Alexa T. McCray and Marie E. Gallagher

PRINCIPLES FOR DIGITAL LIBRARY DEVELOPMENT

Want a library's content to persist and be accessible no matter which computer, browser, or digital format is used? Follow these principles and practices, as well as their implied promises.

Building a digital library is expensive and resource-intensive. Before embarking on such a venture, it is important to consider some basic principles underlying the design, implementation, and maintenance of any digital library. These principles apply not only to conversion projects in which analog objects are converted to digital form, but to digital libraries in which the objects have always been in digital form ("born digitally") and to "mixed" digital libraries in which the objects may be of both types. The principles are, in some sense, self evident, yet it is easy to lose sight of them when under pressure to build a system, despite limited resources and time.

Adhering to the following set of 10 principles (see Figure 1), as well as to the practices that evolve from them, benefits those responsible for the design and continued development of any digital library system, and, perhaps, more important, continues to pay off over the long-term.

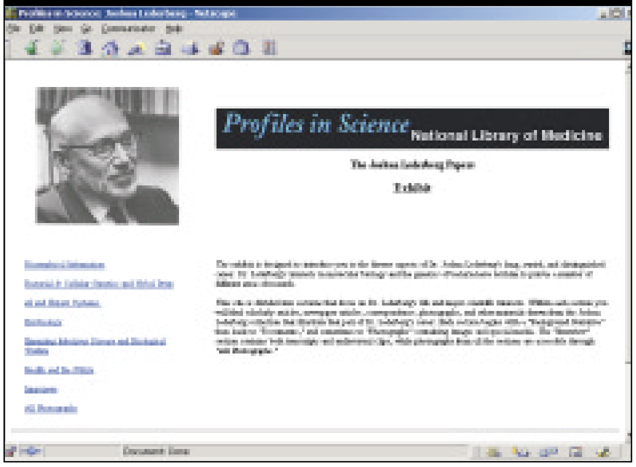
The principles are derived from our experience developing digital library systems over the past decade [7]. We migrated one digital library (created in the early 1990s, before any thought of serving its contents over the Internet) into a more recently created system called *Profiles in Science* (profiles.nlm.nih.gov/). Even though the original hardware and software in the two systems were completely different, the migration was successful, because each was designed with the same basic principles in mind.

Expect change. It may not be apparent why the changing technology landscape is such a thorny problem for digital library projects. Consider, for example, a conversion project in which documents are converted to some digital format. If the chosen format is part of a proprietary system, viewable only through a proprietary interface, when the company that markets the interface no longer supports the system and format, the digitized documents are all but lost. Consider, too, a scenario in which a document is created in a particular word processing program and the document is attached to an email message sent to a notable person. Suppose the goal is to preserve all of that person's email messages for future generations. We are all too aware of our dependence on our email technology for reading such attachments. Imagine what today's platform limitations will mean to future generations, when the content

Figure 1. Principles for building digital libraries.

- Expect change
- Know your content
- Involve the right people
- Design usable systems
- Ensure open access
- Be(a)ware of data rights
- Automate whenever possible
- Adopt and adhere to standards
- Ensure quality
- Be concerned about persistence

Figure 2. Sample exhibit page in the Profiles in Science system.



of the attachments is likely no longer accessible.

Although the Internet, together with the Web, has made digital libraries possible, this fact may also contribute to unforeseen problems if library designers depend too much on today's paradigms and tools. They might be tempted to create a Web site with Hypertext Markup Language (HTML) pages and Web-accessible digital images of objects and documents—that may all be obsolete when HTML changes or is superseded by something else. Changing technologies can quickly outpace the ability of designers to maintain a particular digital library. An approach that anticipates and plans for change is needed to provide lasting access to its information.

Know your content. For users, content is the most interesting and valuable aspect of a digital library (see Figure 2 for a sample exhibit page in the *Profiles in Science* system). Creators of digital libraries need to manage and make decisions about their content, including selecting the objects to be included, digitizing items that exist only in analog form, possibly marking-up items using standard languages like the Standard Generalized Markup Language (SGML), and assigning metadata describing the content and other attributes of each object.

It is important for developers to decide on the nature and number of metadata elements early in a project. Although some elements may be added over time, significant costs might be associated with assigning metadata retroactively to already tagged and cataloged items in a collection. Some metadata elements describe the content of an item, including, say, its title, creator, date of publication, and subjects discussed. Other elements might be assigned for managing the collection; examples include scan status, quality-control status, and internal notes, as well as the technical aspects of the digital objects, such as file format and size.

