Published in Library Hi Tech
Special issue to commemorate 30 years of the journal, vol:30, iss:4

The 1990s: The Formative Years of Digital Libraries

William Y. Arms
August 2012

1. Background

During the past thirty years digital libraries have gone from a curiosity to mainstream. The 1990s were a particularly formative decade. Before 1990 computing in libraries had concentrated on metadata. MARC cataloguing had reached its zenith, indexing services such as Medline had developed sophisticated search languages, and Science Citation Index was the state-of-art in linked data, but with very few exceptions the actual collections were physical items such as printed documents. About 1990, computing reached a level where it became economically possible to mount large collections online and to access them over networks. The result was a flurry of experiments and prototypes. Many are almost forgotten, yet the libraries of today were formed by the energy and creativity of these efforts. This article describes some of these projects and the impact that they have had on modern libraries.

There was nothing inevitable about which prototypes succeeded in the long term. The Internet uses the TCP/IP family of protocols, but for many years the Open Systems Interconnection (OSI) framework was the choice of every major company. The early web was one of several competing ways to mount information online and had many weaknesses. Web searching did not have to be a free service paid for by advertising. Engineers might argue that the successful technology was technically superior, but cultural, economic, and social forces were at least as important in its adoption.

The article is not a full history of the period that is covered. Many important projects have been skipped. A more complete article would have additional examples from the commercial sector and would do more to place the American experience in a worldwide context.

1.1 Before 1990

Before there were computers, libraries were pioneers in using technology such as typewriters and microfilm. Beginning in the 1960s computing was widely used in libraries, but general-purpose computers such as the IBM 360 series were a long way from serving the needs of libraries. As a result, libraries developed specialized software that diverged from the mainstream of software development. Two early examples were the MARC format and Z39.50.

MARC dates from the 1960s as a format for exchanging catalog records on magnetic tape (Avram, 1975). To encode these records, Henriette Avram and her colleagues at the Library of Congress introduced several ideas that were innovations then but are now commonplace. From the beginning, MARC supported the characters found on a typical catalog card, including upper and lower case letters, and the diacritics used by European languages, soon followed by Greek and other non-Latin scripts. Mainstream computing was at least fifteen years behind. At a time when data input was on punched cards and fixed formats were almost universal, MARC allowed fields to be of variable length and repeated as appropriate. MARC's cryptic system for tagging fields looks strange today, but it was state of the art when memory cost a dollar per byte.

Z39.50 was one of the first application protocols for distributed computing. It began in the early 1980s as the Linked Systems Project for linking large bibliographic utilities (Avram, 1986). It was originally designed within the OSI model, but was adapted to run over TCP/IP. Z39.50 was used extensively by libraries and commercial vendors for distributed searching, but was never adopted by the wider computing community. One reason is that it is a cumbersome protocol to implement, with numerous options and a complex mechanism for maintaining state between two computers.

Technology such as MARC and Z39.50 enabled libraries to develop valuable computing services, but created a problem of legacy software. A whole generation of professional staff was trained in these legacy systems and several poorly capitalized companies had their product lines based on them. In 2011, when the Library of Congress announced a strategic move to replace the MARC format (Library of Congress, 2011), it noted that MARC is "based on forty-year-old techniques for data management and is out of step with programming styles of today." On technical grounds MARC would have been phased out many years ago, but its universal adoption by libraries around the world made it hard to replace.

2. Architecture

The architecture of distributed information systems was a theme of computing research throughout the 1990s. For libraries, a particularly important question was interoperability. When collections are managed by many independent providers, how can they appear to a user as a single digital library?

2.1 Early digital libraries

The Mercury Electronic Library at Carnegie Mellon University was the first digital library at an American university (Mercury, 1992). When it was released to the campus in 1991 it had a dozen textual databases and a collection of page images of journal articles in computer science. The articles were licensed from ACM, IEEE, and Elsevier. An important part of the project was the conversion, storage, and delivery of page images over the campus network.

