

Gold at the End of the Digital Library Rainbow: Forecasting the Consequences of Truly Effective Digital Libraries

Michael A. Keller

Green Library, Stanford University, Stanford, CA 94305
makeller@stanford.edu

Abstract. This paper contemplates a truly effective digital library from the user's point of view. It will contain vast amounts of information, comparable to, but different from, the public web; it will help the user make real sense of that information by organizing, evaluating, and testing the reliability or authenticity of needed information; it will help the user manage, adapt, and reuse the information gathered. We will have to understand the digital library as both consumer of information and as publisher and partner with the scholarly author. Digital preservation is and will be a paramount concern.

Creating digital libraries is not easy or inexpensive. The literature on the subject is vast and growing, and a considerable part of that literature is concerned with architecture, design, methods of digitizing and building digital collections, meta-data for digital objects, intellectual property issues and digital rights management, and other "technical" or "professional" matters. We must not lose sight of the true goal of academic digital libraries and that is to make research and teaching better, faster, more penetrating, more intuitive, more powerful. This exhortation concerns itself with that goal and only in passing with the "technical" and "professional" matters. The metaphor of gold at the end of the rainbow is a Western literary allusion, for many of us one that is associated with Irish culture, no doubt in part because there is enough rain and sun mixed in Irish days to make rainbows more common than perhaps they are elsewhere. The problem with the metaphor is that rainbows are not constructed and operated by men, but by nature. And one who chases the rainbow to find its end is chasing an illusory goal, for it is very difficult to pin down the end of a rainbow. For many, the end of the rainbow keeps moving, resetting itself. That part of the metaphor, at least, is apropos of building digital libraries!

As this conference and many others like it attest, we, as information professionals, are challenged deeply by the practices, assumptions, and implications of an increasingly digital and global information space. In my own library, we are struggling to change the whole organizational structure in light of digital collection and service issues. Over the course of 2004, we have been working to structure our support of digital information. Bear in mind, we have been aggressively pursuing digital technologies, collections and services for perhaps two decades, but we find that the many currents and streams have become such a torrent that the very foundations of library

services require examination. It is probable we will create a Digital Services Bureau within the Stanford Libraries, separate from, but of lesser stature than, the Collections and Services and Technical Services division.

In fact, some of my staff have suggested humorously that we should rename Technical Services as the Analog Services division, in recognition that much of their work continues to revolve around the physical artifact. That may be premature, but it may be instructive. The core of what we do continues to depend on the two pillars of Collections and Services: the traditional Technical Services function shares with the Digital Services function a supporting role in making collections available to our readers and providing them with the services necessary to assure access, discovery, interpretation, and continuity to those collections. On the one hand, I have estimated that creating the Digital Services Bureau would shift the organizational position of perhaps 25 individuals. On the other hand, the changes implied by this reorganization will directly influence the working environment for the other 350 or so staff.

I will address here some issues pertaining to “a truly effective digital library” (per the title of this presentation), but the first thing to say is that a truly effective digital library will not be entirely digital: it will complement, supplement, enrich, and inform collections that remain, for whatever reason, in analog form, whether that form is cuneiform tablet, parchment manuscript, silk scroll, printed book, silver-halide image, annotated typescript, or magnetic tape. The original of a text, to the extent it can be preserved, needs to be preserved and kept accessible for scholarly use. After we digitize it, of course.

For some years, I have been referring to the “both/and” dilemma: we must maintain our collection development and services for printed books and serials, while also growing (both by acquisition and creation) our digital collections and services, despite the relatively static budgets universities have provided. The “both/and” dilemma is shifting: we will be buying more digital material and less analog material. But the dilemma will, if anything, become more pointed in these coming years of the early digital age: we must not only spend on digital services and materials, we must continue to acquire what exists only as analog material. Concurrently we must digitize collections and preserve digital materials. At the same time, we must also retain access to the hardcopy, as a safeguard and as an alternate form of access. The stacks in libraries are not going away anytime soon. My provost wishes to believe the digital age will decimate the space needs of the campus libraries; it is my duty to contradict those wishes.

Three Key Aspects of the Digital Library

Let us look briefly at a truly effective digital library from the user’s point of view. In the first place, it will provide truly vast amounts of information. The public web now contains an unaccountably large volume of information – Google indexes over four billion pages, and worries rightly that it doesn’t access nearly enough content. The digital library I envision contains a similar or larger scale of information not presently on the Web at all.

In the second place, the effective digital library – very much unlike the public web, despite the Googles of the world – helps the user make real sense of that vast amount of information. It not only helps retrieve, but organize, evaluate, and test the reliability or authenticity of needed information.

In the third place, the effective digital library will help the user manage, adapt, and reuse the information gathered. In fact, we will have to understand the digital library as both consumer of information and as publisher, or at least as a form of partner with the scholarly author.

The remainder of this talk will expand these three aspects. First, though, we as information professionals must ponder something the user probably assumes or ignores: preservation. We know that digital collections, while inherently reproducible, are also inherently unstable and at risk of catastrophic loss. I will not dwell here on the fundamental mandate for us to preserve the digital collections we assemble, but I wish you to bear in mind this mandate critically influences nearly everything we do.

Building Digital Collections

Unlike conventional collections, digital collections may be accumulated in four ways:

- **by Purchase** - we may buy digital sets much as we do books or serials. Doing so, creates the need to house the materials on our servers or other technology. It also provides us with control, but the corollary is that we hold the responsibility not only for preservation, but also for format migration, an issue that will come to occupy much of our attention in years to come.
- **by Licensing** – The popular rubric is “access, not ownership.” This is an important and attractive – indeed inevitable – method of providing information to our readers/users. The downside, of course, is that we in the library pay, but have no say or control in the long-term access to the licensed content. If a publisher (here using that term to mean the provider of licensed content, whatever it calls itself) decides to increase its prices arbitrarily, if it removes content, if it goes out of business, the library has little option, and its prior investment is for naught. On the other hand, issues such as format migration, archiving, and hosting remain with the publisher, thus limiting the managerial complexity to the library. It may be a pact with the devil.
- **by scanning** or other digitization means - There are many mass book digitization schemes afoot today – I am deeply in discussion with several parties about this effort – each of which is hampered by financial, technological and intellectual property (copyright) issues. We hope that key libraries will be able to undertake mass digitization projects so as to make millions of books available to the world in digital form. Whether this is something practical within the next few years, as opposed to a decade or more hence, has yet to be determined, but I am confident that there will be successful mass digitization efforts to the benefit of the world’s readers sooner or later. And the impact of such an expansion of access to the world’s lit-

eratures is almost incalculable and tremendously exciting. But under any scenario, there will be a tremendous volume of material held by collections that will not be covered by such mass schemes: there is a great deal of locally unique material. I think there is at least as much promise in digitizing local, rare, manuscript, media, and other special collections as there is in converting general collections. And the quantities are enormous.

- **by Exchange** – Collaboration among libraries in collection access can acquire a tremendous life of its own in the digital age, dependent on, of course, and limited by copyright and related issues. If your institution has digitized a half-million works, and my institution has digitized some similar volume in a different range of subject areas, we may strike a barter agreement, either by exchanging data (and thereby beginning to address digital preservation concerns) or by exchanging the right for our users to access the other's collection. The Digital Library Federation, of which I am pleased to be an active member, is currently planning a Distributed Online Digital Library, or DODL, the premise of which is to make the digital holdings of each participating library available to the others' user community. I suggest you track our progress with this cooperative model. Needless to say, there are many other collaborative models in discussion at present, some of which are reported upon in this conference.

We know from many sources, not least of which is the "E-Journal User Study" that Stanford conducted with support of the Andrew W. Mellon Foundation a few years ago, that scholars are extremely interested in depth and breadth of the serial literatures.¹ We know from usage logs of the HighWire Press client publications that, while the number of clicks on a given article taper off very rapidly after publication, the number of clicks does not go to zero. Rather, there is what statisticians call a very long tail to the usage curve. One of the primary desiderata for e-journal users is online access to the backset, to old issues of journals, in some cases back the full history of the journal. And, of course, they want the same flexibility of searching, linking, viewing options they have come to rely on for the current issues of the same journals. The demand is there, I assure you.

We are spending a great deal already on digital resources, though in ways I find almost reprehensible, as I have argued elsewhere.[1] As reported in the most recent *Charleston Report* (v 9, no. 1), a recent report by ARL (the Association of Research Libraries) notes that expenditures on electronic resources between FY95 and FY02 for the typical university research library grew nearly 400% to almost \$1.4 million. During the same period the overall library materials expenditures grew 61%. Electronic journals accounted for 92% of the e-resource expenditures in FY02 and 26% of the library's overall serials expenditures, up from 5% in FY95. Expenditures on electronic books or other one-time purchases account for less than 4% of current monographic expenditures.[2]

¹ See: <http://ejust.stanford.edu>

At Stanford, we are also dedicating increasing proportion of our internal effort to digitizing materials – at the rate currently of hundreds of thousands of documents and thousands of books a year. Although this is not the place to go into detail, I have been very deeply involved in trying to increase the rate of digitization efforts and funding for such efforts, by several orders of magnitude. I believe we are on the threshold of a revolution in conversion of analog content to digital form. Frankly, the problem space – which began as mainly technical – focuses even more on intellectual property issues than on funding. I think you will see some exciting headlines in this area in the next months or at most years.

Making Sense of Digital Collections

We know quite a bit about how users find research material, online and otherwise, and we can say that the digital age introduces literally unimagined possibilities for discovery and analysis. Clearly, we do not know enough about how users behave and which tools of the future they will embrace. Here are just a few of the potentially profound issues having to do with making one's way through the literatures:

- Searching vs. browsing
- Searching across genres, across collections, across institutions, across artificial boundaries – even across the digital-analog frontier
- Searching tools: Google et al. work where they work, and fail dismally otherwise.
- Controlled vs. uncontrolled lexicons
- Taxonomic vs. text searching
- Organizing search results, including visualization tools
- Saving, as well as refreshing, search results – both links and content
- Annotating search results persistently
- Using technology to make value judgments, e.g., to favor trusted sources and suppress dubious material in transparent and controllable ways
- Text analysis
- Image analysis
- Linking of references within the literatures, especially Toll-Free Linking
- Social evaluation of literature – knowing how many others have read or cited a work and even knowing who those readers are (notwithstanding issues of privacy)
- Alert services and other customized means of selective dissemination of information.

We cannot delve into all of these issues at this time, but the messages I would leave you with are that:

- Digital discovery is very much in its infancy.
- Despite some seminal work, at Yahoo, at Google, and even within the library-supported areas of abstracting and indexing and other subscription-driven resources, librarians have not assumed a leadership role to date in developing discovery methods or interpreting user needs and behaviors for developers.

- Librarians can and should have a vital role in helping to develop, test, and promulgate emerging modes of digital discovery. Not to do so would be a real loss to the academy, to the profession, and to the world at large.

Let me focus on searching for a moment, as much of what is listed above bears on what we understand generally as searching. If I wish to search a query exhaustively at Stanford, depending on the discipline, I may need to conduct dozens of separate queries, using dozens of user interfaces, dozens of search rules or conventions. And each search result will come up in a different form, in a separate page, with different caveats as to coverage, currency, depth, etc. Some of the separate search results will be more relevant than others, of course, but some will be filled with spurious hits or duplicates or dead links or links to sources I am not allowed to view. And, of course, no matter how diligent I am using the considerable wealth of sources available to me at Stanford, the probability is low that my efforts will be truly exhaustive.

Despite our efforts as librarians and information professionals to provide controlled and competent tools to our users, the net result is fundamentally uncontrolled and uncontrollable, just like the public web. As some of my colleagues enjoy saying, "Get used to it." The price of vastly increasing access to information in an "information age" is reduction of control.

Please do not misunderstand me: I do not advocate a chaotic information space; in fact, I strongly support our utmost diligence in assisting our readers/users to navigate, understand, and exploit information in relatively controlled ways. But I also think we need to be clear that the library traditions of highly controlled information spaces is a dying anachronism.

The key idea, I believe, is not to try to control the information space, as we used to, but rather to try to provide the user with an armamentarium of tools and techniques to forge a path of meaning through the sector of the information space they wish to explore. Specifically, we need to provide better searching capabilities to our readers. For starters, the tools we offer on our home pages, the OPAC catalog searches and the like, need to become a great deal more powerful. Federated searching, a fancy way of saying searching multiple databases at once, really must become the rule, rather than the exception, such that we don't force the user to predict where the material of interest is hidden or what form of ownership it takes. It is a simple and fundamental point, and one that we can influence, in developing services and encouraging our system and service vendors to improve. There will be much behind-the-scenes development of search standards, proxy management, and like technical support for federated searching. I do not think its importance can be overemphasized.

There will also be a great deal of organizational cooperation to make federated searching more effective. Let me mention the DODL in the context of cross-institutional discovery, retrieval, and sharing. Even though the DODL is not yet underway (early technical meetings are ongoing), the very fact that a dozen institutions are committed to trying to develop some means of crossing boundaries is an encouraging sign. I look forward to the day – perhaps not too far in the distance – when a researcher at Stanford can find and retrieve a digital text from, say, Virginia, without even realizing that it was not local.

Taxonomic indexing, as an adjunct to uncontrolled term searching, is a challenging, but very valuable method of organizing a literature. The easiest way to explain it is to show an example, and the HighWire Press topic map serves well.² By clicking on “TopicMap” at the lower center of the HighWire home page, we see graphically the relatedness among topics, clusters of meaningful likeness, and a means of understanding relations among sub disciplines. This topic map – shared among the 361 journals supported by HighWire – was built by one person, an ex-cataloger, and accommodates some 1.4 million articles. Note also there are over 54,000 topics in this relational set – a relatively literal “web of science.” Populating these topics has required a great deal of coordination with the publishers, but it is in essence an automated process, and one we think very powerful and promising.

Next, sorting and making sense of the search result are vital areas. At Stanford, for example, we are in the middle of a year-long joint development with a software producer called Groxis to provide another visual aid in organizing search results. Their product is called Grokker.³

Compare a Grokker set map, a visualization of a search result, against the now-traditional set of search results: it is virtually impossible to get a sense of the hundreds or thousands of hits from the latter, all one can do is scroll through a page at a time. As a means of finding a known site or a common fact, this is a fine technique – I am sure we have all been pleasantly surprised to find the one page we need at the top of the Google search result list. But as a means of understanding an information space or a range of sources, it is completely inadequate.

Let me mention in passing that Grokker allows a form of federated searching and enables handling several sorts of data to coexist and be managed together by the user.

We can understand the taxonomic or topic map as a form of browsing, as opposed to searching. We know that browsing – more or less in its traditional definition independent of Web “browsers”- is a valued and productive form of library discovery. For a century and more, we have been putting like together with like in the hope that this collocation assists the reader. That is still an important aspect of the digital age. For example, a recent article in the *New York Times* [3] talks to this point: “the Internet cannot replace many of the built-in benefits of the library, like browsing the stacks for related information that could add spark and depth to an essay or a report.” That article goes on to discuss tools being developed at Berkeley and at IBM to assist “cross-linking” descriptive elements of collections, in the case of Berkeley’s Flamenco system, and to explore use of a “collection layer” in organizing folders and files and links, in the IBM example. I think we will see a great deal of important growth in ways to simulate the human capacity for association, fuzzy logic and sets, and partial or sloppy matching of ideas, as opposed to matching of terms, which is much of what we have available now.

A critical element in facilitating research is citation linking. This is something HighWire Press has emphasized in its e-journal services since 1995, and our research verifies that users value this extremely highly. Good scholarship requires checking

² <http://highwire.Stanford.edu>

³ Grokker can be downloaded at <http://www.groxis.com>

and rechecking, including examination of worked cited by other scholars. Given the explosive growth of all the literatures in recent decades, resulting in the literal impossibility for scholars to cover all the relevant literatures, making it easy to get to referenced citations is absolutely critical to scholarly productivity. Thus, with HighWire, not only can the reader get from a reference in one article to the full text of that article in another journal with a single click, her or she may do so without charge, actually free to the reader, and whether or not the scholar otherwise has access rights to the journal in which the cited article appears. This is genuinely profound, and I am extremely proud that HighWire led this development. Of course, this requires both technology and a rethinking of stakeholder interests, but it is certainly the only way to go. There are rumors currently that JSTOR is about to re-code its whole system and corpus of literature so citation linking can be accomplished, and I fervently hope the rumors are true. I have long felt that the lack of such linking has compromised the utility or even viability of the otherwise exemplary JSTOR model.

Another form of assessing one's potential interest in a particular source is learning how many others (and perhaps who) have made use of it. This is actually a very popular and common technique: vide the New York Times Best Seller List, amazon.com reader reviews and rankings, Oprah Winfrey's Book Club list, and simple word of mouth. In fact, this is the dominant mode of evaluating literature in the general public: "Everybody's reading it"; "This summer's must-read novel"; even, "Soon to be a major motion picture." These are absolutely rational determinants (even if they are easily manipulated by marketers) for deciding what to pick up from the airport bookshop. And while this does not sound dignified enough for the scholarly realm, analogous methods could be used in our world. If I knew that the futurist Paul Saffo and the business guru Tom Peters both have studied a certain new book on information practices, you can be fairly sure I would make it my business to check it out. If I knew that Stanford electrical engineer Stephen Boyd personally subscribes to a certain journal, I would certainly advise my friend's daughter, a promising young engineer, to start reading it. If I were an ambitious graduate student in bio-informatics casting about for promising research areas, I would be very interested in what Stanford's Doug Brutlag has been reading lately, whether or not he cites it in recent research results. We all need to know what is hot, whether deservedly or not.

Thus, social aspects of the literatures are important. We as research librarians have all but ignored this – our colleagues in public libraries buying multiple copies of best sellers are far ahead of us in this respect. We do have some tools, however, such as the concept of knowledge communities. At Stanford, we have worked with publishers and scholars to develop several Knowledge Environments, which support in various ways the idea of virtual knowledge communities.⁴ I believe there is much more we could do in this area, ranging from statistical reporting on demand ("this item has been downloaded 34 times this quarter") to supporting peer-to-peer communications based on the literatures. Amazon.com does this extremely well, and it would be tragic for us as librarians to ignore these models of success.

⁴ See, for instance: <http://stke.sciencemag.org>

Another form of discovery that works well – both for Amazon and for HighWire Press – is individual email with readers, based on sign ups for alerting services, whether periodical – e.g., advance Tables of Contents– or topical, e.g., abstracts of newly published articles that may be of interest to me, based on preferences I established. Within a few years, major research institutions will be adding access to multiple millions of digital objects a year – databases, articles, books, images, models, syllabi, drafts, presentations, etc. If we don't have means in place to notify users of material they may need, but not know about in advance, no amount of search capability can enable them to keep up with this onslaught. There is a huge opportunity for us, and one that our traditional efforts in this area cannot begin to address.

Integrating Digital Material

Perhaps for the best of reasons, librarians have really not started to think very much about the results of research except in the context of published results. Let's take a look for a moment at what we may call the traditional scholarly cycle:

- Scholar performs research (including examination of literature via the library)
- Scholar writes paper
- Scholar submits paper to journal (or book to editor)
- Editor manages peer review, editing, format, etc.
- Publisher publishes paper or book
- Libraries subscribe to journal or buy book
- Libraries make materials accessible, discoverable, etc. – perhaps by subscribing to discover services – so that cycle can begin again

There are other, somewhat similar, cycles for student papers, conference proceedings, and other essential elements in academia.

As we have seen over the past few years, this cycle is under attack, in large part because certain publishers – and I emphasize certain publishers, particularly European for-profit publishers, but not the scholarly society publishers – have extracted obscene profits with outrageous subscription price increases. Certain others, including scholars and bureaucrats, have countered by promoting alternate publication models, taking some slight advantage of the fact that online editions eliminate printing and mailing costs. As many of you may be aware I strongly support the publication models of responsible publishers, mainly not-for-profit scholarly societies, such as those that use the Stanford HighWire Press service for online editions. I also advocate and show by example resistance to the “big deals” and escalating pricing of those certain other publishers, in other words, to encourage libraries not to remain hapless and passive victims of commercial predation.

But what I wish to emphasize here is another kind of activism, namely, the active role libraries can and should have in the future of the scholarly cycle. Here is one scenario among many to illustrate the role of a more active and effective digital library:

- Scholar searches online digital information acquired or created by library, using discovery tools, such as federated or taxonomic searching, provided by library.
- Scholar organizes search results using visualization tools provided by library.
- Scholar organizes and stores digital information in library-provided online workspace. Library-supported tool allows local, remote, and original materials to be handled in common, secure, flexible environment.
- Scholar uses digital tools provided by library to extract, model, reformat, and embed earlier digital information in work, which becomes the “research paper.”
- Scholar posts draft “paper” to library-supported institutional repository, along with supporting data and literature, and makes them available to specific colleagues.
- Scholar revises “paper” and releases final version to the repository and the public.
- A virtual electronic journal, after a review process, links to the library repository as the published edition of record, and the library provides various kinds of access to the paper and its supporting documents.

Note that the scholar in this scenario may or may not physically visit the library. The scholar may or may not have discussed what material is of interest with library curators. The key materials may be primary source documents, digital representations of primary source documents, private notes, or the research literatures. They may be texts, hypertexts, databases, visual materials, models, simulations, or some other carrier of information. Indeed, the scholar may not fully realize.

Conclusion

Our traditional measures of library services and, indeed, library empires – e.g., the infamous ARL annual statistics – are absolutely silent on effectiveness. At best, most of us conduct an occasional “satisfaction” survey. Such surveys are important and reveal some interesting insights, but the brutal truth is we librarians have never yet begun to measure results. That may persist into the digital age.

I come back to the user or reader as the measure of a truly effective digital library. We could stipulate as an easy approximation that a truly effective digital library is one in which the user gets what he or she needs or wants. It should be clear to you that I do not think this begins to address the challenge. Perhaps the right measure is something like this:

The truly effective digital library is one which anticipates, addresses, and tests, the way scholars seek, understand, handle, store, annotate, synthesize, manipulate, and publish information such that it is integral, whether or not visible, to the entire scholarly enterprise and supports the productivity of scholars’ research, thinking and expression.

In sum, the truly effective digital library, freed in part from the shackles of physicality, is the information infrastructure of the academy, not a support service, but rather the backbone of intellectual effort, whether applied to teaching, learning, or research. I don’t know whether libraries are capable of this transcendent mission, and I know I can at best dimly understand its implications, but I am fairly sure that is what constitutes a truly effective digital library.

References

1. Keller, M. Casting Forward: collection development after mass digitization. Fiesole Collection Development Retreat, March 18th 2004
http://digital.casalini.it/retreat/2004_docs/KellerMichael.pdf
2. E-Journals Take Lion's Share of E-Resource Budgets. *ARL Bimonthly Report* 235, August 2004 <http://www.arl.org/newsltr/235/snapshot.html>.
3. Eisenberg, A. Making a Web Search Feel Like a Stroll in the Library. *New York Times*, August 19, 2004.