Digital Libraries and Their Challenges

DANIEL GREENSTEIN

ABSTRACT
This article derives from a review of key challenges confronted by libraries that are actively investing in online collections and services. Conducted in the first instance to help refine the programmatic goals of the Digital Library Federation (DLF), it took account of the digital library developments, successes, needs, and challenges perceived by professionals working at the DLF's twenty-four member libraries. Methodologically, the review relied on two research tracks. The first involved desk-based research into the strategic documentation and technical reports that inform members' work in this area. The second involved extensive discussions convened at some twenty-seven sites.

After a brief summary of some key findings related to the digital library—definitions of the digital library are possibly premature and will underrepresent the extent to which its activities are shaped by local institutional, legal, and business imperatives—this article reviews five key challenges offering some thoughts about how each may be confronted in the future.

INTRODUCTION
The digital library extends the breadth and scale of scholarly and cultural evidence and supports innovative research and lifelong learning. To do this, it mediates between diverse and distributed information resources on the one hand and a changing range of user communities on the other. In this capacity, it establishes "a digital library service environment"—that is, a networked online information space in which users can...
discover, locate, acquire access to and, increasingly, use information. Although access paths will vary depending on the resource in question, the digital library service environment makes no distinctions among information formats. Books, journals, paper-based archives, video, film, and sound recordings are as visible in the digital library service environment as are online catalogs, finding aids, abstracting and indexing services, e-journal and e-print services, digitized collections, geographic information systems, Internet resources, and other "electronic" holdings.

In constructing a digital library service environment, the library becomes responsible for configuring access to a world of information of which it owns or manages only a part. Accordingly, the digital library is known less for the extent and nature of the collections it owns than for the networked information space it defines through its online services. In the world of commercial publishing, aggregators compete on the basis of the value-added services that they layer on top of overlapping electronic collections. Similarly, digital libraries establish their distinctive identities, serve their user communities, emphasize their owned collections, and promote their unique institutional objectives by the way in which they disclose, provide access to, and support the use of their increasingly virtual collections.

The digital library service environment is not simply about access to, and use of, information. It also supports the full range of administrative, business, and curatorial functions required by the library to manage, administer, monitor engagement with, and ensure fair use of its collections whether in digital or non-digital formats, whether located locally or off site. The digital library service environment integrates (and interfaces with) information repositories that are characterized by open-access shelving, high-density book stores, and availability via interlibrary loan, and include data services and digital archival repositories. It manages information about collections and items within collections often throughout their entire life cycle. It incorporates patron, lending, and other databases, and integrates appropriate procedures for user registration, authentication, authorization, and fee-transaction processing. The digital library service environment may also evolve into a networked learning space, providing access to, and a curatorial home for, distance and lifelong learning materials. The digital library service environment is, in sum, an electronic information space that supports very different views and very different uses of the library. It is designed for the library's patrons as well as for its professional staff and with an eye on the needs and capacities of those who supply it with information content and systems. It is built in the full knowledge that information technologies will continue to change rapidly as will our understanding of how they can be used to support education and cultural engagement. Finally, it is evolving as the library's defining function and as such is developed with a view to its financial and organizational sustainability.
ARCHITECTURAL AND SYSTEMS CHALLENGES

The digital library typically relies on a narrow base of appropriately skilled professionals to keep abreast of the rapid pace of technical change while maintaining, indeed extending, robust and fully operational online services and collections. In both respects, it is stretched beyond capacity with evident deleterious effect. Lacking the resources to develop core systems components (e.g., search and retrieval tools, user interfaces and user profiling services, user authentication and authorization services) that work across individual collections and services, the digital library adopts a tendency toward a more ad hoc approach that meets the most pressing demands involving development work. Although viable in the short term, the strategy threatens severely to undermine a position over which the library exerts only a tenuous hold—that of the trusted provider of high-quality information services. Where pure research and development activities are concerned, the rate and pace of technical change diminishes the time between the identification of a potentially valuable new technology and its deployment in a digital library service environment while the risks and costs associated with any decision to deploy a new technology remain stable or increase. Accordingly, libraries are investing in more technologies, more often, and with less information than at any time in the past.

