ESTABLISHING A DIGITAL LIBRARY

White Paper
February 2009

Michael A. Keller
University Librarian
Stanford University

Preface

Seeing a growing need from my partner and customer discussions globally, I asked Michael Keller of Stanford, one of the leaders in the Digital Library Community and the co-founder of the Sun Preservation and Archiving Special Interest Group (Sun PASIG; www.sun-pasig.org) to set out his vision on how to establish a digital library in today's technology environment. We wanted to have a document that could give both librarians and IT professionals an overview of the key functions of a digital library and how they map into the requirements of the 21st Century information society.

We both also want to use this white paper as a 'living document' that can be extended and deepened over time through input from both the Library and IT communities. We openly invite comments and elaborations on threads in this ‘getting started’ piece.

I would like to thank Michael for his work, commitment, and guidance. I hope you find “Establishing a Digital Library” useful!

Art Pasquinelli
Education Market Strategist
Sun Microsystems, Inc.
art.pasquinelli@sun.com

About the Author

Michael A. Keller is Stanford’s University Librarian, Director of Academic Information Resources, Founder of HighWire Press, and Publisher of Stanford University Press. He has led libraries at Cornell, UC/Berkeley, Yale, and Stanford. Keller’s board service includes Hamilton College, Long Now Foundation, Japan’s National Institute for Informatics, and National Library of China. He is a guest professor at the Chinese Academy of Sciences, Senior Presidential Fellow of the Council on Library and Information Resources and 2008 Fellow of the American Association for the Advancement of Science. Advisor and consultant to numerous scientific and scholarly societies as well as for the city of Ferrara, Italy, Newsweek magazine, Princeton and Indiana Universities, as well as the national Library of China, and King Abdullah University of Science and Technology, he was a Siemens Stiftung Lecturer in 2008. Keller, with his colleague Art Pasquinelli of Sun Microsystems, is the co-developer and co-chair of the Preservation and Archiving Special Interest Group.

Table of Contents

This Page Intentionally Left Blank

Chapter 1

Introduction

Fifteen years after the introduction of the Mosaic browser, almost 35 years since the term “Internet” was first used, and almost 20 years since the phrase World Wide Web was coined, progress in information technology developments, innovation in publishing and communication, and enough experience by users of the World Wide Web have led us to some common understandings of what digital libraries should be and should do. This paper outlines some of the expectations and requirements for digital libraries as well as some observations about the implementation phase for what might be regarded as first and widespread attempts to construct and operate full integrated digital libraries on the basis of those expectations and requirements.

Expectations and requirements of users will be described and illustrated. Insights into the functional specifications necessary for digital libraries to be considered successful in this new phase will be cited. Components without which digital libraries in this coming phase might fail both for current expectations and for stage setting for the next phase will be described and illustrated. Among these components are essential ones like digital rights management, authentication and authorization of users, preservation of digital objects for the long-term, and digital archives for convenient and flexible access to all sorts of digital objects.

The perspective of this author is that of a senior officer at Stanford University responsible for the university’s libraries, academic computing, and publishing organizations, operations, and enterprises.

Chapter 2

Situation Report

Let's look first at the stage we are in now, the precursor to the integration phase of the integrated digital library.

In the publicly available Web as of June 2008, there may be as many as 63 billion indexed web pages from about 104 million sites. However, the vast majority of documents on the Web are in the deep web, the access-controlled web; they number more than 550 billion documents. So, at best, Google is indexing as much of the publicly accessible web, but that amount is roughly 12.5% of the total size of the web as measured in documents. At most large universities, upwards of 1,000 databases on various subjects are provided to authorized members of each university community. Those databases are part of the deep web, as are the tens of thousands of e-books, e-journals, and other access controlled sites, including those for movies and music. At those same large universities, databases of meta-information provide access to the contents of physical collections. There are on-line public access catalogs (OPACs), those improved versions of the old card catalogs. There are indexing and abstracting services that help crack the contents of lots of anthologies, collections, and journals. There are as well numerous reference works, both those that are re-cast from pre-net versions, such as the Oxford English Dictionary on-line, the Encyclopedia Britannica on-line, the Grove Dictionary of Art, and the Engineering Index, as well as those that have only existed in digital form accessible through the Web, such as CSA's Illumina, a collection of databases that cover major areas of research, including materials science, environmental sciences and pollution management, biological sciences, aquatic sciences and fisheries, biotechnology, engineering, computer science, sociology, art history, and linguistics and the Children's Literature Comprehensive Database.