Important, too, is deciding on the basic conceptual units, or objects, the system will include, such as individual documents, photographs, videos, or lab notebooks. This decision affects the level at which metadata is assigned (for example, to an entire book or to each chapter in the book) and how the materials are organized, accessed, and archived. Sometimes overlooked is the practice of assigning each conceptual object its own unique identifier linking it to its metadata record and to other objects in the collection.

Involve the right people. Ideally, individuals from a variety of backgrounds and offering a variety of expertise contribute to building a digital library. In practice, this may not be the case, but even when it's not, knowing that building the system requires insight from a number of fields yields a better digital library.

The two fields involved most directly are computer science and library science. Computer scientists appreciate the possibilities, as well as the limitations, of technology and are generally the ones who actually build the system. Librarians, including catalogers, indexers, and archivists, have long been the custodians of information resources, understanding not only the information needs of diverse audiences but the issues involved in preserving materials for continued access and use. Digital library research and development have meant that each group has had to understand the other groups' perspectives.

Illustrating the importance of multiple perspectives is the development of the Dublin Core metadata standard [8]. Computer scientists are concerned with the semantic interoperability a digital library metadata standard affords in the very large Internet information space; librarians already have deep experience indexing and cataloging and recognize the importance of these concepts for information retrieval. Moreover, because so much valuable data exists in a variety of metadata systems, including the Machine Readable Catalog

(MARC 21) standard, methods to map the data between and among these systems using automated "crosswalks" have been developed [10].

Also important when embarking on a project within an organization is whether its senior management supports the effort. Because most digital library projects are long-term efforts, they require the commitment of long-term financial and human resources. Beginning such a project involves an implicit, if not explicit, commitment to the continuation of the work and a promise that the digital materials will continue to be available. Lacking organizational commitment, it may not make sense to even begin a project.

Design usable systems. Most digital libraries are made available over the Internet through Web technology, though, strictly speaking, this is not a necessary attribute of a digital library. However, as the advantages of the Web are so great, most library systems today are designed to be Web-accessible. The most successful Web site designs account for a number of factors, including the technical differences among computers and browsers, including speed of access, and differences among users, including Web navigation preferences. Browsers differ in the way they display information, even though they use the same basic communication protocols (such as the Hypertext Transfer Protocol, or HTTP, and File Transfer Protocol, or FTP) and standard markup languages (such as HTML and perhaps the Extensible Markup Language, or XML). Since users may change default settings, including font size and other parameters, it is always preferable to create as simple an interface as possible and avoid server-side control of the exact display of the data. Providing multiple access points not only makes a digital library more interesting, it also acknowledges the differences among its potential users.

Accessibility for users with a range of physical disabilities should also be a concern when developing the interface to a digital library. This includes user access to all content; documents that are clear and simple; user control of styles; the availability of context and orientation information; inclusion of clear navigation mechanisms; and standard markup.

Ensure open access. Ensuring open access is closely related to usability concerns, including access to the information in the digital library, as well as to the digital library itself. Christine Borgman defines access to information as "connectivity to a computer network and to available content, such that the technology is usable, the user has the requisite skills and knowledge, and the content itself is in a usable and useful form" [2]. Michael Lesk writes that open access to information raises a number of public policy issues, including whether or not all segments of society are given equal access to information [6].

One way to ensure open access to content is to avoid proprietary hardware and software solutions whenever possible. That is, while it may be reasonable to create content using commercially available systems and tools, avoid requiring special software or hardware to access that content. A number of companies, while charging for the tools used to create digital images, make their interfaces available for free. As long as the software is easy to download and install, and the developers of the digital library make it clear which software is needed to view the images, the content will be accessible, as long as the interfaces are available. In all cases, however, for continued accessibility and use, open, nonproprietary systems are preferred.

Be(a)ware of data rights. A possible threat to open access to information arises because of intellectual property concerns. Existing intellectual property and copyright law provides economic and legal protection to publishers of physical artifacts. "Fair use" (allowing libraries to make, say, single copies of portions of books or journals) and "first-sale" rights (allowing individuals to, say, lend or resell copies of books they have purchased) have promoted greater access to physical artifacts than might be possible otherwise, but these notions are only indirectly applicable to networked information. A recent National Research Council report states: "The information infrastructure offers both promise and peril: promise in the form of extraordinary ease of access to a vast array of information, and peril from opportunities both for information to be reproduced inappropriately and for information access to be controlled in new and problematic ways" [3].