Mercury's architecture was typical of distributed computing before the arrival of the web. It combined the Z39.50 protocol for distributed searching with services provided by the university network such as authentication. The fundamental paradigm was to search a text database to find information, followed by online delivery of a document or an abstract. Z39.50 was used to send queries to the servers on which the information was stored. For interoperability, Mercury used a reference server, which kept metadata about the online collections, the fields that could be searched, the indexes, and access restrictions. Since the library was attached to the Internet and most materials were licensed from publishers, security was a central part of the architecture.

2.2 Computer science research

For a period beginning in the 1990s, digital libraries were a recognized topic of computer science research. All parties gained from this partnership. Modern information systems rely heavily on computer science, while computer science benefits from the focus on usability introduced by information scientists. The management of vast amounts of information has become one of the driving forces behind data-driven supercomputing.

The first project supported by the US government was the Computer Science Technical Reports project funded by the Defense Advanced Project Research Agency (DARPA) in 1992. (Why DARPA should support research into libraries is a question beyond the scope of this article.) This was unashamedly a computer science research project. It was coordinated by the Corporation for National Research Initiatives (CNRI), whose leaders, Robert Kahn and Vinton Cerf, had been central to the development of the Internet. Other members of the project came from five universities: Berkeley, Carnegie Mellon, Cornell, MIT, and Stanford.

Kahn and Cerf had written an architectural proposal, "The World of Knowbots", which was widely circulated, despite efforts to keep it confidential (Kahn and Cerf, 1988). They were advocates of a detailed architecture that would be used by every library, but the university members knew the difficulties of gaining consensus within even a single campus and argued for a looser framework. These architectural discussions clarified several important concepts (Arms, 1995). CNRI created the handle system for managing persistent names, which is the technology behind the Digital Object Identifier (DOI) used by publishers. Jim Davis and Carl Lagoze from Cornell implemented NCSTRL, a distributed library for technical reports (Davis and Lagoze, 2000). Today's Fedora system is a direct descendent of their work. As a subproject, Michael Mauldin at Carnegie Mellon developed the Lycos web search engine, whose impact is discussed below (Mauldin, 1997), and the Stanford CSTR group was later the incubator for Google.

In 1994, the National Science Foundation (NSF), DARPA, and NASA joined together in the first digital libraries initiative (Griffin, 1998). The NSF funds came from the computer science division, but the division was remarkably open in supporting information science and library research. The NSF's imprimatur also established the name for this field as the plural "digital libraries", replacing earlier names such as "electronic library" and "virtual library".

2.3 The early web

In the early 1990s, several groups created computer software that allowed authors to bypass publishers and mount their own information online. Three are particularly noteworthy: Gopher, WAIS (which used a stateless variant of Z39.50), and the World Wide Web. The first two are footnotes in history, but there was nothing inevitable in the success of the web.

It is doubtful whether the web would ever have become popular without the Mosaic browser developed by Marc Andreessen, a student at the University of Illinois, and released in 1993. Mosaic brought color to the Internet. It brought formatted text and embedded images. These alone made it an instant success, but even more importantly it was a portable interface. Before the web every networked application required a separate user interface for each type of computer. These interfaces had to be modified for every new version of the operating system. For example, Mercury developed a pipelined algorithm to retrieve page images, transmit them across the network, and display them very quickly, but it was designed for Unix computers on the Carnegie Mellon network and was a burden to port to other systems. The first release of Mosaic was for Unix but it was soon followed by versions for Microsoft and Apple operating systems. Its commercial successor Netscape provided a browser for every computer system, so that application developers could write a single version of the user interface and reply on the browser to deliver it to the users.

For applications such as digital libraries, the early web had many flaws. It had some valuable features – a simple version of HTML, URLs to locate information, and MIME types to specify formats – but its underlying protocol, HTTP, was very limited. Mosaic had support for a few older protocols, e.g., Gopher, and FTP, but in practice there was little choice but to use the primitive versions of HTTP and HTML. There was no concept of state, no security, and no form of identifier except the location field in a URL. The set of MIME types was fixed and there was no way to add more types or to download code to a browser. In due course most of these problems were solved by additions to the core technology of the web. State was added through cookies. HTTPS emerged as a secure form of communication. Plug-ins provided a flexible way to add new types of data. Netscape introduced JavaScript and Sun developed Java applets to download code to a browser. The incremental way in which these features were added explains the messy architecture of the web that we live with today, but the overall result is that modern browsers and web servers provide a rich set of facilities for networked applications, including libraries.

In the 1990s these developments of the web were far from obvious. Most computer scientists and librarians considered that the web was a toy, useful in the short term, but destined to be replaced by something much better.

2.4 Transitional digital libraries

Early web servers provided a connection to external systems via the common gateway interface (CGI). This allowed developers to use a web browser to display information and to interact with the user, but to write their own code for everything else. An important library from this period was American Memory from the Library of Congress (Arms, 1996).

Earlier experiments at digitization by the Library of Congress had data stored on video disks and a user interface based on Apple's HyperCard. HyperCard was also used for the first version of the Perseus Digital Library of classical texts, developed by Greg Crane at Harvard (Crane, 1998). In 1995 the Librarian of Congress established a project to digitize five million items and make them available on the web within five years. Some creative counting was needed to meet this target, which included large numbers of items already digitized by the University of Michigan and Cornell, and collections from a grant program funded by Ameritech Corporation. Most of the items were digitized photographs, maps, and other non-text materials, but some books were converted to SGML, using a DTD based on the Text Encoding Initiative.

The architecture of American Memory had three major components: a web browser, a data store, and a server to search and browse the collections. Nowadays it would be called a three-layer architecture. Access to the collections was entirely through the indexes. In a departure from previous practice, American Memory combined access through metadata with full text indexing where possible. Another important departure was explicit recognition that the metadata varied greatly for different categories of material. MARC records were used when available, but many collections relied on finding aids, which were encoded using the Encoded Archival Description.

With the technology that is available today a replica of this system could be built with off the shelf components, but in the 1990s the technology was not available. Numerous tricks had to be used so that early web browsers could display the materials. They included a specially designed page-turner and a TIFF viewer provided as a plug-in. In an offline batch process, the SGML texts were rendered into HTML, which was stored in the data store. Separate thumbnails were kept for all images.

As an interesting footnote to history, American Memory was one of the first web sites to recognize the importance of graphic design, and employed professional graphics artists. Previous web sites had been designed by computer specialists, few of whom had any graphical expertise.

2.5 Modern digital libraries

By the end of the 1990s, digital libraries were no longer a novelty. The DSpace software for managing institutional repositories is an excellent example of a library system from this period (Smith *et al.* 2003). When the definitive history of modern libraries is written, institutional repositories may be recognized as a visionary achievement or relegated to a footnote as an experiment that failed, but there is no doubt about the quality of the DSpace software. The architecture is firmly within the framework of the modern web, but

extends it with concepts developed by the library community. From the first release, DSpace used Dublin Core metadata, XML for data exchange, the Open Archives Initiative protocol for metadata harvesting, and the CNRI handle server for persistent naming. Within its three-layer architecture, the system was designed around the workflows of a large university and was explicit in its support for a very wide range of formats.

DSpace was developed by MIT Libraries under the leadership of MacKenzie Smith with support from Hewlett Packard. It was first released in 2002. Eventually it too will become a legacy system. Until then it is a fine example of how web technology can be extended to meet the needs of a specialized community.

3. Searching

A theme of the past thirty years is the steady substitution of automatic processing for tasks that used to be carried out by experts. Fields from chess playing to machine translation have discovered that simple algorithms plus immense computing power often outperform human intelligence.