Wikipedia and similar products of the net provide mostly free and most often quite relevant, if not entirely authoritative information on millions of topics. YouTube, FaceBook, and MySpace provide services that enable "netizens" anywhere with the capacity to make public, that is "to publish", videos, biographical information, and commentary that may or may not be authoritative and accurate. Beyond those, there are hundreds of thousands of still and moving images available on the web now too. In short, whereas there has been an extraordinary increase in information and knowledge available for research and study as well as significant improvements in the means to discover information that is potentially relevant because of the advances and accomplishments of the digital world, the truth is that readers and users, students and

professors encounter difficulties in penetrating the thicket of information resources, especially in conducting deep and systematic searches. Google, Yahoo, and the other indexers and catalogers of the Web have helped a great deal through their services, but their efforts are largely limited to web sites and web documents that are publicly accessible.

Libraries and librarians, publishers, indexers and abstractors have gone a long way in organizing the chaos of academic information resources, but there are too many catalogs, indices, finding aids, guides, and knowledge maps for anyone but the most assiduous subject specialist to master. Google, Yahoo, and the other indexers and catalogers of the Web have helped a great deal through their services, but their efforts are largely limited to web sites and web documents that are publicly accessible. Some of the work of the dogged scholar seeking to master all the literature of a particular topic has been made easier by Web services, especially Web indexing, but some surveys show that that ease is traded against the superficiality or shallowness of the content on the Web, so that the need for real sleuthing by scholars is still very much needed. Perhaps the Google Book Search project, which is now publicly presenting the settlement of grievances against it by authors and publishers, in the long run of time will make scholarly sleuthing ever easier and reduce that detective work to archives and rare books, information sources quite unlikely to be digitized in the coming several decades.

As an example, if one were to wish to gather information about the composition, literary-historical sources, and performance history of Carl Orff's Carmina Burana, one would look in two major journal indices, several OPACs, numerous encyclopedia and dictionaries (in multiple languages including Czech and Russian!), indices and collections of newspapers, a good dozen recorded music reviewing services, as well as various monographs on Orff and Medieval German poetry. There is much on the Web too. It is no wonder, then, that superficial, if mostly accurate and too brief, sources of information like the Wikipedia, are employed by novices in any discipline or topic. Worse yet, some readers turn to such limited, if convenient, collections of information as searchable titles in the Amazon on-line book store.

Another way to look at the situation is to think of all the information, knowledge, and opinion ever recorded in this world as a kind of jungle. Google, Yahoo, and other web indexers provide some ways to identify individual flowers, plants, lizards, butterflies, birds, and monkeys in the jungle, while librarians, editors, publishers, and similar knowledge or information managers provide routes and trails through the jungle so that one might find and select the sites and the objects of one's interests. There are lots of words in lots of languages in the indexes and lots of trails through the jungle.