There are no straightforward answers to the enforcement of intellectual property rights for information available in digital form. The Internet and Web have emerged from communities that believe in sharing information, rather than restricting access to it. This has led to the perception, and perhaps even expectation, that anything available on the Web is freely available and may be redistributed at will. Some argue there will be a realignment in the way intellectual content is published and distributed worldwide (for example, so authors are the primary "publishers," largely replacing existing commercial publishers). Others argue for technical solutions to restrict illegal copying. Still others argue that copyright law needs to be strengthened to prevent unauthorized use of digital information. And still others argue that perhaps the legal protections should come from contract, rather than copyright, law, so data is licensed, rather than sold.

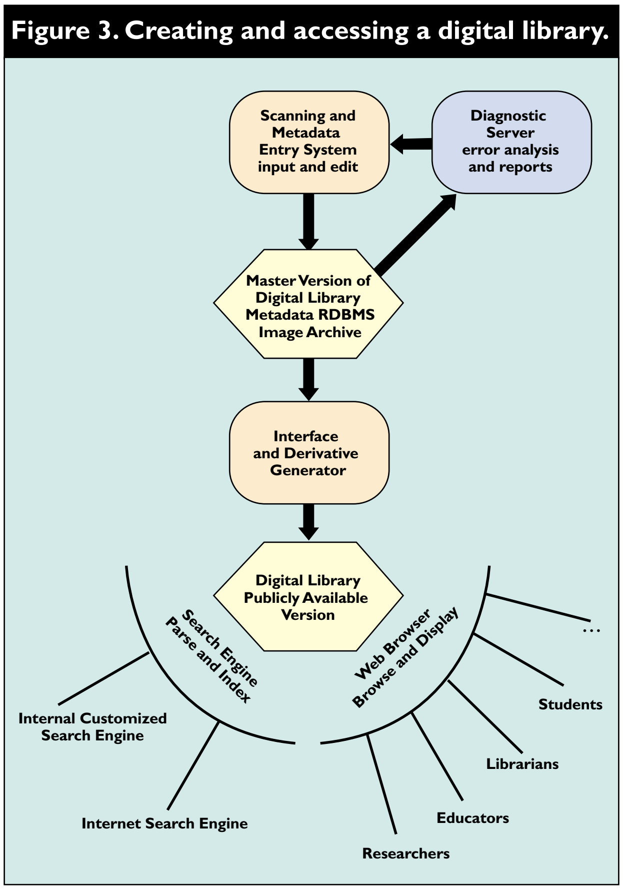
Some resolution to these problems will undoubtedly arise in the coming decades. In the meantime, however, those creating digital libraries need to be aware of the issues, participate in the public debates about their resolution, and establish procedures to manage them in order to protect their collections to the extent possible [1]. For example, in conversion projects, every attempt should be made to seek permission from the copyright owner for the materials that are to be digitized. Privacy issues should also be considered when, for example, the full papers of a prominent individual are being digitized. Ideally, the donor will have marked the items that are sensitive in some way and left instructions about how they should be handled. In all cases, however, careful tracking of permissions and privacy information ensures the collection will not be at risk at some future time.

Automate whenever possible. Because building a digital library requires significant intellectual effort on the part of the system's creators, the more automated tools that can be built and used, the better will be the use of precious human resources. These tools need to be easy to use and incorporate real-time aids, including data validation, pull-down lists, report generation, and other time-saving devices, thereby allowing the content expert to concentrate on the intellectual tasks at hand (see Figure 3 for the design outline of *Profiles in Science*). Content experts use the metadata entry system to add metadata to a master database, entering the information only once. Subsequently, the information is extracted and combined as needed from the master database to generate HTML pages, search indexes, and reports. Entering the data only once saves human time and effort, reduces the error rate, and allows maximum flexibility. Nearly the entire Web interface is generated from the database, allowing regeneration whenever necessary, while adhering to the latest Web standards. The system is designed to be modular, allowing existing modules to be modified easily and new modules added for additional functionality. For example, we added the diagnostic server to provide a preliminary view of the master *Profiles in Science* database, allowing content experts to discover and correct errors and inconsistencies before public release.