3.1 Automated indexing

Information retrieval is a prime example. In 1990, somebody who wanted information about computing might use the INSPEC indexing service. Ten years later, the same person typically began by searching Google. By all traditional criteria this was wrong. The publications indexed by INSPEC had been selected by experts; the individual articles had been reviewed and edited by publishers; and the index records were created by trained indexers. Yet all this expertise was frequently not as useful as the fully automated approach taken by Google (Arms, 2000).

The first step in the transition to modern information retrieval was to replace descriptive metadata with indexes built from the actual text of documents, and to replace boolean searching with ranking based on statistical measures. As early as 1967, Gerald Salton had demonstrated that full text indexing of documents could be more effective than boolean searching of metadata (Cleverdon, 1967), but his work had little immediate impact in libraries. One reason was that full text indexing strained the capacity of existing computers, but perhaps more important was skepticism. How could the application of simple statistics to the text of a document be more effective than metadata created by experts? Repeated experiments such as the TREC conferences in the 1980s confirmed Salton's approach, but there was strong resistance against moving to the new methods.

The web changed all of this. Indexing services in fields such as medicine and law had assumed that the user was a trained professional or a reference librarian. Their search languages used complex boolean combinations of index terms. In skilled hands, these languages were very powerful, but they were totally unsuited for the casual user. From the beginning web search engines such as Infoseek, Lycos, and AltaVista were designed

for untrained users and supported queries that were simple lists of search terms. These early web search services followed Salton in ranking search results by applying simple statistics to the words in a document.

Experts trained in conventional information retrieval were quick to criticize this approach. The early engines did quite well in indexing a web of 25 million pages, but struggled with duplicates and had problems in ranking widely different pages. It is hard to remember now that the impetus for the first Dublin Core workshop was a belief that automatic indexing would be inadequate for huge numbers of pages. Richer records, created by content experts, were considered necessary to improve search and retrieval (Weibel, 1995).

The experts failed to predict that new developments in the technology of web searching would leave human indexing far behind. One of the breakthroughs was a paper by Sergey Brin and Larry Page, the founders of Google, while they were still students at Stanford (Brin and Page, 1998). Their paper was reportedly rejected by the journal to which they first submitted it – presumably because of its naïveté about conventional information retrieval –, but it contained some remarkable new ideas.

One new idea was to redefine the objective of web searching. Previously, information retrieval had assumed that all documents in a collection were of equal value, for example peer reviewed papers in aeronautics journals. For a given request, some documents were relevant and others were not, but all relevant documents were important. Success in boolean searching was measured by how many of the documents retrieved were relevant (precision), and how many of the relevant documents were retrieved (recall). Full text searching replaced the concept of a set of retrieved documents by a ranking of all the documents, but the underlying metrics were still precision and recall. The Google paper pointed out that on the web documents vary greatly in importance. Most searchers are not looking for the set of all relevant documents, they are looking for a few items that best satisfy their needs. (The paper was quite vague in defining "best".)

A second major contribution of this paper was to show how the relationship between a web page and its neighbors could be used to predict its usefulness. Other search engines had begun this development, for example by analyzing the structure of URLs, but the authors made it a central part of their approach. The paper is particularly well known for introducing the PageRank algorithm which ranks a page based on the pages that link to it.

In recent years, web search engine such as Google or Bing have made extensive use of concepts from artificial intelligence. Natural language processing is used for spelling and other language analysis, and machine learning is used to adjust the numerous parameters that control the algorithms. The search services use their huge numbers of users to experiment with the search algorithms. A simple experiment is to vary one of the parameters and to use the new version for a fraction of the searches. Observing which hits the users selects provides information about whether the new version is an improvement on the old. Machine learning is then used to optimize the combination of features that are used by the search engine.

3.2 Metadata

Because of the successes of automated indexing and natural language processing, descriptive metadata is less central to libraries than it used to be, though it remains important for non-textual materials and in specialized domains such as law or medicine. Overall, it is probably fair to say that during the 1990s library researchers put too much energy in trying to create general-purpose metadata standards. Metadata is expensive and it is easy to underestimate the difficulties in getting a format adopted. Of the attempts to develop standards for specific types of digital material, only a few have established themselves, such as SCORM for courseware, ONIX used by publishers, and the various formats for geospatial information. One line of research, which offered promise but eventually led nowhere, was to extract metadata automatically from web pages.

There has been one outstanding success, the Dublin Core metadata initiative, led by Stuart Weibel of OCLC (Weibel, 1995). As already mentioned, Dublin Core was originally seen as a low cost way to create metadata for indexing web pages. While web indexing has succeeded without manually created metadata, Dublin Core has satisfied a broad wish for a lightweight metadata format that fits with the technology of the web. Just as MARC was technically advanced in its day, much of the success of Dublin Core is the thoroughness with which it has integrated modern technology for automatic processing. Today it is used in areas that have nothing to do with libraries by companies such as Microsoft and Abode who use it as an embedded format for their software products.

From its inception, Dublin Core aimed to reduce the cost of creating metadata by minimizing the role of professional cataloguers. The report of the first workshop in 1995 stated, "Since the Internet contains more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves." This was one of the first specific suggestions that non-experts could create useful metadata.

Metadata harvesting is a good example of the complementary strengths of computer science and information science. The idea that repositories should expose metadata for others to harvest was introduced by the Harvest project in the early 1990s (Bowman *et al*., 1994). This was a computer science research project. It developed some basic software, but did not meet the practical needs of libraries. Several years later, Herbert Van de Sompel and Carl Lagoze revived the concept and developed the successful Open Archives Initiative protocol for metadata harvesting, using concepts from Harvest (Van de Sompel and Lagoze, 2000).

4. Content

Libraries are nothing without content. Twenty years ago nobody anticipated the pace at which library information would become available in digital formats. The earlier retrospective conversion of card catalogs to MARC had taken decades and there was nothing to suggest that the conversion of library collections would be different.

4.1 Financial models

The biggest surprise between today's libraries and the predictions made twenty years ago is the immense amount of material that is now available with open access. The financial model for much of the content – often of very high quality -- is that it is placed online with no expectation of payment. This was completely unexpected. Early digital libraries assumed that high quality content is inevitably expensive. Therefore libraries must pay for content, and digital libraries have to prevent people from accessing materials that they have not paid for. The fact that the early web had no barriers to access was one of the reasons that it was dismissed as a toy. In retrospect it is obvious that many organizations and individuals benefit from having their materials widely read and will use their own resources to make it happen, but most things are obvious in retrospect.

For a while, micropayments were seen as a promising way to charge for access. Every item of information would have a small charge associated with it. Recording huge numbers of tiny transactions led to interesting research in security and transaction processing, but the idea was never widely adopted. It is now almost forgotten (Sirbu and Tygar, 1995). Instead, subscriptions emerged as the standard way that libraries pay for collections of online materials, particularly journals.

An early business strategy was colloquially known as "the addiction model": give the service away free until the users were hooked and then charge them for access. As a daring alternative, Lycos decided to build its business on advertising. When web search engines such as Infoseek tried to develop a base of paid subscribers they could not compete with the free service provided by Lycos and also turned to advertising. This became so profitable that it is hard to remember that there was ever an alternative.

4.2 Selection

The economies of large scale computing have reduced the need for selection in publishing and in library collections. With printed materials, libraries could not afford to collect everything because of the costs of acquisition, cataloguing, and storage. Therefore, specialists selected the items to be acquired for library collections. However, selection is labor intensive and hence adds costs. It also adds delays and sometimes valuable material is rejected. With digital libraries, users may be better served by eliminating selection and collecting everything. Here are two important examples of open access services that operate with minimal selection.

In 1996, Brewster Kahle began to archive periodic copies of the web (Kahle, 1997). From the earliest days, when he copied web pages on to streamer tapes in his San Francisco house, the Internet Archive has been unselective. The web is too big for the

archive to collect everything, but the decisions about what to collect are made on broad criteria, such as formats, structure of web sites, and requests from copyright owners. At no time has any attempt been made to select pages based on criteria of quality or their archival value. Kahle came from a supercomputing background. He had the insight to realize that the equipment costs of comprehensive collecting were manageable and he had the technical skill to develop ways to do so. For many years his approach received widespread skepticism. Today, few would deny that the Internet Archive is the definitive archive of the web.

The e-print archives now known as arXiv.org also have no formal selection or review. The archives reserve the right to reject papers that are clearly out of scope, but make no judgment on their scientific quality. The first archive was established in 1991 by Paul Ginsparg at Los Alamos National Laboratory to serve the needs of high-energy physicists (Ginsparg, 1997). There are now archives for other branches of physics, mathematics, and related disciplines such as computer science. High-energy physics laboratories had previously circulated preprints of their papers to each another by email, but arXiv.org goes much further. In essence anybody can post an unreviewed paper claiming that it is original research. Timeliness is central to the success of the archives. Because there is no selection and no editing, papers are made available to researchers around the world within a day of being submitted.

Open access costs money and the business implications are still far from resolved. The Internet Archive was originally funded by a private foundation set up by Kahle and his wife. In the early years, funds for arXiv.org came from the National Science Foundation and the Department of Energy. Now it is managed by Cornell University Library. Cornell is a rich university, but still faces the question whether its operating budget should be supporting scientists around the world.

4.3 Experiments in publishing

The 1990s saw numerous experiments with new forms of academic publishing. Some were motivated by the belief that when the taxpayer supports research the public should have open access to the results. Others were conscious efforts to lower the cost of scholarly publishing, with its time consuming editing and peer review, transfer of copyright from authors, and high prices for library subscriptions.

First Monday was an early example of an open access journal (Valauskas *et al.*, 1996). It uses a traditional model of peer review, but is entirely online and open access. Since its first publication in 1996, it has maintained a high academic standard. *D-Lib Magazine* uses a different approach that emphasizes timeliness (Friedlander and Arms, 1996). Originally, manuscripts could be submitted as little as three days before publication, which left time for the editor to make suggestions, the authors to approve them, and final formatting. Because of this fast publication, many important events in digital libraries were first published in *D-Lib Magazine*. As with arXiv.org, neither *First Monday* nor *D-Lib Magazine* expects the author to transfer copyright.

There have been many other attempts to provide open access to scientific research. Recently there has been some progress, particularly in bio-medical literature where the UK *BioMed Central* and the *Public Library of Science* lead the way. Overall, however, the attempts have not been particularly successful. Publishers have been very creative in protecting their revenues, but the real problem is the conformist university community that judges faculty almost entirely by their publications in traditional journals. In some fields, but far from all, the problems of open access and timeliness have been solved by authors simply posting their papers on personal or departmental web sites, at the same time as they submit them for peer review. This leads to the strange situation where fellow researchers read each other's work on their private web sites, yet they cite later versions of the papers that are published by a conventional journal.

4.4 Digitization

Printed documents can be converted to digital formats either by transcription or by scanning. The seminal project in scanning document collections was CORE, which ran from 1991 to 1995 (Lesk, 1992). This was a research project by Bellcore, Cornell University, OCLC, and the American Chemical Society. CORE converted about 400,000 pages from twenty chemistry journals. It introduced several ideas that have since become popular in conversion projects, including two versions of every article, a scanned image and a text version marked up in SGML. The SGML text was used to build a full-text index for information retrieval and for rapid display on computer screens. CORE was one of the first studies to emphasize the importance of browsing in addition to searching.

Since big digitization projects are beyond the financial resources of universities, many of the early projects were by publishers to convert their back runs of journals. Elsevier's Tulip Project, which began in 1991 with nine university libraries, was an important pilot that demonstrated the benefits to academic users of having the scanned text of journal articles available online (Gusack and Lynch, 1995). Tulip's special contribution to the development of modern libraries was to find ways to track usage patterns while ensuring that the individual users remain anonymous.

Two not-for-profit organizations have made special contributions in this area. The first is the Andrew Mellon Foundation, which has been a generous supporter of libraries for many years. Realizing that many society publishers did not have the resources to convert their collections, the foundation created JSTOR and helped it develop a business model that allowed it to become self sustaining (Bowen, 1999). The other is the University of Michigan. The university has a long history of digitization, beginning with the Making of America project in 1995 in partnership with Cornell. The university has been particularly attentive to affordable ways to organize and manage large collections of digitized documents. By a happy coincidence, Larry Page, one of the founders of Google, is a Michigan alumnus. Through his involvement, Michigan has played a central role in Google's work to digitize huge numbers of books from major research libraries, and in the creation of the Hathi Trust to manage these materials on behalf of libraries.

4.5 Transcription

Two very different fields were pioneers in transcribing comprehensive collections of documents: law and classics. Each has a small core corpus that is not protected by copyright. Law was the first discipline to have an extensive collection of primary materials online. Jerome Rubin and his colleagues launched Lexis as a commercial service in the early 1970s, with access to the full text of statutes and case law (Rubin, 1973). In classics, the Thesaurus Linguae Greacae has transcribed almost every Greek text from the classical period to the fall of Byzantium.

Some time in the mid-1980s Michael Hart formalized his long-standing interest in transcribing texts into the Gutenberg Project with the goal of having volunteers transcribe ten thousand books and making them available to everybody. Hart was widely viewed as an eccentric and his goal as quixotic, but he persevered and the Gutenberg corpus is now one of the most important collections of digitized books.

5. Change

The 1990s were a decade of disruptive change, but rapid change has continued since then. CrossRef, the Open URL, and Fedora date from about 2000 and the major developments of institutional repositories were after the period discussed in this article. Wikipedia began in 2001.

At times of change, successful organizations often have difficulty in accepting that assumptions that were central to their past achievements may be barriers to success in a new environment. Libraries have been no exception and almost every development in digital libraries has been met with resistance. The reluctance to change is aggravated by the fact that the early versions of new ideas often do not reach the standards of the established alternatives. It is easy to focus on the short term problems and ignore the long term potential. Scholars examined the first books scanned by Google and rejected the entire program because of technical problems that were being worked on at that very moment. Cynics delighted in entering false information into Wikipedia and writing articles to show that the entire concept was flawed.

Yet, despite this inertia, libraries have succeeded in embracing much of the potential of online information, often in ways that were not predicted. Some experiments failed, often because of external social and economic forces, but others have survived and become mainstream. The 1990s were a period of affluence for America universities and libraries were able to experiment and make changes. Universities are now under severe financial pressures, but thanks to the pioneering work done in the 1990s their libraries appear well placed to face a period of uncertainty and restraint.

Notes on the References

Wherever possible, the references given below are to the earliest paper that describes each project. There are no references to several of the projects, which were never

formally described in the academic literature or the articles about them were written many years later. Many of these projects are described in Wikipedia or on modern web sites. Some of these descriptions are excellent, such as the Wikipedia article on Mosaic (July 2012). For other projects, and particularly those that became commercial products, the histories that are now online are frequently self-serving or distorted.

The URLs in the references were checked on July 20, 2012.

References

Arms, C. (1996), "Historical collections for the National Digital Library", *D-Lib Magazine",* Vol. 2, No. 4. http://dlib.org/dlib/april96/loc/04c-arms.html.

Arms, W. (1995), "Key concepts in the architecture of the digital library," *D-Lib Magazine*, Vol. 1, No. 1. http://www.dlib.org/dlib/July95/07arms.html.

Arms, W. (2000), "Automated digital libraries. How effectively can computers be used for the skilled tasks of professional librarianship?", *D-Lib Magazine*, Vol. 6, No. 7/8. http://dlib.org/dlib/july00/arms/07arms.html.

Avram, H. (1986), "The Linked Systems Project: its implications for resource sharing", *Library Resources and Technical Services*, Vol. 30, No. 1, pp. 36-46.

Avram, H. (1975), *MARC; its history and implications*, U.S. Government Printing Office, Washington D.C.

Bowen, W. (1999), "JSTOR and the economics of scholarly communication", *Journal of Library Administration*, Vol. 26, No. 1-2, pp. 27-44. DOI:10.1300/J111v26n01_05.

Bowman, M., Danzig, P., Manber, U., and Schwartz, M. (1994), "Scalable Internet Resource Discovery: Research Problems and Approaches," *Communications of the ACM*, Vol. 37, No. 8, pp. 98-107.

Brin, S., Page, L. (1998), "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, Vol. 30, pp. 107- 117.

Cleverdon, C. (1967), "The Cranfield tests on index language devices", *ASLIB Proceedings*, Vol. 19, No. 6, pp. 173-194.

Crane, G. (1998), "The Perseus Project and Beyond", *D-Lib Magazine*, Vol. 4, No. 1. http://www.dlib.org/dlib/january98/01crane.html.

Davis, J. and Lagoze, C. (2000), "NCSTRL: design and development of a globally distributed digital library", *Journal of the American Society for Information Science*, Vol. 51, No. 3, pp. 273-280.

Friedlander, A., and Arms, W. (1996), "D-Lib", *Ariadne*, No. 5. http://www.ariadne.ac.uk/issue5/dlib/.

Ginsparg, P. (1997), "Electronic research archives for physics", in *The Impact of Electronic Publishing on the Academic Community*, Portland Press, London.

Griffin, S. (1998), "NSF/DARPA/NASA Digital Libraries Initiative: a program manager's perspective", *D-Lib Magazine*, Vol. 4, No. 4. http://dlib.org/dlib/july98/07griffin.html

Gusack, N., and Lynch, C. (1995), "The TULIP project", *Library Hi Tech*, Vol. 13, No. 4, pp.7–24.

Kahle, B. (1997), "Archiving the Internet", *Scientific American*, March.

Kahn and Cerf (1988), *The world of Knowbots*, Corporation for National Research Initiatives, Reston VA. http://www.cnri.reston.va.us/kahn-cerf-88.pdf.

Lesk, M., *et al.* (1992), "Better things for better chemistry through multi-media", *Proceedings of the Eighth Annual Conference of UW Centre for the New OED and Text Research*, Waterloo, Ontario. http://www.lesk.com/mlesk/waterloo92/w92.html.

Library of Congress (2011), *A bibliographic framework for the digital age*, news announcement, October 31. http://www.loc.gov/marc/transition/news/framework-103111.html#ftn1.

Mauldin, M. (1997), "Lycos: Design choices in an Internet search service", *IEEE Expert*, Jan-Feb, pp. 8-11.

The Mercury Team (1992), *The Mercury Electronic Library and Library Information System II: the first three years*, Mercury Technical Reports Series, No. 6, Carnegie Mellon University, Pittsburgh. http://www.cs.cornell.edu/wya/papers/Mercury6.doc.

Rubin, J. (1973), "LEXIS: An automated research system," in May, R. A., *Automated law research*, American Bar Association, pp. 35-42.

Sirbu, M. and Tygar, D. (1995), "NetBill: an Internet commerce system optimized for network delivered services", *Proceedings of the IEEE CompCon*.

Smith, M., *et al.* (2003), "DSpace an open source dynamic digital repository", *D-Lib Magazine*, Vol. 9, No. 1. http://dlib.org/dlib/january03/smith/01smith.html.

Valauskas, E., Dyson, E. and Ghosh R. (1996), "Editors' introduction", *First Monday*, Vol. 1, No. 1. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/464/385.

Van de Sompel, H. and Lagoze, C. (2000), "The Santa Fe convention of the Open Archives Initiative", *D-Lib Magazine*, Vol. 6, No. 2. http://dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.

Weibel, S. (1995), "Metadata: the foundations of resource description", *D-Lib Magazine*, Vol. 1, No. 1. http://www.dlib.org/dlib/July95/07contents.html.