:

- Editorial board
- Rights and usage terms
- Copyright statement
- Journal description
- Advertisements
- Reprint information
- Editorials
- Events lists
- Errata
- Conference announcements
- Various sorts of digital files related to individual articles (data sets, images, tables, videos, models).

Which of these need to be archived and preserved for the future? Some of these types of materials will pose problems for publishers. Not all of these items are controlled in publishers' asset management systems. Some are treated as ephemeral "masthead" information and simply handled as Web site content. When such information changes, the site is updated and earlier information is lost. For example, few if any scholarly e-journals provide a list of who was on the editorial board for an issue published a year or two ago. Deciding what of all that is seen on periodical sites today should be archived and maintained will require careful consideration by archives, publishers, and users.

Should content be normalized? The variety of formats of digital objects in an archive will affect the cost and complexity of operation. To control such complexity and cost, an archive may want to normalize deposited objects into a set of preferred formats whenever possible. Such normalization can happen at two levels:

- File formats. An archive may prefer to store all raster images in TIFF, for instance, and convert JPEG or GIF images into that format. Controlling the number of file formats will reduce the complexity of format monitoring and migration.
- Document formats. Many publishers encode article content in SGML or XML (or plan to do so soon). Most publishers create their own DTD (or modify an existing DTD) to suit their specific needs and delivery platforms. An archive may choose to

⁹ For a discussion of emulation for preservation, see the following Web sites: <http://www.clir.org/pubs/reports/rothenberg/contents.html>, and <http://www.dlib.org/dlib/october00/granger/10granger.html>.

normalize all such marked-up documents into a common DTD, reducing the complexity of documentation, migration, and interface software.¹⁰

Normalization and translation always involve the risk of information loss. Archiving may well involve a difficult trade-off between information loss and reduced complexity and cost of operation.

Should a standardized ingest format be developed? The OAIS model uses the concepts of “information packages,” that is, bundles of data objects and metadata about the objects that are the unit of deposit, storage, and distribution by an archive. The model allows transformation of objects as they move from one type of package to another (see figure 2).

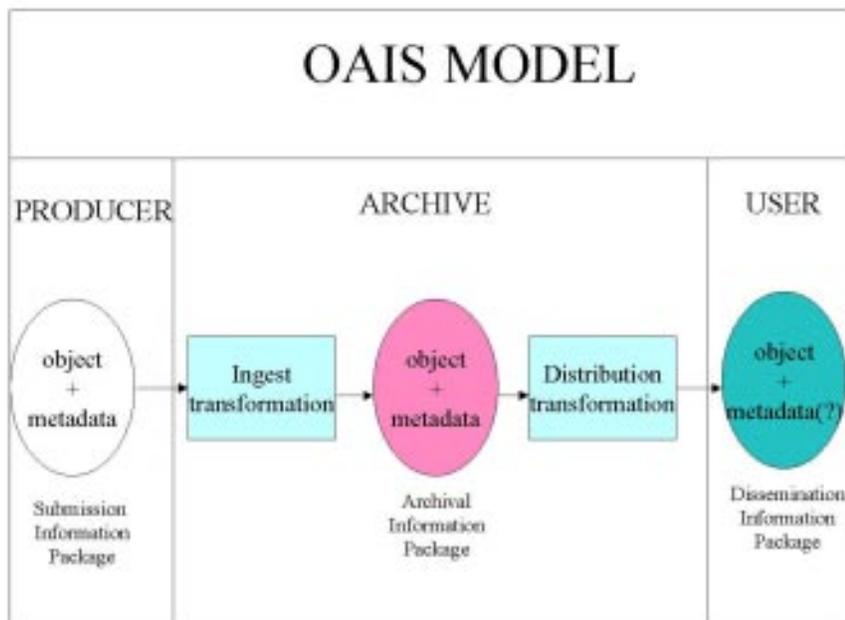


Fig. 2. Information packages in the OAIS model

If, as expected, any given publisher is depositing content into a number of different archives, and any given archive is accepting deposits from a number of different publishers, standardizing the format of “Submission Information Packages” may reduce operational cost and complexity for both communities (although at the cost of devising and maintaining such a standard).

Preserve usable objects, or just bits? A key element in digital preservation is maintaining the usability of digital objects in current delivery technology as the environment changes over time. This process is usually assumed to be one of “format migration,” that is, the transformation of objects from obsolete to current formats, although it can also be carried out through emulation, that is, maintaining current programs capable of emulating older technology and thus rendering obsolete formats.

¹⁰ As part of a current journal archiving project at Harvard, a consultant is examining the feasibility of creating an “archival e-journal DTD,” which would be a preferred format for article deposit.

However it is accomplished, the cost of preservation will be sensitive to the number and types of formats in an archive.

Digital periodicals can contain a wide range of technical formats. Whether it will be practical for archives to maintain current usability for such a diverse range of formats is far from clear. It is possible that archives will need to differentiate between formats where usability is maintained and those for which the archive only ensures that the bits are maintained as deposited and that their documentation is kept useable to support future “digital archeologists.”

SUMMARY

There is tremendous variety in the players, content, and technology that will naturally shape any program to archive digital periodicals and make program planning difficult. However, plan we must, or face losing over time a significant portion of the formal literature of our time. If that happens, future generations will be left with a much poorer understanding of our age than we have of our nineteenth- and twentieth-century ancestors.

FURTHER READING

Council on Library and Information Resources, Digital Library Federation, and Coalition for Networked Information. 1999. Minimum Criteria for an Archival Repository of Digital Scholarly Journals. Available from:
<http://www.diglib.org/preserve/criteria.htm>.

Based on the Open Archival Information System model, these criteria were developed in a series of meetings involving libraries and journal publishers.

Flecker, Dale. 2001. Preserving Scholarly E-Journals. *D-Lib Magazine* 7(9) (September). Available from: <http://www.dlib.org/dlib/september01/flecker/09flecker.html>.

This article describes an initiative of The Andrew W. Mellon Foundation to create several demonstration archives for scholarly digital journals, and enumerates some difficult issues raised in planning such archives.

Mark Bide and Associates. 2000. *Standards for Electronic Publishing: An Overview*.

Available from: <http://www.kb.nl/coop/nedlib/results/e-publishingstandards.pdf>.

This report was commissioned by the Nedlib project (see below), and reviews the current state of practice in using standardized formats for digital books and periodicals.

Nedlib (Web site). Available from: <http://www.kb.nl/coop/nedlib/>.

Nedlib is a project of the European Community involving a number of national libraries. It is intended to describe a framework for electronic copyright deposit and archiving.

“Springer-Verlag joins with international library community in creating electronic information archive for mathematics” (Press release, July 23, 2001). Available from: <http://www.library.yale.edu/~license/ListArchives/0107/msg00088.html>.

This notice describes an international effort to archive the literature of a specific field, mathematics.