Figure 4 outlines the relationship between the underlying master version of the system's data and its derivatives. High-quality lossless TIFF images, for example, serve as archival copies and as master copies for creating a variety of derivatives, such as PDF or JPEG files, for optimal use on Web sites. Then, as better compression algorithms are developed, it will be easy to discard the derivatives (precisely because they are derivatives) and generate a newer file format from the original master.

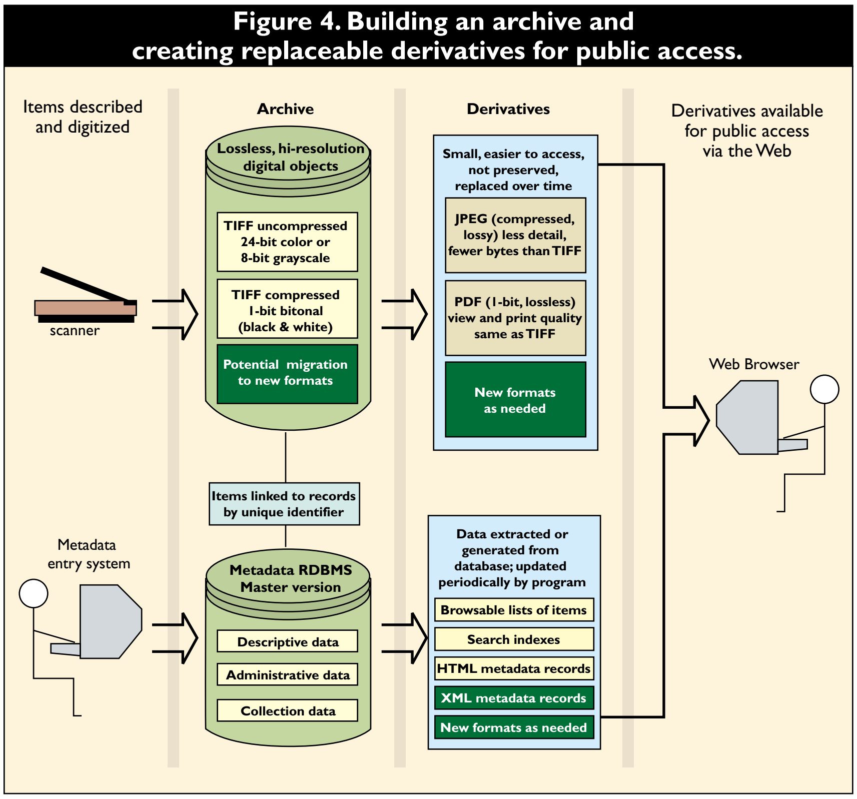
Adopt and adhere to standards. The use of standards in system building has many benefits. Applications are more readily scalable, interoperable, and portable [11]; these characteristics are all important for the design, implementation, and maintenance of digital libraries. Using standards is especially important for the aspects of digital libraries that are most labor-intensive. Scanning, metadata entry, and document markup, all involving the evaluation and handling of individual items in a collection, are resource-intensive and best done carefully and only once. Data might still have to be migrated to other forms and formats in the future, but migration will be easier, because standards have been used consistently. Images scanned using standard file formats, such as TIFF, or texts saved in the open ASCII or Unicode formats will be more easily accessible in the future than images or texts encoded in proprietary formats.

Figure 3. Creating and accessing a digital library.


Scanning and Metadata Entry System input and edit
Diagnostic Server error analysis and reports
Master Version of Digital Library Metadata RDBMS Image Archive
Interface and Derivative Generator
Digital Library Publicly Available Version
Search Engine Parse and Index
Web Browser Browse and Display
Internal Customized Search Engine
Internet Search Engine
...
Students
Librarians
Educators
Researchers

The benefits of using standards for interoperability should be clear to any developer. For example, if all finding aids for historical collections are marked-up in SGML using the Encoded Archival Description document type definition, one can easily imagine distributed access to all the finding aids held by the members of a university library consortium. Delivering the contents of the digital library on the Web, using standard, valid, and current HTML, including metadata tags, and other standard Web technology, increases the chances that other Web search engines will be able to find the library, as well as the specific items in it.

Figure 4. Building an archive and creating replaceable derivatives for public access.