Adding to the density of the jungle, to the complexity and depth of the information and meta-information available digitally are the numerous large scale digitization projects, among them the Million Books Project, the Open Content Alliance, and the Google Book Search Project. Tens of millions of books could be treated by these projects. Beyond those widely known efforts to transform the contents of printed, physical books to digital objects for indexing and in the cases of public domain works and books whose rights holders have permitted on-line views of pages, for reading, there are literally thousands of projects digitizing all kinds of specialized material, from Ancient and Medieval manuscripts to rare printed books to archival documents to government publications to glass slides of historical events or persons. A few examples of these are worth mentioning. The Matthew Parker Online Library project is providing Web-based images of pages from the 537 manuscript books given by Elizabeth I of England's first Archbishop of Canterbury to his alma mater, Corpus Christi College at Cambridge University. The project is providing improved descriptions, bibliographies of secondary literature and modern editions, as well as numerous navigation tools as well as superb digital images in several resolutions of all the pages of these manuscripts. One should be aware of the investigation of Romanesque church buildings in the Bourbonnais conducted on the basis of data gathered through laser surveying equipment; that project is delivering new insights into Medieval construction methods. There are private projects digitizing rare books too, a fine example of which can be seen at http://www.rarebookroom.org.

Other problems of access to published works abound. For instance, while there is a terrific text base of over one million Chinese works published in the People's Republic of China (i.e. From 1949), it has only a few works from the Ming and Qing dynasties and very little from the period of the Republic of China (1912-1949) and, of course, very little from the Republic of China (Taiwan). Much work remains to be done there. Another excellent example is that of the classic literature of Spain. The Biblioteca Virtual Miguel de Cervantes, based at the University of Alicante in Spain, has amassed by keyboarding about 22,000 published works of classics of Spanish literature over the past 10 years. There is much to be done there too. In the category of archives, the government of Alsace has undertaken the conversion of thousands of volumes of notarial records using the automated digitizing hardware and software of 4 Digital Books, in order to provide deeper information about marriages, deaths, the transfer of real property, among the many other events in the lives of Alsatians over the past two centuries.

Stanford and the World Trade Organization have digitized most, but not all, of the archival records of the General Agreement on Tariff and Trade, the predecessor organization of the WTO in order to facilitate scholarship and thereby understanding of the bi- and multi-lateral trade agreements that shaped and stabilized the global trading environment from the end of the Second World War. Understanding and appreciating those developments will result from the scholarship yet to be performed on that digitized archive, for much of it is still restricted. Finally, documents of the distant past, in virtual versions of themselves, not modern editions of their texts, are beginning to appear on the Web. There are marvelous projects underway at the Stifts bibliothek St. Gallen digitizing the manuscripts of that famous scriptorium that was never over-run or raided in its long history. There are similar projects at the British Library, Corpus Christi College, Cambridge on the Matthew Parker Library and at the Archdiocese of Cologne, among many other locations in Europe, that will make more research on more Ancient, Medieval, Renaissance, and Early Modern manuscripts and archives possible for scholars and students for whom travel to the repositories holding these treasures is difficult, constrained by time, and expensive. The wealth of information and source material on the Web is ever expanding, though authenticity and accuracy are still issues, and all searchers of the Web are advised to doubt and then verify what they are reading.

In short, in the proliferation of information and knowledge available in physical and digital libraries, neither the meta-information tools, whether in traditional or digital forms, nor the huge indexing and discovery services on the Web — Google, Yahoo, and the like — have made ferreting out the information from the combined set of all possibilities very easy. Neither the commercial sector nor their not-for-profit counterparts have succeeded in simplifying the discovery process.

There have been some noble efforts in simplifying discovery involving the collecting of a small number of relevant meta-information data sets and re-coding them so that all the data records look very similar in the synthetic set. The most successful of these is the one engineered at the Research Library of the Los Alamos National Laboratory under Rick Luce, who is now Vice Provost and Director of Libraries at Emory University. It is known as SearchPlus and supports searching using a single search argument on BIOSIS (1969-), Engineering Index (1884-), Inspec (1898-), and the ISI citation indices (Arts & Humanities, 1975- ; SciSearch, 1900- ; and Social SciSearch, 1973-).

SearchPlus is remarkably sophisticated in many ways and the services it offers to its users (basic and advanced search, cited browse, cited search, marked records, search history) provide some markers for functions desirable in the next phase of the development of digital libraries, the integrations phase. LANL's Research Library has built on top of SearchPlus a service known as FlashPoint, a good example of the other approach to providing federated searching to numerous meta-information sources. FlashPoint supports searching using a single search argument across the consolidated databases in SearchPlus and two others that are not consolidated into a single database, MathSciNet (1940-) and PubMed (1951-).

There have been several attempts to develop a federated search engine that could address many meta-information databases, but none have been markedly successful...yet.

The chaos that is the Web can be navigated after a fashion by using Web indexing and cataloging services, such as those provided by Yahoo, Google, A9, Ask.com, Dogpile, and so forth, but the results of searches in those services vary in relevance to the interests of the searcher. Google's results are conditioned by the nuances of the elements in the Page Rank scheme, many of which elements are not known, for instance. Re-ordering and applying visualization functions to the results of searches may help speed the assessment of relevance after a search has been performed. A good example of the possibilities of this post-search approach is that of Grokker, a product of the Groxis Company; Groxis gathers results based on common words in the html headers so that in the visual map web documents appear clustered together concentrically, permitting rapid choices of documents containing words of direct relevance and rapid discarding of document clusters containing irrelevant words.

Libraries and librarians do their best to identify effective meta-information publications and databases. They become experienced in the vagaries of the Web search services so they can advise readers on what works and why, but as well what the hidden pockets of information might be. An ordered set of numerous information resources are assembled by libraries and librarians as subsets or even counterpoises to the disorderly Web, which is in constant turmoil and constantly growing, albeit in unknown ways.

This recitation is meant to illustrate briefly the current situation those involved in the knowledge generation and communication trades face. The integration phase of development of digital libraries is underway now.

Chapter 3

Elements of the Integration Phase in the Development of Digital Libraries

The previous section mentions some of the elements that must be combined to comprise the integrated digital library. Assembling those elements in a design in the following graphic is a way to illustrate both the constituents as well as their relationships in one version of the integrated digital library.

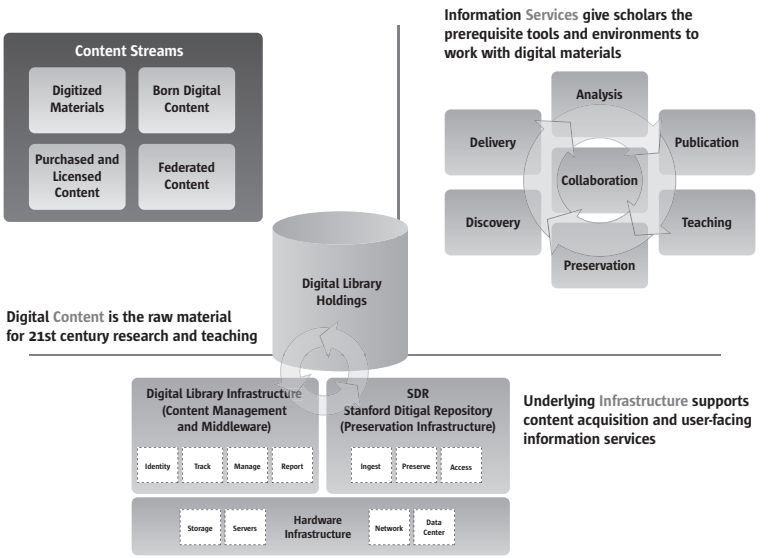

Figure 1.

Digital content and meta-data about physical content must be contained, controlled, and in many ways understood by the digital library infrastructure in order for the basic functions of content management and content preservation to occur, even as the collections of digital objects, digital information, and meta-information grow. Among the digital genres of content are e-journals (and their essential parts, e-articles), e-books (both born and circulated digitally as well as digitized ones), newspapers, government documents, Media objects (films, still images, sound recordings, diagrams, plans and the like), Java™ models, spreadsheets, and so forth. Much of that content is and will be

licensed; some of it will be owned outright and some of it will be in the public domain. To the greatest extent possible, the digital objects forming the content of the digital library will be held locally, in the local digital library and digital preservation infrastructures. Some of it will be held in distant locations, in which cases, meta-information about the digital objects will be locally held.

The digital library infrastructure will also support the basic functions of libraries: selection and acquisition; provision of intellectual access for discovery (cataloging and indexing); circulation and distribution; interpretation and instruction (reference and research consultation services as well as basic information literacy and information heuristic instruction); preservation; analysis, manipulation, presentation, and display (services supporting the use of digital objects in academic, commercial, and practical settings).

Paired with the digital library is the digital archive, the assemblage of functions and processes involving ingestion, preservation, and provision of access.

Scholars, students, and inquirers are provided with a range of services, many of them digital that operate on and/or through the digital content. Among the services required are: discovery (searching in several modes and using several methods, indexing in a variety of ways); delivery of data, texts, and images in usable modes through as few as possible reading or viewing environments including streaming media; alerts and recommendations related to known or experientially based (perceived) parameters; data- and text-mining, extraction, and analysis; combining digital objects and data to form new views, analyses, and expressions (e.g. overlay of data in geospatial information systems approaches); annotation and labeling of digital objects for private and broader distribution; publication (formal and informal); managing courses and presenting course materials (including all the administrative aspects of teaching as well as testing and assessment); research collaboration (local and global); portfolio management for educational, career, and personal purposes.

Implementation of the integrated digital library requires a hardware and software infrastructure that has only lately become possible to assemble. Of course servers, spinning magnetic disks, tape drives, and the like are required, but more, much more is necessary as well. Workflows conceived of from the consideration of selection for the inclusion of digital content are necessary. RAID disks and easy management of digital objects into tiered storage are required; on-line, near-line, and off-line digital copies are needed to make cost-effective use and access possible, as well as to effect

storage, and long-term preservation of digital objects. Robust, high bandwidth, low latency internal networks with good links to the public networks are required. Digital rights management (and the related capacity to respond to misuse of resources), authorization of users, firewalls, computer and network use policies with effective enforcement mechanisms, uninterruptible electrical supplies, secure and capacious data centers, backup services, and failover systems are all required.

Stanford makes use of a raft of Sun equipment and software, both in the Stanford Digital Repository as well as HighWire Press. We employ a range of Sun SPARC® servers, about two dozen of them, running Solaris™ 8 & 9 in support of databases and applications. We also employ about 30 Sun™ x86 servers running Red Hat Linux and Solaris 10; among these are X2200s, X4100s, X4200s, X4500s, and X4600s. There is a Sun L700 tape library at work as well as two Honeycomb arrays. And we are beginning to use SAM-QFS. Naturally for years we have employed Java, MySQL™, and Solaris. At HighWire Press, there is a similar array, but as well about 4 dozen Sun Blade™ 100s, numerous storage arrays, and, of course, many servers, from the Sun Ultra™ line to E, T, V, and X series machines; in all HighWire employs 219 Sun devices and of course makes use of Solaris, Java, and various of the Sun storage applications managing storage.

Two major inefficiencies have plagued the development of the integrated digital library.

One is the management of digital objects with as little input from humans as possible. For years, complex systems involving multiple and divergent decisions in numerous places in the workflow of digital libraries and archives were planned. Recently a new approach, SAM-QFS, was made an open source product by Sun. SAM-QFS is a shared file system and a storage archive management application. It allows digital objects and data to be managed across a tiered storage array based on the value assigned to the data or digital object. In the integrated digital library, this combination of functions will reduce human intervention, but provide known assignments of value and location essential to long-term preservation and access. Because Sun has made the Solaris operating system open source as well and because Sun servers, magnetic disk arrays, and tape robots, most of the main components of the hardware and software architecture necessary for the digital library and digital archive are in place. When combined with the open source Fedora package of modules (repository, preservation, semantic, and enterprise services), which are both scalable and configurable, there is little that remains to be brought to the architecture to enable a robust digital library or archive.

The other inefficiency is that of creating or capturing descriptive meta-data for each digital object and versions thereof. Presently we are all dependent upon the application of skilled catalogers to the task. They apply various rules and principles, assemble the descriptive information in standard formats, and relate objects in various ways. All of this takes time and frankly, the flow of digital objects to the digital libraries systems is so great and becoming greater that this approach takes too much time and thus costs too much. Among the questions that are asked in assessing alternate methods or the re-engineering of the present way are these: what can the originators of the born digital content do to provide meta-data; what can be captured automatically; what descriptive elements can be discarded and replaced with alternate modes of intellectual access (such as search in which the texts or selected portions of them become part of the meta-data), automatically derived abstracts, or even reference to related objects; what collaborative efforts resulting in coordinated meta-data creation and then sharing through a trusted third party might be set in place? Specifying digital rights meta-information for each object by hand is another inefficiency

A minor inefficiency involves ingestion. Some digital archives are experiencing overload at the ingestion stage beyond the problem of meta-data creation. Batch processing, high performance computing solutions, and startling increases in ingestion pathways are all considered or under test.

Chapter 4

Regarding Preservation and Access

The roles and functions of libraries and archives over the long run of history have always involved the transmission of the ideas and expression emanating from prior generations to future ones. Until the last quarter of the 19th-century, paper, made almost always from good rag content, and parchment were good media for long-term retention, provided of course, books and manuscripts were not subject to the privations of fire, water, rodents, insects, and war. Paper made from wood pulp in the late 19th-century and therefore laden with acid salts is now self-destructing. While there are methods for dealing with that physical and chemical problem, in fact, there are few mass deacidification programs in place and operating. The expense is too great and the methods as yet too inefficient. Mass digitization of printed materials is coming to the rescue. In the longer run, paper laden with acid salts eventually will self destruct. Yet most of those books are not valuable as artifacts; surely we should save some of them, but not all of them. Mass digitization captures the pages, potentially with great accuracy, and permits us to save page views independent of the fragile paper carrier of the information. And of course because we have a digital image of the pages, we can understand the words on those pages by way of optical character recognition. Indexing of the words and phrases then become possible, as do associative, taxonomic, and semantic indexing, all done algorithmically. Mass digitization has become an essential step, then, in preserving knowledge current and available today in one or another physical medium, into knowledge and information available today, tomorrow, next year, next century, and perhaps next millennium for use, re-use, mixing, and re-mixing.

A step beyond mass digitization and a step necessary for information and knowledge born and distributed digitally is the incorporation of those digital objects in a digital library for use by humans. In order, however, to assure that the digital objects in the digital library are available for generations to come, those objects must reside as well in a digital archive, in which the objects are not subject to modification, are understood as to the versions of the operating systems, applications for reading or viewing, and the data formats embodying them so that as those key technologies change, the digital objects required technical contexts and the data itself can be migrated to newer versions. One alternate strategy for dealing with the rush of revisions natural to digital technologies are encapsulation of digital objects and their required technical contexts in a sort of digital envelope, one that is ready to be opened in the future for use and study. Another strategy involves the creation of very robust and highly flexible emulation software so that the essential functions and characteristics of any digital object deposited now in a digital archive could be re-invigorated, substantiated again, sometime far in the future.

Regardless of the strategy, and migration is the one most widely adopted today, we can be sure that not all of the digital objects we deposit and not all of the functions and characteristics they presently exhibit will survive. For this reason, more research and development is necessary as well as the careful selection of what is to be placed in the digital archive. On the principle that what is used is effectively preserved, the selections we make now will determine the use ultimately to be made of them and from them the new knowledge that inevitably will be created.

Here is a schematic of a digital archive whose first principle is the preservation of digital objects, with access to them provided across an air gap via copies made from the original files. This method provides no opportunities for the mother files to be corrupted by acts of omission or commission.

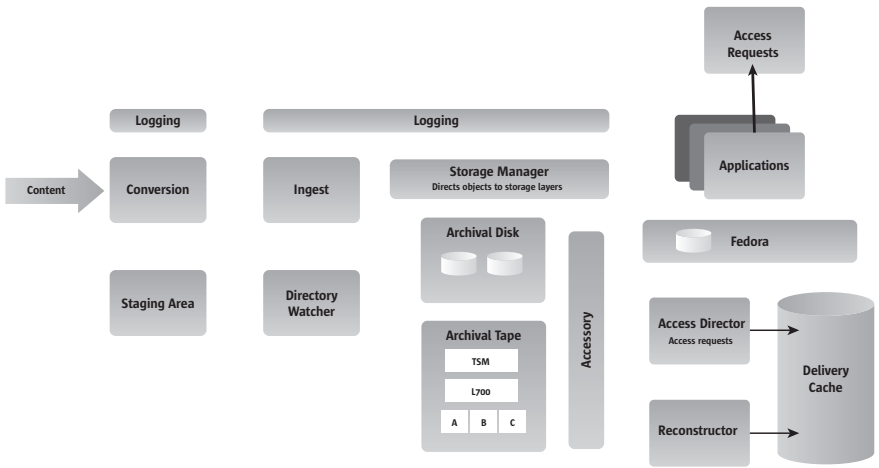

Figure 2.Digital Archive Logical Diagram

Chapter 5

Integrated functions in the Integrated Digital Library

The functional requirement that all digital objects, their content or their metadata be subject to the same discovery mechanisms simultaneously when a search is launched makes all the difference. Realizing that functional requirement will take some effort. New designs, new applications or much revised versions of existing ones, and new tools will be necessary. From the users perspectives, we know that browsing in various contexts will include:

1. Browsing in classification order (accounting for unclassified items, as dwell as items in various classification systems will be special challenges)
2. Browsing among authors and responsible entities
3. Browsing among titles, including series titles
4. Browsing among subjects
5. Browsing among uniform or conventional titles
6. Browsing among imprint information elements (place, publisher, date)

And a related question is how to make that sort of browsing work in such a way to present options for looking more closely at the title page and perhaps the front matter, including tables of content and introductions of a work of potential interest.

A variety of searching and indexing methods need invoking too. The sort of searching options provided by most on-line public access catalogs in libraries involving choices to search by author, title, classification, subject headings, and so forth are necessary. Taxonomic indexing such as provided by HighWire Press for articles published through its services to e-journal publishers (provide instructions on how to find and use), while excellent by itself because of its basis on ideas, not just on expressions, is important because other services, including graphic navigation among objects and ideas as well as matching, alerting, and recommendation services can be built on it. Semantic indexing provides the means to extract proper nouns of all sorts and thus the practical possibility of combining indexes to arrive at more precise matches to a users' needs. Finally, associative searching as created by Professor Akihiko Takano of Japan's National Institute for Informatics in Tokyo (see WebCat Plus) is the realization of a non-semantic matching principle based on word frequency, word relationships, and word positions; associative searching matches users search terms, even very large sets of them such as paragraphs, chapters, and whole articles very precisely to all texts in very large text bases.

With regard to searching and indexing, it must be noted that there are substantial problems in multi-lingual searching and indexing. New scholarship by linguists, computational linguists, and computer scientists is needed to create approaches to enabling multi-language searching and indexing with closely relevant results. Another approach might involve the creation of a true meta-language, research on which is underway by those involved in the universal network language research.

In addition to browsing, searching, and indexing in integrated and integrative ways, the integrated digital library will also provide Web-based services, such as hyperlinking citations to cited references (as originated and implemented by HighWire Press), graphical navigation (as implemented by Groxis, HighWire, others), and virtual post-it notes (private, shared closely, shared widely, broadcast). Also, personalization and customization must be supported, along with the services users will need to create their own knowledge environments in numerous topics of interest, for their own convenience as well as for collaborating or sharing with others.

The recently announced settlement of the suit against Google Book Search by authors and publishers[1] has several ramifications for libraries, first in the U.S., but ultimately everywhere. First is the necessity of functional integration of any library's catalog and digital resources licensed from other sources than Google with the e-books in the Google database as well as the basic Google search services and especially Google Scholar. A second ramification applies mainly to participating and cooperating libraries with Google regarding the local archiving of digital copies of books sent to Google, both for local analysis, indexing, searching, and web services as well as for long-term archiving, a hedge against the inevitable morphing and, who knows, the demise of Google as a corporation. The third ramification is that of the services that one or two participating libraries will offer text mining services to scholars everywhere based on copies of all books provided to Google by any library. The research that may be performed on that database will mark an expansion of the services digital libraries may offer, from natural language processing tasks to semantic indexing and searching to data mining to use of the database of e-books as a testbed for new applications, including browsers and search engines to the development of better OCR applications for non-Roman character languages to the development of Web 2.0 and 3.0 services. Of course, with the availability of site licenses to the totality of e-books in the growing Google book database, all libraries (first in the U.S., but ultimately everywhere) will have the opportunity to offer a much larger collection of out of print books, both in copyright and in the public domain, than any single library could provide to their patrons. What will be the effects of the availability of that huge collection on teaching, learning, and research, whether in K-12 settings (though that is not a target market for Google), in higher education, or, most interestingly in public library and non-institutional settings. The on-going assessment of the effects of the availability of millions more books will be an important task that will benefit from library engagement.

1. See http://books.google.com/googlebooks/agreement/ and http://library.stanford.edu/about_sulair/special_projects/GoogleBooks.html for more information and more links.

Chapter 6

Conclusion

As much conceptual and functional ground as has been traversed since the beginning of the World Wide Web not two decades ago will be traversed in even less time in the next phase of digital library development. Among the challenges libraries will need to face is the acceptance, even the embrace of services and publicly available information resources that are far beyond what any individual library could provide. At the same time, libraries operating digital avatars of themselves need to look to the functions that major and minor corporations cannot or will not provide, perhaps cannot or will not provide over the long run of time.

The role of custodian of the digital emanations of our culture for access and for long-term preservation is and will be essential. The notion that any one or two or three companies or institutions, whether acting for a collective or driven by profit motives or acting under self-appointed eleemosynary principles, can and will suffice for the long-term is simply not supportable after even the most casual examination of the fate of such organizations. What really counts is the application of the professional attentions of librarians and cybrarians and their particularized systems in concert with their information technology colleagues in numerous institutions with long histories and longer futures to the traditional roles of libraries, albeit using new methods and tools on new formats and new media. The proliferation of methods, types, environments, and approaches spells the survival of the digital artifacts of civilization.

The practice of digital librarianship, cybrarianship, in many different ways and in many different places is exciting now and will be even more so in the future. The role of attentive, responsive and, yes, leading technology companies in providing hardware, software, architecture and design possibilities, to the community of cybrarians will be as influential and helpful as it has been in the past. Among the best of those companies has been Sun Microsystems. The combination of technologies produced for many different purposes in many different settings, technologies which are reliable and consistent, which can account for interoperability and modularity, and are usable in open systems architectures will be the ones that cybrarians and the particularized systems, approaches, and policies will adopt.

The course of the future cannot be clear, so the innovation arising from mutual understanding, from principles of responsiveness and responsibility, and the devising of sustainable and scalable digital libraries will be crucial to continuing the contributions we make to our local communities.

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN (9786) **Web** sun.com

SunWIN #555379 Lit. #SWWP14775-0 2/09