The palliative measures that are currently offered are only partial at best. Mechanisms that encourage greater information sharing offer some promise, though one cannot under estimate the severe constraints under which they operate. One can envisage (can already identify, for example, in the Coalition for Networked Information) forums that allow appropriately skilled professionals to:

- inform one another about potentially valuable new technologies;
- share results of any local experimentation with and assessment of such technologies; and
- foster shared experimentation and assessment where appropriate.

More systematic information sharing requires more formal exchange and review of the rich technical literature that is beginning to populate digital library's internal Web pages.

Research and development efforts led by OCLC and other organizations that are grounded in, and responsive to, the library community are also promising. Yet one cannot overestimate the contributions that such organizations can make in an area that cannot easily be made to sustain itself financially.

Perhaps greater reward will accrue from recent efforts to delineate which of the panoply of systems necessary to support digital libraries need to be developed and maintained by the library per se. Initial investigation into certificate-based authentication, reference-linking, and name-resolu-
tion services, for example, suggest the existence of a class of infrastructural services that is required by the digital library but more effectively mounted on an institutional, or even cross-institutional, level. In the commercial world, such high-tech utility services are the stuff of the business-to-business sector. Encouraging a similar developmental trajectory for digital library services that are commonly required but beyond the capacity of any single library organization to supply would require libraries to articulate requirements and aggregate demand for such services in order to create incentives for third party suppliers to move in and supply the market.

STANDARDS AND BEST PRACTICES

The need for “standards” and “best practices” is universally felt but so differently defined as to render the objects of desire almost meaningless. The emergence of guidelines that lay equal claim to objectivity and authority, a welcomed sight when only a trickle, begin to compound the obfuscation through their proliferation. In this context, it may be helpful to reveal three related, but distinctive, needs:

1. for information that helps digital libraries flatten their own learning curves;
2. for some community-wide agreement about the minimum level of data creation practices that promise to support the library in its various roles of integrating access to, supporting use of, and managing electronic information content;
3. for benchmarks that help “consumers” evaluate digital library collections and services.

Flattening the learning curve is a main source of concern that can be explained in part with reference to the fact that the digital library’s ambitions frequently exceed its research and development capacity. Whether launching an initiative to construct EADs, digitize illuminated manuscripts, or develop proxy authentication services, the digital library has a natural inclination to learn from, rather than to relearn, the experiences of others. Satisfying this substantial demand is probably more a matter of information sharing than some other complex effort aimed at identifying standards or even best practices. What is required is not so much prescriptive documentation (e.g., how to use the Core Categories supplied by the Visual Resources Association–VRA) so much as decision tools that guide project planning and introduce and signpost alternative solution strategies.

Agreement about preferred data creation practices is necessary to support digital libraries in their efforts to supply services (e.g., resource location and retrieval, data analysis and long-term management, user support) that mediate between end users on the one hand and extensively distributed, deeply heterogeneous, information content on the
other. Diversity of data content is not tied exclusively to the existence of different data formats (e.g., raster graphics, ASCII texts, GIS) and metadata schemes (e.g., the Text Encoding Initiative's [TEI] Header, the VRA core, the metadata standards recommended by the Federal Geographic Data Committee—FGDC) but to the fact that there are few common implementations of any single format or scheme. Data resources are typically developed to meet the very specific needs and interests of particular end-users (one is all too familiar with the diversity borne of the phrase "fitness for purpose"). They rarely take into account the library's needs as an organization responsible for layering services across a cacophony of electronic content. The impact on libraries is as considerable as it is predictable and once again encourages a tendency already noted toward ad hoc solutions that are developed to meet the need of specific collections. The more cost-effective and scalable approach that relies on the development of core technologies requires a greater degree of consistency across networked information resources than currently applies.

Here one suspects that libraries would do well to work cooperatively with one another and with the specialist communities that develop and implement data creation standards to articulate the minimum level characteristics they prefer (even require) for the different classes of information content they are expected to serve. The result need not be unduly prescriptive or constrain the idiosyncrasy and innovation that has emerged as a hallmark of the online information resources that are produced by or for scholarly communities. It may also be welcomed by the data creators who would at least be supplied with some formal statement of those practices that are most likely to promote access, exchange, security, and longevity of their own content.