Items described and digitized
Archive
Derivatives
Derivatives available for public access via the Web
Lossless, hi-resolution digital objects
TIFF uncompressed 24-bit color or 8-bit grayscale
TIFF compressed 1-bit bitonal (black & white)
Potential migration to new formats
Small, easier to access, not preserved, replaced over time
JPEG (compressed, lossy) less detail, fewer bytes than TIFF
PDF (1-bit, lossless) view and print quality same as TIFF
New formats as needed
scanner
Web Browser
Items linked to records by unique identifier
Metadata entry system
Metadata RDBMS Master version
Descriptive data
Administrative data
Collection data
Data extracted or generated from database; updated periodically by program
Browsable lists of items
Search indexes
HTML metadata records
XML metadata records
New formats as needed

Because technology is changing so rapidly, it is important to question whether it will be possible to port an existing system to another platform, user interface, or software. In principle, while porting should always be possible, in practice it is much easier when the system's design and implementation adhere to certain standards (see Figure 5 for pointers to standards resources). It's also possible that sometime in the future the entire digital library will be part of another more-encompassing system, either within or outside the original institution. The transition will be smoother and more successful if the system is designed with open standards in mind.

Ensure quality. Quality metrics can be applied to all the processes and outcomes involved in creating a digital library. They are relevant to selection, metadata entry, image capture, and the overall usability of the system. Complete and correct metadata yields many benefits; incomplete or incorrect metadata affects the quality of the entire digital library. Metadata plays a vital role not only in resource discovery but in managing the collection. If, for example, subject codes are applied haphazardly or incorrectly, access could be more difficult, and attempts to generate browse hierarchies based on these codes could be foiled.

If scanning procedures and guidelines in conversion projects involve immediate review and evaluation of the scanned images for appearance, including orientation, resolution, color, and tone, there will be fewer future problems. Images that are skewed, dark, or cut off are not welcome in a digital library. Digitized video and audio need to be reviewed periodically for adherence to evolving standards and for their ability to be viewed and heard using current tools.

Some quality-control metrics can be automated; others require careful human review. Digital library projects should define and then carry out quality-control methods as part of their normal procedures. Adhering to such methods ensures that quality assessment becomes an integral part of building and maintaining the digital library.

Be concerned about persistence. A recent article by Jeff Rothenberg describes a year-long effort to understand the issues of digital preservation, noting: "The conclusion reached by the impressive group of 21 experts was alarming; there is, at present no way to guarantee the preservation of digital information" [9]. While preservation has long been a concern of archives and libraries, it has only recently been of interest to a much larger community.

Anyone creating a digital library has a stake in the outcome of preservation approaches, yet only some address the problem directly. A survey of the member institutions of the Research Libraries Group, which is interested in the question of digital-object preservation, ranked "technological obsolescence as the greatest threat to digital collections" [5]. Several suggestions have been made by researchers to address preservation of digital objects. Perhaps the one discussed most often is the "migration strategy," which entails the transformation of data from one file format to another, converting it from one software environment to another, or moving it from one physical medium to another. Migration implies a strong and enduring commitment on the part of an institution to continually refresh its collections to keep up with the technology. Another proposed method is the "emulation strategy," which entails the emulation of an entire software system so it runs on future unknown systems [9].

The Digital Library Federation, a consortium of libraries interested in electronic information technologies, draws an important distinction between preservation and persistence of digital objects [12]. Preservation refers to an object's technical longevity and quality; persistence is a much broader notion, encompassing preservation, but also referring to whether the object would still exist in any form at all in the future. Persistence implies a commitment to both maintaining the object and keeping it accessible. It is even possible that entire digital libraries will disappear if efforts are not made to maintain them. Several years ago, an analysis of existing Web sites found the average lifetime of a URL was only 44 days [4]. This discouraging statistic may be accounted for in a number of ways, including that data has been moved, not deleted, but also that we, as a community, are right to be concerned about these issues.

Conclusion

When creating digital library systems containing valuable content, we are making important promises to both current and future users. Seriously attending to the principles discussed here and to the practices that evolve from them places us in a much better position to keep these promises. Valuable content should be handled with care and rendered in the highest quality possible. Valuable content should not disappear. We need to understand how to preserve and safeguard digital material, so it doesn't become obsolete simply because we didn't pay attention. Finally, we need to strive for continued open access to all knowledge. There is no better time to start than now and no better place to start than with our own valuable collections.

Figure 5. Useful URLs to standards resources.

