Into the Future:
The Challenge and Promise of Technology for Digital Libraries

John Wilkin, University Library, University of Michigan

Abstract

For too long, we have conceived and built digital libraries as self-contained worlds consisting of digital object creation, management, discovery, delivery and use—walled gardens in the midst of thriving communities. This model runs counter to an intensively networked world where users encounter information through large discovery systems like Google and Amazon, and where "use" takes place in a variety of environments outside of the systems that manage those resources. As users shift to a networked world, digital libraries are challenged to break with their history of stubborn insularity, to adapt to that world or lose relevance. There is much that is on the horizon that holds promise for helping digital libraries change and increase their relevance. Wilkin will discuss the current paradigm of digital libraries and the consequences of that paradigm, will discuss the way that we can adapt our work in digital libraries, and will briefly examine specific examples of technologies that can contribute to more successful digital library efforts.

1. The problem—what do our digital libraries look like?

For too long, we have conceived and built digital libraries as self-contained worlds isolated from other library work and from the network at large. These digital library worlds, consisting of the entire lifecycle of work—from digital object creation, management, discovery, delivery to use—seem like walled gardens in the midst of thriving communities. In them, we see:

- **Large numbers of isolated efforts**: This is not news to anyone. Even in the Digital Library Federation, a group of 30 institutions, their registry of collections shows more than 500 distinct collections, not searchable as a whole or even (in many cases) in clusters by institution.
- **Wild variation in architectures:** Nearly every institution prides itself on having a unique architecture, created around a distinctive philosophy.
- **Equally great variation in software, often homegrown:** The pride in architecture is echoed in software development efforts. It is as if we are saying "if you cannot build it yourself, you should not be in the digital library business." Although some software (e.g., Greenstone, Fedora, DSpace and Michigan's own DLXS) tends to predominate, we see many examples of homegrown software with relatively few development resources and small collections.
- **Little commonality in services:** Aside from the most obvious services—search and display—there is very little agreement on the types of services that might be needed. One obvious exception is OAI metadata exposure, but even here many major efforts do not share metadata through OAI.

Is there evidence that this approach to building digital libraries is isolating?

- Nearly 70% of content represented by records in OAIster is not represented in Google.
- Moreover, my research assistant recently (unscientifically) selected twenty-three "digital libraries" from one online list and searched for their contents in Google.

In all but six, some or all of their contents were not findable in Google. In at least two, none of their contents were in Google. Said another way, 74% had significant content not findable through Google.

We would like to believe that if our digital library efforts are successful, our users should reasonably expect to find the contents of digital libraries in Google or any other broad Internet search. The reality is that they do not find that content: our digital library content remains isolated, perhaps with superior functionality and even superior search capabilities within the various systems, but as a world unto itself, increasingly separated from the users we care about most.

2. What does the world outside of our digital libraries look like?

- **Intensively networked:** Google, Amazon, Flickr, iTunes and others dominate the discovery and use environment of our users. As Lorcan Dempsey argued, the "massive computational and data platforms [of Google, Amazon and EBay] exercise [a] strong gravitational web attraction," a sort of undeniable central force in the solar system of our users' web experience. OCLC environmental scans make clear that our users look to services like Google and Amazon before turning to those things we purchase or build.
- **Increasingly scholarly:** Three of the more interesting emerging developments of late have been OCLC's WorldCat Local, Google Book Search, and Google Scholar. What has happened with WorldCat Local, Google Book Search and Google Scholar has extended that same sort of pull to key scholarly discovery resources. Now, however, mainstream "network services" like Amazon and Google web search, deficient in their ability to satisfy scholarly discovery, are complemented by similarly "massive computational and data platforms" that specialize in just that—finding resources in the scholarly sphere.
- **Increasingly divergent from us:** As mentioned, our approach has been to build walls rather than connections. These forces (Amazoogle), and perhaps more like them in the future, should influence the way that we design and build our systems. If we ignore these types of developments, choosing instead to build systems with ostensibly superior characteristics, *systems that sit on the margins*, we effectively ensure our irrelevance, building systems for an idealized user who is practically non-existent.

In the library world, our resources, skills and investments have helped to create an opportunity for us to shape a next generation of library systems, simultaneously cognizant of the strong network layer **and** our needs and responsibilities as preeminent research libraries. We have designed and built our past systems in partial isolation from each other system, reflecting the state of library technology and our response to user needs. We were not *wrong* in the way that we developed our systems, but rather we were right for those times. What our libraries must do *now* is reconceive our efforts in light of the changed environment. And the reconceptualization should not only be built with an awareness of the new destinations our users choose, but also with a recognition that we have a special responsibility for the long-term curation of library assets. As good as the Amazoogles are, they are at best incomplete. Even at its most successful, Google Scholar

does not include all of the significant investment in electronic resources that we purchase for our communities, and Google Book Search is not designed to support the array of activities that we associate with scholarship.

3. Principles for a newly-shaped digital library world

Knowing that we must re-direct where we invest our resources is one thing; knowing *where* we must invest is another. I do not believe I should (or could) paint an accurate picture of the sorts of shifts we should make. On the other hand, I can lay out here a number of key principles that should guide our work.

- **Balanced against network services**: I believe this is probably the most important principle in the design of what we must build. We must not try to do what the network can do for us. We must find ways to facilitate integration with network services and ensure that our investment is where our role is most important (e.g., not trying to compete with the network services *unless* we think we can and should displace them in a key area). For example, we have recognized that Google will be a point of discovery, and so rather than trying to duplicate what they do well for the broad masses of people, we should (1) put all things online in a way that Google can discover; and (2) because we recognize that Google will not build services in ways that serve all scholarly needs, work to strategically complement what they do. In the first instance (i.e., making sure that Google can discover resources), we will always need to block them, for legal or other reasons, from discovering content. In this, libraries and archives share a common problem. These types of exceptions should add nuance to what we do in exposing content. In the second instance, when it comes to building complementary services, we will need to be both smart (and well-informed) and strategic.
- **Openness**: What we develop should easily support our building services *and*, even more importantly, should allow others to build them. It should take advantage of existing protocols, tools and services. Throughout this document, I want to be very clear that these principles or criteria do not necessarily point to a specific tool or a specific way of doing things. Here, I would like to note that the importance of openness, though great, does not necessarily point to the need to do things as *open source*. As O'Reilly has written in his analysis of the emergence of Web 2.0, this is what we see in Amazon's and Google's architectures, where the mechanisms for building services are clearly articulated, but no one sees the code for their basic services: the investment shifts from shareable software to *services.* Similarly, our being open to having external services built on top of our own should not imply that our best or only route is open source software. What is particularly important is the need to have data around which others would like to build tools and services: openness in resources that few wish to include is really only "putting lipstick on a pig."
- **Open source**: Despite what I noted about openness above, wherever possible, we should try to do our work with open source licensing models *and* we should try to leverage existing open source activities. In part, this is simply because, in doing so, we will be able to leverage the development efforts of others. We should also aim for this because of the increasing cost of poorly functioning commercial products in the library marketplace. Note, though, that when we choose to use open source

software, it is important to pick the right open source development effort—one that is indeed open and around which others are developing. Much open source software is isolated, with few contributions. We should aim for openness in our services over slavish devotion to open source.

- **Integration**: Tight integration is not the most important characteristic of the systems we should build, nor should this sort of integration be an end in itself; however, we have an opportunity to optimize integration across all or most of our systems, making an investment in one area count for others. In Michigan's MBooks repository, we have already begun to demonstrate some of the value in this type of integration by relying on the Aleph X-Server for access to bibliographic information, and we should continue to make exceptions to tighter integration only after careful deliberation. I am also a firm believer in the value of "loose" integration (e.g., automatically copying information out of sources and into target systems), but the example of the Aleph X-Server has been instructive and shows the way this sort of integration can provide both increased efficiency and greater reliability in results. It may also be obvious, but a focus on integration will also lead to more modular systems, and I have a word to say about modularity later.
- **Rapid development**: If we take a long time to develop our next generation architecture, it will be irrelevant before we deploy it. I know this pressure is a classic tension point between Management and Developers: one perspective holds that we are spending our time on fine-looking code rather than getting a product to the user, and the other argues that work done rapidly will be done poorly. This dichotomy is false. The last few years of Google's "perpetual beta" and a rapidly changing landscape have underlined the need to build services quickly, while the importance of reliability and unforgiving user expectations have helped to emphasize the value of a quality product. We cannot do one without the other, and I think the issue will be scaling our efforts to the available resources, picking the right battles, and not being overambitious.

4. Directions and Promise

These sorts of defining principles are familiar and perhaps obvious, but what is less obvious is where all of this points. If we in digital libraries do not position ourselves to take advantage of the types of changes I mentioned at the outset, we will enhance our existing investments for a few years until our systems or even our libraries are entirely irrelevant. If we make the right sorts of choices in the current environment, we should also be able to capitalize on the efforts of others, thus compounding the return on each library's investment.

- We must move toward articulating an overarching integrated environment.
- This approach does not presume that we will replace our existing technologies with something different. Many libraries have made many good choices on technologies that are serving their institutions well, and to the extent that they are the best or most effective tool for aligning with the principles I have laid out, we should use them. The X-Servers of Aleph and MetaLib are excellent examples of tools that allow the sort of integration we imagine.
- Where there is a shared development community, we can benefit from a community of developers.

In all of this, we will need a strategy, and a strategy that remains flexible as the landscape changes.

There are some clear indications that these sorts of principles I have described are at play, and I would like to conclude with some hopeful examples of trends:

- **Open source development:** Evergreen and LibraryFind may only carry forward the existing model of library technologies in isolation, but the fact that they are open source and that development energy comes from some of our brightest colleagues bodes well.
- **Services at the network level:** The development of WorldCat Local seems especially promising, despite significant shortcomings. For many reasons, it does not compete with Endecca, Primo, and the rest of the NextGen discovery tools, but it does promise to situate itself as a powerfully centralized version of this type of discovery.
- **Smart architectures:** The repository system, Fedora, has been designed right for this kind of layers and integrated approach. The community that has grown up around it is evidence of its success. The mere example of VITAL, or integration of Fedora in VTLS's library management system is precisely what we would hope for in at least one regard.
- **Modularity:** All of these examples are promising, but perhaps the most promising to me is the recent attention to extend OAI to add support for Object Re-use and Exchange (OAI-ORE). As the OAI web site notes, the effort intends to "develop specifications that allow distributed repositories to exchange information about their constituent digital objects. These specifications will include approaches for representing digital objects and repository services that facilitate access and ingest of these representations. The specifications will enable a new generation of cross-repository services that leverage the intrinsic value of digital objects beyond the borders of hosting repositories." Sometimes we think too big, and this approach of adding to our services a small component is likely to have tremendous and far-reaching impact.

It is time to see our environment as being comprised of a set of inventory management responsibilities (both print and digital, both local and remote) that leverages a growing and maturing array of network services so that our users can effectively discover and use the resources available to them. I think that requires a change in the way we think about our technologies and a much more strategic arrangement of those technologies in relation to each other. We may be stuck with a bunch of local print "repositories" because of the nature of print and the history of library development. That is not the case for our digital repository, however. On top of our repositories, we need to conceptualize the sorts of services we need (e.g., ingest, exposure, other types of dissemination, archiving, etc.) and the tools that can best accomplish these things. Digital libraries and digital archives are presented with some fundamentally different problems, but they share many of the same challenges, and I hope that these discussions help to point to some issues and approaches in the area of archives.