Benchmarking is required because digital libraries operate in a networked environment where they are both consumers and suppliers of digital collections and services. As consumers, they will want to know, for example, about the quality, persistence, and longevity of the collections and services that are offered by commercial third party suppliers of subscription-based journal content. As suppliers of collections and services, libraries should expect to confront (and to meet) the same demand from their own users. Those users definitely include individuals but also institutions—other libraries for example—which link to, or wish to interoperate with, their locally managed collections. As a supplier of networked information services, it is simply not enough for a digital library to disclose a collection of Web-accessible images created as surrogates for items in its special collections, for example. Institutional users must know about image quality, persistence, and longevity as will scholarly users whose academic endeavors are built in part on the ability to identify and relocate information sources.
Benchmarking standards are not, one suspects, a short-term goal. They are more likely to develop as a logical outgrowth of the other activities that are indicated in this section. Indeed, the developmental path is likely to be an incrementally progressive one. Exchanging information about current practices will help flatten the learning curve encountered by digital libraries as they launch new collections and services. By reviewing current practice once assembled, it may be possible to identify those preferred practices that support the digital library in its development of mediating services that work across deeply heterogeneous information content. Comparative evaluation of those mediating services as they mature may supply a foundation upon which some benchmarks may emerge.

**Collection Development**

Had this review been conducted seven, five, or even three years ago, it might have encountered a more heroic approach to the development of digital collections. Whether compelled by the business logic of electronic publishing, the proliferation of public-domain Internet content; the prospects of enhanced access to special, rare, and archive holdings; or by an interest in gaining core competence in key technical areas, libraries were found enthusiastically creating and acquiring access to digital information content. Some years on, work in each of these areas is as enterprising. Enthusiasm for it, however, is tempered by an equal measure of reticence that is borne of experience and reflects concern about cost, longevity, integration, and scale. One consequently encounters a very real concern to understand better how decisions to create or acquire access to a digital resource will impact on how and at what cost that resource will be used, how it will be integrated into existing library collections and services, and how it will be maintained and supported over time. In effect, one encounters the earliest stages of what may emerge as a fundamental revision of traditional library collection policies and practices.

Such revisionism can only be encouraged even if it transpires that the same high-level considerations effectively govern the development of traditional and hybrid (digital and nondigital) collections. From our present perspective, the differences on either side of the digital divide are more apparent than the similarities. For digital formats, the rate and pace of technical change, the volatility of digital media, and the implications that access licenses have for collection development and use forces fundamentally new considerations, e.g.,

- the costs involved in accessioning a data resource into a collection (a process that may itself include data reformatting, metadata creation or amendment, systems design or modification, development of any documentation that may be required by end users, public service librarians, systems librarians, etc.).
the copyright and licensing issues and associated system requirements (e.g., to maintain security, process registrations or payments);

- the hardware, software, and networking environments that are required to provide access to a resource and of the stability, maintenance, and potential migration of those environments;

- the methods and costs involved in migrating data through changing technical regimes with as little information loss as possible;

- the development and provision of appropriate user support services;

- the impacts that a digital resource once accessioned or created will have on the work of departments concerned with cataloging, licensing and administration, public service, and library systems.

Here, as elsewhere, it may be prudent to assemble and critically assess current practice emerging within those libraries that are beginning to think strategically and programmatically about developing their digital collections. A library-based literature already exists to guide the planning and implementation of digitization projects. If synthesized and supplied with a life-cycle approach that anticipates how design and implementation decisions impact longitudinally over time and horizontally across the full range of library services, it would contribute enormously to any revision of library collection policy. A similarly systematic review might be conducted for other components of the library's increasingly digital collection including, for example, the electronic information that is supplied by commercial third parties and the online finding aids and indexes, including those comprising links to third-party networked information resources.

Clearly, the library needs ultimately to allocate its limited acquisitions budget effectively across a range of very different information resources, including those listed earlier but also the more traditional paper-based and analog formats. In this respect, it needs collection policies that assist in weighing the relative short- and long-term costs, benefits, and value that are associated with very different resource types. Although the effort described here will not supply that policy framework, it will provide some essential building blocks.

Penetrating and Mobilizing User Communities

In a digital library, how information is made, assembled into collections, and presented online affects whether, to what extent, and how it can be used. A truism, perhaps, but one that marks an incremental step for the library into an arena traditionally occupied by publishers. The statement is not intended to re-open a tired and unhelpful debate about whether the future holds a place for either the library or the publisher. Rather, it demonstrates how our understanding of a digital object's life cycle implies a need to engage differently or, perhaps, just more deeply with our users in order to enhance our understanding of their needs,
aspirations, and behaviors as an essential design component of online collections and services; and develop appropriate support services.

There are at least three additional reasons to place user relations on a revitalized—perhaps slightly different—footing. Some user communities that surround the library are producers of digital content, including research data, dissertations, e-prints, and computer-assisted teaching materials. That content has enormous educational and cultural value, but only if it is assembled into professionally managed collections, maintained over time, and made meaningfully accessible to other end users through online portal and other services. User communities may also possess the expertise or knowledge that, when applied effectively to existing collections, digital collections can substantially enrich and enhance those collections, thereby lending new meaning to them and even making them accessible or comprehensible to communities for which they were not initially intended. In this regard, engaging with user communities is a vital component of any effective collection development policy.

Second, some user communities are aware of the tools necessary to manipulate information and are mobilizing, sometimes on a large scale, to supply those tools—especially where they are unavailable from the commercial sector. Ask a professional archaeologist or a geographer about GIS; a musician or engineer about the manipulation of sound data; an economist, social scientist, or astronomer about the management and analysis of large-scale statistical data; a medic or a film producer about the management and analysis of still or moving images. Just as some user communities are poised as net suppliers of digital content, others may be in a position to supply tools to a digital library service environment that can enhance that environment's functionality.

Third, digital library collections and services represent a substantial and growing investment by libraries, educational institutions, and other cultural organizations. Even where there is no commercial imperative to mount such collections and services on a cost recovery or revenue-generating basis, there will be significant pressure to measure performance and value of investment in terms of use. By engaging with user communities more effectively, libraries can inform investment decisions by anticipating their potential benefits (and beneficiaries, where some financial return on investment is sought).

To re-engage its user communities, the digital library will work on several levels and in ways that are dictated by purely local circumstance. Work in at least one area may benefit from some greater community-wide attention, notably in the development and application of quantitative and qualitative methods that help assess users' needs and interests in light of their behavior in, and use of, contemporary online environments. Although such methods exist and are being deployed by libraries to meet some of their user assessment needs, there is substantial room for shared
activity. Work coordinated by ICOLC (International Coalition of Library Consortia, 1998) testifies to the potential that exists for such activity, yet it focuses narrowly on the use statistics that the library community would prefer to gather from its commercial suppliers. The community might benefit from a similarly concerted undertaking that focuses simultaneously in two directions: on measures deployed in a commercial context as a central part of marketing e-services, and on the use being made of existing library-supplied online services and collections. Work along the first of these tracks will require perspective and expertise from outside the library and academic communities. Work along the second will require some systematic assessment of existing but highly fragmented experience.

**LONG-TERM ACCESS TO DIGITAL INFORMATION**

The persistence of digital information remains an essential challenge for digital libraries. A few are poised to develop limited archival repositories. Their progress may rely on the emergence of two elements that are currently absent.

First, there is no widespread agreement about the minimum functional requirements of a digital archival repository. Such agreement is essential. Without defining what maintenance entails (and thus the requirements of the repository), libraries cannot tell suppliers of digital content what is needed to preserve the information. The suppliers need to agree on the requirements of a repository to satisfy any demand that libraries may make with regard to that content's persistence.12 Finally, for emerging repositories to be trusted, whether as suppliers or consumers of digital content, they require a blueprint for the services they need to offer and a benchmark against which their services can be measured and validated.

A second element that is absent from the digital preservation arena is a more realistic understanding of the value of digital information. The costs of maintaining digital information over time are unknown but undoubtedly high. The costs of information loss are likewise unknown, but the potential costs must be considered. For example, a drug company maintains data generated in the development of a new product for as long as those data have value to the company. Such data might be kept as evidence in the case of legal action; the costs of not preserving the data could be ruinous. In this context, preservation may be expensive but less so than the alternative.

It would be difficult for libraries to make similar assessments, given their overwhelming focus on commercially produced scholarly materials (e.g., journals and reference services). Moreover, because of the number of subscriptions they hold, it would be unlikely that any single library or library consortium could take responsibility for preserving such content over the longer term, nor does long-term preservation motivate the commercial supplier. And the commercial supplier's understanding
of "longer term" will understandably be at variance with that of the library.

Might we begin, then, with digital information for which no one is likely to take an archival interest—e.g., with the digital surrogates, for example, that are created by some libraries? This is not to suggest that all digital surrogates must be preserved. The British National Gallery periodically re-digitizes its collection of some 2,500 art objects to take advantage of new imaging technologies. The same strategy is not necessarily advisable for all, especially those conducting projects to digitize tens or even hundreds of thousands of individual objects. The question to be addressed is not only about the costs of preservation but also about the higher costs that are likely to be involved in periodic re-digitization.

And what about the digital content emanating from surrounding academic departments that makes up an increasing proportion of the university's intellectual assets? Computer-based research, learning, and teaching materials have significant value. Yet that value is fully realized only if the materials are assembled into professionally managed collections and maintained over time.

Admittedly, decisions to maintain the university's intellectual assets will not be made by the university library in isolation. The information content that is available from the university's digital library makes up only one part (a very important part to be sure) of the university's portfolio of information assets. To determine its value and the bearable expense involved in its preservation, the entire portfolio needs to be reviewed. In the university context, progress in digital preservation is likely to require institutional ownership of a far broader preservation problem.

Digital library research and development agendas are not difficult to come by and I am certain that others will include more compelling and urgent issues than those included here. The key challenge is in mobilizing efforts behind those agendas, and it is here that the digital library, given its commitment to maintaining legacy services and the dearth of R&D capacity, faces significant obstacles. Membership organizations have a role to play, and many are engaged in relevant activities. There are constraints, however, on what they can achieve. Some have substantial investment in legacy services that need to be maintained in a way that limits organizational capacity for innovation. Others are built and rely on broad group consensus that may not always be conducive to the pioneering spirit that is required.

The digital library's institutional context presents another set of obstacles. In the university sector, the digital library is constrained by the absence of any institution-wide approach to electronic information and knowledge management. Key infrastructural components that are commonly required by administrative, academic, and information service departments (e.g., those associated with rights management and security,
preservation services, aggregating and indexing functions) are offered on a piecemeal departmental basis. They are thus under-resourced (as may be the case for authentication services) or unsupplied (e.g., with respect to preservation services). A balkanized approach also characterizes the creation of educational content, whether construed as online learning materials, research data, or e-prints. As a result, the content emerges in such diversity as to hyper-inflate the cost of its professional long-term management and exploitation. In the meantime, no single organizational unit within the institution is positioned to mount a meaningful defense against the large and growing number of commercial providers stepping aggressively into the provision of educational information and services. If, in the evolving knowledge economy, the university’s future is tied to its ability to capitalize on local intellectual and information assets, its present approach to information management will emerge as its biggest, and potentially most debilitating, threat.

In the higher education sector, the digital library’s challenges are not its own. They belong to its host institution and need to be resolved at an appropriate institutional level. Wendy Lougee (1998) has persuasively characterized the digital library’s development as a staged evolutionary process. I would endorse that view and suggest that the digital library is mature when its activities are woven into the fabric of its host institutions—when its challenges and its successes are no longer its own. One suspects that maturation will at some point require a bold step on the part of the digital library—one that entails giving up to a higher or broader authority its hold over its own strategic planning process. Only such an authority will be able meaningfully to guide the mission-critical activities on which the digital library’s development currently encroaches.

NOTES

1 The Digital Library Federation (2000) is a consortium of digital libraries operating under the umbrella of the Council for Library and Information Resources and seeking to leverage their collective reputations, investments, and research and development capacities for their own benefit and for that of the library community at large.

2 Institutions visited in the course of the review include the Library of Congress, the New York Public Library, the National Archives and Records Administration, and the university libraries at Berkeley, California Digital Library, Carnegie Mellon, Chicago, Columbia, Cornell, Emory, Harvard, Indiana, Michigan, Minnesota, New York, North Carolina State, Pennsylvania, Princeton, Southern California, Stanford, Tennessee, Texas at Austin, Virginia, and Yale. Visits were also made to the Research Libraries Group and OCLC. Documentation pertaining to digital library developments is included in a bibliography of technical reports and strategic documents available from the DLF’s Web site (“Documenting,” 2000).

3 The typical digital library will maintain few, if any, professional staff devoted exclusively to research and development activities. At the same time, access to research capacity in academic departments is severely curtailed in part because digital library research applications are perceived as being less attractive to researching scholars than those with immediate application in industry, and in part because the exigencies of producing scalable, robust, and stable online services rarely fall within the researching scholar’s purview. Interesting counter examples exist at several libraries which have, sometime
in conjunction with other departments, established units with an applied research orientation. Examples include university libraries at Columbia in its Academic Information Systems Research and Development unit (n.d.) and its Center for Research on Information Access (2000), at Harvard through its Library Digital Initiative (2000), the California Digital Library through its Advanced Technology Unit, Michigan through its Digital Library Production Service (2000), and the Cornell Institute for Digital Collections (n.d.).

Counter examples may be found at Columbia (through ACIS, n.d.), Michigan (where the DLPS is developing middleware for various classes of digital content including text, image, bibliographic, and finding aids, Harvard (through the work of its Library Digital Initiative, 2000), Virginia, Cornell ("Cornell Project Prism," 2000), Berkeley, and the California Digital Library. Here, work is underway on so-called "middleware" that supplies core digital library system components.

One cannot help but to be astonished by the persistent failure of academic faculty to recognize the library as a trusted supplier of high-quality information content capable of offsetting the disturbing tendency they detect among students to conduct their research and seek support for their assignments through commercially supplied Internet search engines and reference services.

The directory of technical reports referenced above represents a narrow attempt by the DLF to develop such an exchange ("Documenting," 2000).

This approach is being adopted by the University of Michigan among others and is most fully developed for data in text and image classes. See, for example, the documentation surrounding the image class at Image Services (2000).

For a few examples, see the Library of Congress' "Collection Policy Statement: Electronic Resources" (1999); "Electronic Resources at the University of Michigan" (2000), and the several publications that exist at "Managing Electronic Resources at Yale University" (2000). The latter site, currently under early development, is an aggregation of Yale University Library policies and procedures related to electronic resources. These documents have been created, variously, by CoDGer (the Committee on Digital General Resources), the Collections Development Council, the e-cataloging committee, the Medical Library, Kimberly Parker (Electronic Collections Librarian), and others. This page is a product of CoDGer's deliberations and is maintained by its chair, Kimberly Parker (kimberly.parker@yale.edu); Cornell University "Report of the Committee on Electronic Resources" (World Wide Web Working Group, 1996).


The inter-relationship of qualitative and quantitative methods is explained in Rieger & Gay (1999).

The text of this section appeared in Greenstein (2000).


REFERENCES


To survive in the digital age and stay relevant, public libraries need to be brave and innovative. They must embrace both the physical and virtual. Libraries must offer more than just books. In lower-income countries, like Bangladesh, the context for libraries is different but some of the same challenges remain. Bangladesh itself is a small country with a huge population, estimated at around 162 million people. Bangladesh has plenty of libraries, but the quality of their services is inconsistent. Bangladesh’s current library and information services provide access to information to varying degrees, but they are fragmented and unaligned in what they offer. For instance, there are 68 government public libraries, all in urban areas, which have long been in need of investment.