Accessibility guidelines
- Web Accessibility Initiative at W3C; www.w3.org/WAI/
- Federal Information Technology Accessibility Initiative; www.section508.gov/

Initiatives and resources
- Digital Libraries Initiative Phase 2; dli2.nsf.gov/
- Open Archives Initiative; www.openarchives.org/
- Digital Library Standards at Berkeley SunSITE; sunsite.berkeley.edu/Info/standards.html
- Open Archival Information System (OAIS); www.ccsds.org/RP9905/RP9905.html
- U.K. Interoperability Focus; www.ukoln.ac.uk/interop-focus/

Intellectual property
- Copyright, Intellectual Property Rights, and Licensing Issues; sunsite.berkeley.edu/Copyright/
- U.S. Copyright Office; www.loc.gov/copyright/

Markup standards
- Hypertext Markup Language (HTML); www.w3.org/MarkUp/
- Extensible Markup Language (XML); www.w3.org/XML/
- Standard Generalized Markup Language (SGML); www.w3.org/MarkUp/SGML/

Metadata standards
- Dublin Core; dublincore.org
- MARC 21; lcweb.loc.gov/marc/
- Encoded Archival Description (EAD); lcweb.loc.gov/ead/
- International Federation of Library Associations and Institutions Metadata Resources; www.ifla.org/II/metadata.htm
- Metadata and Resource Description at W3C; www.w3.org/Metadata/

Network standards
- The Digital Object Identifier; www.doi.org/
- Hypertext Transfer Protocol (HTTP); www.w3.org/Protocols/
- Persistent URL Home Page; purl.oclc.org/
- Uniform Resource Names (urn) Charter; www.ietf.org/html.charters/urn-charter.html
- Z39.50; www.loc.gov/z3950/agency/

Standards organizations
- Association for Computing Machinery (ACM) Technical Standards Committee; www.acm.org/tsc/
- American National Standards Institute (ANSI); www.ansi.org/
- Coalition for Networked Information (CNI); www.cni.org/
- Digital Library Federation (DLF); www.clir.org/diglib/
- Institute of Electrical and Electronics Engineers, Inc. (IEEE) Standards Association; standards.ieee.org/
- International Organization for Standardization (ISO); www.iso.ch/
- Internet Engineering Task Force; www.ietf.org/
- National Information Standards Organization (NISO); www.niso.org/
- World Wide Web Consortium (W3C); www.w3.org/
- Unicode Consortium; www.unicode.org/

Subject access and control
- Medical Subject Headings (MeSH); www.nlm.nih.gov/mesh/meshhome.html
- Getty Vocabulary Program; www.getty.edu/research/institute/vocabulary/introvocabs/
- Library of Congress Subject Headings (LCSH); lcweb.loc.gov/cds/lcsh.html

References

1. Arms, W. *Digital Libraries.* MIT Press, Cambridge, MA, 2000.
2. Borgman, C. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World.* MIT Press, Cambridge, MA, 2000.
3. Committee on Intellectual Property Rights and the Emerging Information Infrastructure. *The Digital Dilemma: Intellectual Property in the Information Age.* National Academy Press, Washington, DC, 2000.
4. Kahle, B. Preserving the Internet. *Sci. Am.* (Mar. 1997); see www.sciam.com/0397issue/0397kahle.html.
5. Kenney, A. and Rieger, O. *Moving Theory into Practice: Digital Imaging for Libraries and Archives.* Research Libraries Group, Mountain View, CA, 2000.
6. Lesk, M. *Practical Digital Libraries: Books, Bytes, and Bucks.* Morgan Kaufman Publishers, San Francisco, 1997.
7. McCray, A., Gallagher, M., and Flannick, M. Extending the role of metadata in a digital library system. In *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries* (Baltimore, May 19–21). IEEE Computer Society, Los Alamitos, CA, 1999, 190–199.
8. National Information Standards Organization. *The Dublin Core Metadata Element Set: Draft Standard Z39.85*; see www.niso.org/Z3985.html.
9. Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Rep. to Council on Library and Information Resources, Jan. 1999; see www.clir.org/pubs/ reports/rothenberg/pub77.pdf.
10. St. Pierre, M. and LaPlant, Jr., W. *Issues in Crosswalking: Content Metadata Standards.* National Information Standards Organization, 1998; see www.niso.org/crsswalk.html.
11. Strand, E., Mehta, R., and Jairam, R. Applications thrive on open systems standards. *StandardView 2,* 3 (Sept. 1994), 148–154.
12. Waters, D. *The Digital Library Federation: Program Agenda.* A program of the Council on Library and Information Resources, June 1998.

Alexa T. McCray (mccray@nlm.nih.gov) is the director of the Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD.

Marie E. Gallagher (gallagher@nlm.nih.gov) is a computer scientist at the Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD.