

2 Outcomes of digital libraries' first decade

Karen Calhoun

Cornell University Library (retired)

ksc10@cornell.edu

Note: This is a preprint of a chapter whose final and definitive form was co-published in *Exploring Digital Libraries: Foundations, Practice, Prospects* by [Facet Publishing](#) (2014) and [ALA Neal-Schuman](#) (2014).

Overview

This chapter identifies and discusses a set of significant outcomes from the first decade of digital library research and practice (1991 to 2001). It describes accomplishments that set the dominant themes and continue to shape the field of digital libraries today. The chapter's overall purpose is to offer a framework for understanding the productive work of thousands of people during that period, one that reveals the interplay of people (producers and providers of digital libraries); enabling technologies; and the collections, services and communities they support. Figure 2.1 visualizes the framework and seven elements within it. The chapter discusses the elements in the following order:

1. A new field of research and practice
2. The transformation of scholarly communication processes
3. Open access
4. Technological innovations
5. Digitization and digital preservation
6. Metadata and standards
7. Working digital libraries and the communities they serve

These are the elements that formed the foundations as digital libraries moved into their second decade.

Keywords: Digital libraries—Evaluation; Scholarly communication; Open access; Metadata; Digitization; Interoperability; Repositories; Cultural heritage collections

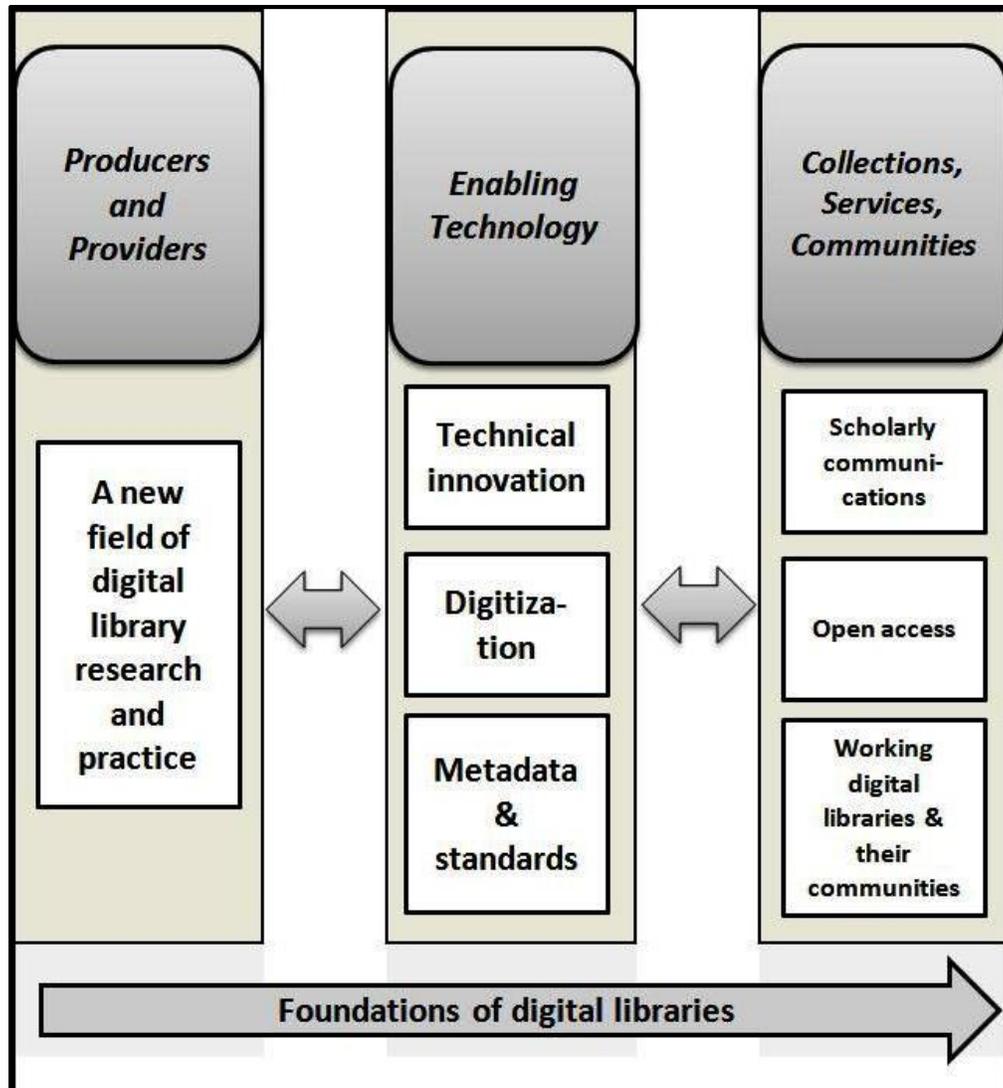


Figure 2.1 Key outcomes of the first decade of digital libraries

A new field of research and practice

The disciplines of digital libraries

As noted by Lynch, the first decade of research, development and practice in digital libraries was characterized by “an enormous, exhilarating flowering of innovation, creativity and experimentation” (2000). From 1991 and into the new millennium, large numbers of projects were generously funded internationally and nationally by government agencies and foundations, institutions, public- and private-sector organizations and individuals around the world. At local

levels, universities invested considerable funding in digital library research, prototyping and operations. The flowering was plentiful but diffuse: Lynch begins a later article with the remark that “the field of digital libraries has always been poorly-defined, a ‘discipline’ of amorphous borders and crossroads” (2005). In a preprint of a conference paper, Nguyen (2011) offers a long view based on a systematic study of twenty years’ development of the peer-reviewed literature. Nguyen’s results from an analysis of Scopus suggest that peer-reviewed papers have come from computer science (63%), library and information science (26%) and many other fields (11%).

Community building and organizational support

Early digital library community building, which took place through conferences, foundations, associations, cooperatives, partnerships and projects, brought far-flung digital library developers and practitioners together and contributed substantially to their efforts. These early community building efforts produced an active field of digital library research and practice as well as working digital libraries.

With respect to conferences, the computer science section of IEEE and the Association for Computing Machinery (ACM) began hosting conferences in 1995 and 1996 respectively. National and international library associations as well as many other associations and organizations now host digital library conferences; interested individuals could attend one or more conferences each month, if they so desired (*D-Lib Magazine* (dlib.org/groups.html) maintains a list of digital library conferences).

The foundations, associations, membership organizations and others that have been major supporters of digital library development are too numerous to describe in this short section, but without their contributions, digital libraries would not have emerged.

Education for digital librarianship

A variety of training programs as well as formal courses in digital libraries had begun to appear by the end of the first decade, and more developed over the ensuing years. Ma, Clegg and O'Brien (2009) provide an overview of trends and the results of their study of education for digital libraries from 1999 to 2006, as digital libraries were emerging. Ma's results echo the earlier findings of Spink and Cool (1999) and Liu (2004). One well-known cooperative project to develop a digital library curriculum combined experts from both LIS and CS (Yang et. al 2009). Tamaro (2007) reported on work being done in Europe to develop digital library education; earlier, Liu (2004) had reported on programs being offered in the UK, the Netherlands and elsewhere. Sheila Corral (2011, 57-60) offers a more recent evaluation of progress and the continuing debate around educating library professionals for the specific requirements of digital library environments.

The literature of digital libraries

The founding of D-Lib Magazine and Ariadne

Bill Arms and some colleagues founded *D-Lib Magazine* (dlib.org) in 1995. It has proved to be a key resource tracking the progress of digital libraries and the interdisciplinary field that grew up around them (see also chapter 3). Much of what was happening in the NSF-funded projects was reported in *D-Lib*. *Ariadne* (ariadne.ac.uk) grew out of the eLib program in the UK (Dempsey 2006a) and has served a similar function (Tedd 2002) for UK projects, particularly those funded by JISC, a very important agency supporting UK higher education and libraries, computing and research. Published by UKOLN, *Ariadne's* first issue is dated January 1996.

Blogs and e-discussion lists

Since its beginnings the digital library community has embraced the web and its new forms of communication and participation. Roy Tennant has been blogging about digital libraries since 1997 (Tennant 2004b, vii) and his blog *The Digital Shift* (thedigitalshift.com) has been widely influential. Since 1990 the current awareness newsletter *Current Cites* has been a good source for digital library topics. Charles Bailey Jr. (digital-scholarship.org/cwb/) has created and maintained online bibliographies since 1996; they have been a valuable source for learning about and tracking selected digital library topics. The Dublin Core Metadata Initiative has maintained public online news since 1995 and a public e-mail list since 1996.

Publication patterns over time

An informal quantitative analysis of publications on digital library topics suggests that articles began to appear in the early 1990s and grew to a peak in 2005 and 2006. This publication pattern is illustrated by figure 2.2, which is a snapshot of the count of items indexed by Google Scholar with either “digital library” or “digital libraries” in the title for each year from 1990 to 2012. The counts were captured in June 2013. The reader should consider that there may be a time lag before newer papers are indexed in Google Scholar; this time lag contributes to the lower number of articles found for 2011 and 2012.

For comparison the same search was done using Scopus (an Elsevier-owned subscription database of citations and abstracts from primarily peer-reviewed academic journals and conference proceedings). Scopus is a competitor to another commercially-available product for tracking scholarly citations, the Web of Science (WoS) from Thomson-Reuters. The pattern of the Scopus curve is the same as the Google Scholar curve, with articles growing to a peak in 2005 and 2006, but remaining fairly steady through 2010. This analysis was inspired by Christine Borgman’s keynote address at a Joint Conference on Digital Libraries (2009), in which

she briefly noted the clustering pattern of the usage of “digital library” in Google Scholar indexes.

The results visualized in figure 2.2 are not unexpected: they are consistent with studies comparing Google Scholar, Web of Science and Scopus from Meho and Yang (2007), De Sutter and Van Den Oord (2012), and Harzing (2012 and 2013). Google Scholar’s coverage of many document types (e.g., dissertations and theses, reports, conference presentations, working papers and posters, preprints, and more) is the reason for the higher number of documents indexed in Google Scholar compared to Scopus, which indexes articles from primarily peer-reviewed sources.

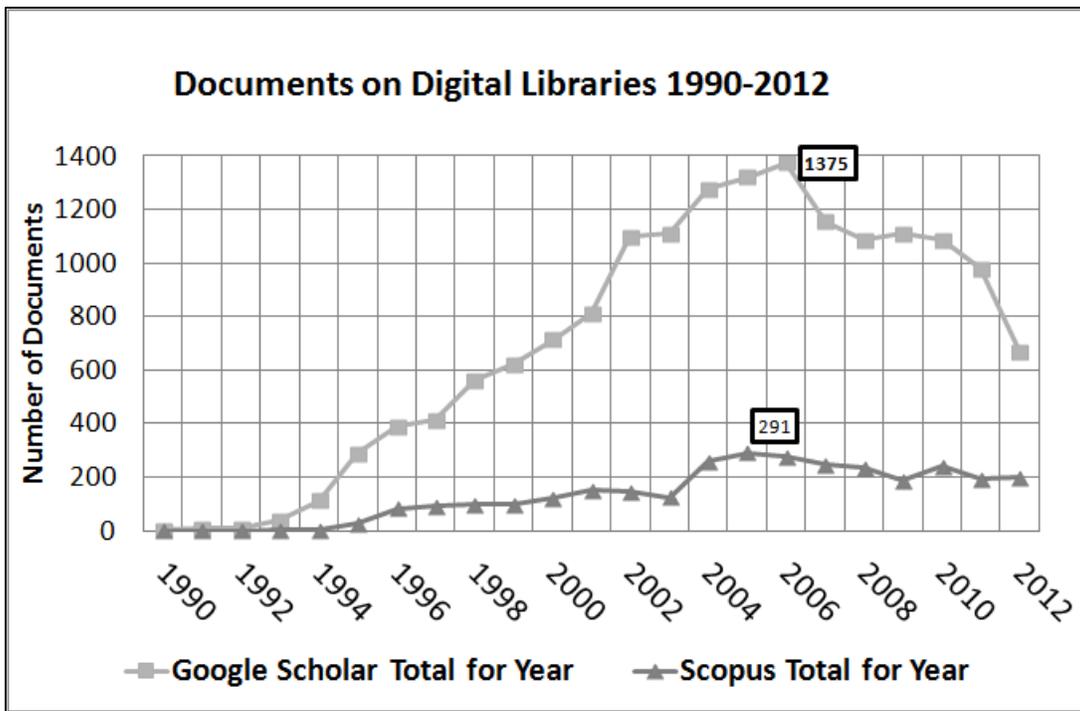


Figure 2.2 Digital library documents indexed in Google Scholar and Scopus, 1990-2012*

*Articles with the words “digital library” or “digital libraries” in the title

The results of this quantitative analysis are also consistent with the history of national-level research and development funding for the digital library field, as described in chapter 1. Ambitious projects produced a growing number of publications during the years following 1994—when large-scale funding began—until about 2006, when the many research findings produced under the largest grants had appeared in the literature.

The impact of shifts in funding

After 2005, large-scale funding for digital library research from US federal agencies diminished. In a 2005 article Stephan Griffin, then a program director at NSF, noted plans for a third US digital libraries program (beyond DLI-1 and DLI-2), for example as documented following the 2003 NSF-sponsored Chatham Workshop (Larsen, Wactlar, and Friedlander 2003). This third program did not materialize as expected. Similarly Lynch, reviewing a decade's work in digital libraries, noted

As of 2005, it seems a virtual certainty that substantial programmatic US government funding of digital libraries research in terms of the construction of prototype systems is at an end, at least for the near future. The novelty of constructing digital libraries as a research end in itself has run its course... (2005)

After 2003 other NSF funding priorities came to the fore such as cyberinfrastructure, e-science and the stewardship of digital data (Atkins et al. 2003). Chapter 8 returns to these topics.

The literature of digital library practice

Around the start of the new millennium, the digital library field of endeavor began to include many more publications reporting the results of practitioners. By 2007 a large number of university research libraries had introduced digital library programs. For example, in a survey conducted on behalf of the Association of Research Libraries (ARL) in January 2006, results

indicated that over half of ARL libraries had or planned to have working institutional repositories of locally produced digital works (Bailey et al. 2006). In February 2006, a little over half of the members of the ARL libraries responded to a survey about digitization activities; nearly all of those who responded (97%) reported they were engaged in these operations in their libraries (Mugridge 2006).

The transformation of scholarly communication processes

Early projects

Robert Wilensky, principal investigator of a DLI-2 project that began in 1999 wrote “our practice of disseminating, accessing and using information, especially scholarly information, is still largely informed by the nature of pre-electronic media” (2002). He, like many others working in the field of digital libraries at that time, advocated the development of new enabling technologies and new publishing models that would transform and substantially improve scholarly information dissemination and use. Hans Geleijnse of Tilburg University, a leader and early adopter of digital library technologies in the Netherlands, provides an excellent description of how Tilburg began innovating its scholarly information services in the early 1990s (Geleijnse 1999).

The first decade of digital libraries research and practice made significant progress toward this set of goals. Chapter 1 discussed Mercury and CORE, two early influential projects. Others include:

TULIP (The University Licensing Program), a project organized by Elsevier in 1991. TULIP tested the (pre-web) networked, desktop delivery of e-journals with nine universities. A parallel experiment with Tilburg University ran from 1992 to 1995 (Elsevier 2012). The work evolved eventually to a web-based service for finding and delivering a large number of scholarly journals. The projects provided Elsevier with technical lessons and the university partners with a

better understanding of e-journal distribution and access issues associated with electronic journals. At the University of Michigan, which had been a TULIP partner, the experiences of project participation, combined with early experimentation with SGML, positioned Michigan to continue contributing to digital library development (Bonn et al. 1999). Participation in TULIP was an important stepping stone for other university participants as well.

Red Sage was supported by the University of California-San Francisco, Bell Labs and Springer-Verlag and ran from 1992 to 1996. The three partners assembled a large group of participating commercial publishers, scholarly societies, and university presses to build a digital library for the health sciences and serve as a laboratory to inform the transition from print-based to digital systems (Lucier and Brantley 1995). The participating publishers in Red Sage benefited not only from technology transfer but also from a better understanding of the economic and social issues associated with the electronic delivery of journals.

UK e-publishing projects. This was a large set of projects beginning in 1995 in the context of eLib (described in chapter 1). The projects were organized into seven program areas, among them on-demand publishing, digitization, electronic document delivery and e-journals (C. Rusbridge 1995). The experiences of these projects provided many UK universities with new skills and abilities to exploit information technology innovations (Kirriemuir 1996).

DeLLiver (Desktop Link to Virtual Engineering Resources). The DeLLiver testbed project was funded by professional societies, commercial publishers and federal agencies and began in 1998. It considerably advanced research and development of web-based access to full-text journals and articles (Mischo 2004, 7-10). Scholarly societies and publishers subsequently used the project's design insights and specific technologies to establish or improve their own full-text repositories and hyperlinking systems.

e-Depot. The Dutch national library extended its responsibility for the *legal deposit* of all Dutch publications to the digital era by making the decision to build an “e-Depot” in 1993 (Oltmans and Lemmen 2006, 63). This was a project that demonstrated to the digital library field what archiving, preservation, and legal deposit programs could look like in the digital age, and how national libraries could strike innovative, large-scale, mutually-beneficial agreements with commercial publishers and online information service providers. In 2010 the Dutch national library announced that it would upgrade e-Depot to become a “National Platform for Digital Publications.” The new platform is intended to aggregate e-content (as e-Depot has done) but also to deliver content from the national library’s ambitious mass digitization program to digitize all Dutch printed publications since 1470, some 730 million pages. The initial stage of the project involves partnerships with Google and Proquest (Janssen 2011).

Open access

The open access movement was another key outcome of the first decade of work in digital libraries. Among the early influencers is Stevan Harnad. In 1990 he published a paper (now frequently cited) that advocated extending the idea of an electronic archive to include digital prepublications of scholarly articles (*preprints*). The purpose was to harness the nascent forms of digital scholarship to take scholarly collaboration to a new level and “substantially restructure the pursuit of knowledge.” The copyright laws were among the obstacles he listed to realizing the goal.

There was reason for optimism: by the next year (1991), the Los Alamos research institute had begun to instantiate such a new model for digital scholarship and collaboration (Harnad 1999). The new model featured *self-archiving* of preprints and final refereed drafts. An early instantiation was the Los Alamos Physics Archive, which eventually became arXiv.org (see

table 2.1). Physicist Paul Ginsparg had developed the physics archive from the idea of a “centralized automated repository and alerting system ... a solution [that would] democratize the exchange of information” (Ginsparg 2011).

The foundation stones for what became a strong global movement for *open access* to scholarship included:

1. The opportunity to greatly improve scientific inquiry and the advancement of knowledge
2. An innovative re-conceptualization of the scholarly communication process for the digital era
3. Open online archives (repositories)
4. The concept of self-archiving

Advocacy and advocates for open access sprang up quickly, as further discussed in later sections of this chapter and chapters 4, 6 and 8.

A new world of scholarly research, teaching and learning

The impacts of these early projects and later investments in new systems by scholarly societies, publishers, indexing services, research institutes and open access advocates were enormous. If a time machine were to transport a set of graduate students and faculty members from 1990 to today, they might find their contemporary colleagues’ scholarly sources and practices almost unrecognizable. As Clifford Lynch, director of CNI, wrote in his review of the ways in which information technology has changed academic libraries (2000):

In the late 1980s, the world of scholarly communication, teaching, and research began to change as a result of networking and advanced information technology. We entered a decade characterized by an enormous, exhilarating flowering of innovation, creativity, and experimentation. The idea of networked information emerged ... International information sharing and collaboration were greatly facilitated. The use of the Net became critical in many forms of scholarly communication. Preprints and technical reports

became widely distributed on the Net, democratizing access to this critical information and speeding up the rate of communication... Scholarly communication became much more interactive through the use of technologies as mundane as mailing lists or as sophisticated as collaboratories.

Rapid adoption and changing work practices

Surprisingly rapid integration of the new systems and databases for electronic resources into everyday practices for research, teaching and coursework followed the digital transformation of the scholarly communication process. By fall 2001, a survey (Friedlander 2002) of over three thousand faculty, graduate students and undergraduates found that while the use of print sources remained important, 35% of faculty and 49% of graduate students reported they were relying exclusively or almost exclusively on electronic sources for their research.

Undergraduates were even more willing to shift to online research practices, with 49% reporting they used electronic sources exclusively or almost exclusively. Over time these trends have grown considerably stronger.

Technical innovations

This section covers the outcomes of first decade research and practice that advanced the enabling technologies of digital libraries. It begins with an outcome of DLI-funded digital library research at Stanford called PageRank. It next turns to outcomes that advanced interoperability. A third subsection considers outcomes that enable interlinking across digital sources. A fourth subsection considers the genesis of open access repositories.

PageRank

The physicist Paul Ginsparg (2011) concludes his retrospective on the 20th anniversary of the physics open archive with the insight “the Internet, World Wide Web, search engines, and other

developments described here all initially stemmed from the academic community's need to transmit, retrieve, and organize information." Indeed, the academic community's information needs drove many technological innovations and influenced what digital libraries produced in their first decade. A new world of scholarship was one of those outcomes. Another outcome that arose from the early work of the digital library field was an innovation that has changed the world for everyone—PageRank.

In April 1998 at the Seventh International World Wide Web Conference, Stanford graduate students Sergey Brin and Larry Page presented the results of research they had conducted as part of a team working on one of the six NSF DLI-1 projects (see chapter 1). That project, under the leadership of Hector Garcia-Molina and others, was called the Stanford Integrated Digital Library Project. Brin and Page's conference paper presented a prototype of Google and its underlying system for efficiently crawling and indexing the web, called PageRank (1998). PageRank uses the link structure of the web to produce what might be called "associative indexing"—an approach much like that envisioned by Vannevar Bush in 1945 when he proposed the "memex."

PageRank had its beginnings in the Stanford project's attempts to discover powerful new ways to find information (Stanford University. Digital Library 2005a, 2005b; Google 2012). Brin and Page presented their conference paper in April 1998. In September 1998 they founded Google, whose well-known mission is "to organize the world's information and make it universally accessible and useful."

Comparing the Google mission to the stated purpose of the Stanford Integrated Digital Library Project, funded under DLI-1 and continuing in DLI-2, suggests that these projects influenced

Brin and Page's bold vision for Google. The following quote from the Stanford award abstract demonstrates this strong connection (National Science Foundation 1998b):

This project - the Stanford Integrated Digital Library Project (SIDLP) - is to develop the enabling technologies for a single, integrated and "universal" library, providing uniform access to the large number of emerging networked information sources and collections. These include both on-line versions of pre-existing works and new works and media of all kinds that will be available on the globally interlinked computer networks of the future. The Integrated Digital Library is broadly defined to include everything from personal information collections, to the collections that one finds today in conventional libraries, to the large data collections shared by scientists. The technology developed in this project will provide the "glue" that will make this worldwide collection usable as a unified entity, in a scalable and economically viable fashion.

Among the outcomes of the first decade of digital libraries, the contribution of the Stanford digital library project to the creation of PageRank made significant progress toward the dream of a universal library. Google "emerged from the [DLI] funded work and has changed working styles for virtually all professions and private activities that involve a computer" (Paepcke, Garcia-Molina, and Wesley 2005).

Early support for interoperability

One vision of digital libraries that fueled this first decade's efforts included the notion that digital libraries would reflect a distributed environment; in other words, they would bring together diverse collections of information on different computer systems in different locations around the world. Interoperability and integration of search results in an understandable display for the user are the prerequisites for cross-searching, retrieval and display of diverse, distributed, complex digital objects.

“Interoperability” (in this context, the provision of uniform, coherent access to diverse information from different, independently managed systems) has proved to be a great and ongoing digital library challenge. Chapter 3 discusses the grand challenge of interoperability and the progress that has been made, starting with efforts using Z39.50, a protocol for information retrieval that pre-dates the web. The following section picks up the interoperability thread with an outcome of early digital library work, the Open Archives Initiative.

The Open Archives Initiative

The Open Archives Initiative (OAI, openarchives.org) was instrumental in defining a new framework for interoperable digital libraries. OAI has had a significant impact on how scholars distribute, share and discover research. Its origin is a meeting held in Santa Fe in October 1999 in response to a call to explore cooperation among scholarly e-print archives (Van de Sompel and Lagoze 2000; Lagoze and Van de Sompel 2001). The technical and organizational framework for OAI that emerged from the meeting came to be called the Santa Fe Convention, which was seen as the key to increasing the impact of open repositories and establishing real alternatives to scholars’ dependence on traditional journal publishing. The group’s work led quickly to the development of the OAI Protocol for Metadata Harvesting (OAI-PMH).

OAI-PMH

Participants at the OAI Santa Fe meeting representing arXiv.org, the California Digital Library, CogPrints, RePEc (a repository of papers in economics) and NCSTRL (a repository of technical reports in computer science) left the meeting with the intention to be early adopters of the Santa Fe Convention. The technical specifications for OAI-PMH (the metadata harvesting protocol) were released in May 2001 (Lagoze and Van De Sompel 2001; Lagoze et al. 2002). The intent was for OAI-PMH to be the “appropriate catalyst for the federation of a broad cross-section of

content providers.“ OAI-PMH represented a fresh, easier-to-implement approach to achieving interoperability for distributed digital libraries.

By adopting OAI-PMH, individual repositories make their metadata accessible in a standards-based way for harvesting by providers of search and discovery services. The framers of OAI-PMH intentionally chose a low-barrier, easy-to-implement approach. Many adopted OAI-PMH to enable interoperability with other metadata providers and allow harvesting of their data stores, thereby making their digital libraries more widely known. This strategy has paid off extremely well for making the content of open access repositories visible in search engine results. Perhaps more than any other first-decade digital library technical innovation, OAI-PMH has been a major factor in the rapid growth of open access repositories around the world.

Identifiers

Digital library researchers and builders have understood the central importance of persistent identifiers from the earliest days of digital library work. Identifiers are an essential component of the Kahn-Wilensky architecture of digital libraries (see chapter 1). Bermès 2006 introduces and explains the critical role of identifiers in the context of digital library projects. The keen appreciation of persistent identifiers continues to be a defining characteristic of digital library research and practice (chapters 3 and 9 further discuss identifiers).

The Handle System and DOIs

The Handle System and DOIs were key outcomes of first-decade digital library research. Kahn and Wilensky first developed the Handle System (handle.net) in 1993 (2006, 115). Today, handles are used to identify journal articles, technical reports, books, theses and dissertations, government documents, metadata and more. The International DOI Foundation's implementation of the Handle System is the DOI (Digital Object Identifier) system (2012). DOIs

were rapidly taken up by publishers and implemented as a critical part of the infrastructure for digital publishing. For example DOIs are now used by CrossRef (crossref.org), a consortium of nearly 4000 publishers.

URLs versus persistent identifiers

URLs are Uniform Resource Locators. Although the word “locator” is embedded in the phrase from which URL is derived, URLs are unreliable for locating and linking to things over time (see for example Nelson and Allen 2002). Everyone is familiar with broken links on the web.

Aware of the difficulties that web developers were having maintaining usable URLs, OCLC Research (1996) developed and made software freely available to help developers manage URLs in a way that would reduce the need for maintenance and provide long-term stability. This is the PURL (Persistent Uniform Resource Locator) software, which provides for flexible naming and resolution of URLs. OCLC completed a collaborative project to re-architect and release the PURL software as open source in 2007 (OCLC Research 2007).

Over time, best practices for web developers and digital library implementers favor Uniform Resource Identifiers (URIs), the technology for naming and addressing resources on the web (w3.org/Addressing; Baker and Dekkers 2003). Chapter 3 returns to the discussion of URIs and how they relate to digital libraries.

Reference linking

There is another enabling technology related to digital library interoperability. Information seekers expect to be able to link directly and immediately between sources like an article and its references, from citations in a database or online index, or from references in a catalog or bibliography. This functionality is called “reference linking.” The Open Journal demonstration project (Hitchcock et al. 1998) confirmed the value of links for providing faster and more direct

access to more information, enhancing the effectiveness of information retrieval, and adding value to electronic resources. It also influenced the development of what became widely used solutions for reference linking by scholarly publishers and online information services.

OpenURLs

Reference linking is particularly important in a hybrid library, where some of the resources may be represented by online citations but the text to which the citation refers is available only in print. Another application of reference linking is providing user access to the appropriate online version of an article, given the set of sources to which that user has access. Caplan and Arms' article on reference linking for journal articles (1999) provides a useful generic statement of the problem that reference linking solves: "given the information in a standard citation, how does one get to the thing to which the citation refers?"

Reference linking works best if the links persistently identify what users want to link to. Unfortunately, persistent identifiers do not exist for everything (or even most things). Early digital library research identified other methods for linking that have become familiar and widely adopted: OpenURLs and services based on them, such as SFX from Ex-Libris, a library system vendor active in the development of OpenURLs. Van de Sompel and Hochstenback (1999a, 1999b, 1999c) noted:

The omnipresence of the World Wide Web has raised users' expectations [for interlinking] ... When using a library solution, the expectations of a net-traveler are inspired by his hyperlinked Web-experiences. To such a user, it is not comprehensible that secondary sources, catalogues and primary sources, that are logically related, are not functionally linked.

An OpenURL provides a standardized way for an information service to capture and transfer metadata about an information object in one location, transport this data to another information service, then display the information object to the user. The digital library community's interest in OpenURL was immediate; it was approved for fast-track standardization by NISO, and it became an approved standard in 2004.

While the utility of OpenURLs in practice suggests some new work to improve linking, as suggested by Blake (2002), Chandler (2009) and Trainor and Price (2010), OpenURLs are now widely deployed by publishers and aggregators, library subscription agents, library system vendors, consortia and libraries.

The emergence of open access repositories

Two first-decade outcomes led to the emergence of open access repositories, a type of digital library that has had a substantial impact on the world's access to scholarly content. The origins of these outcomes can be traced to the OpCit Project and OAI-PMH.

At the end of the 1990s NSF and JISC funded six international digital library projects. One was the Open Citation Project (OpCit) with participants from Southampton University, Cornell University, and the Los Alamos National Laboratory (National Science Foundation 2001). Hitchcock and others (2002) tell the story of OpCit. The key, lasting technical outcome of the OpCit project enabled many to build open access repositories by producing the open source software called GNU EPrints. By the time OpCit concluded, the EPrints software (eprints.org) was being used by nearly 60 archives. As of this writing (June 2013), ePrints software is being used by 505 of the 3,430 repositories tracked by the Registry of Open Access Repositories (roar.eprints.org). EPrints can be said to have stimulated the subsequent development of other

repository software (like DSpace) and the building of many open access repositories (see chapters 4 and 8).

Digitization and digital preservation

Digital content is often created through digital reformatting. *Reformatting* converts an original object (that is, an object in its original form, like text or images) to a digital one that is not only easier to preserve, but also to compress for storage and manipulate with computer programs. This conversion process is called “digitization,” the process of converting a physical item into a digital representation or facsimile. Digitization relies on a number of enabling technologies, including scanning and OCR but also digital photography, re-recording and other techniques. Many types of materials held by libraries, museums and archives might be digitized: maps, music (printed and recorded), manuscripts, photographs and images of many kinds, videos, oral histories, 3-dimensional objects and microfilm or microfiche.

The following sections briefly describe the key outcomes of first-decade digital library work involving digitization and preservation:

- Large-scale digitization of scholarly journals
- Some early defining projects that established the value of digitization
- National library programs for cultural heritage materials
- Contributions to preservation
- The emergence of digitization specialists and best practices

Scholarly journals: JSTOR and other initiatives

JSTOR (see table 2.1) is an example of an organization with roots in the first decade of digital libraries. JSTOR (jstor.org) is not a publisher, but an independent non-profit organization

founded to help academic libraries and publishers. JSTOR, initially funded in 1994 and officially launched in 1997, is a key first-decade outcome because of its substantial influence on the development and creation of digital libraries of scholarly content, how journals are preserved, library management of shelf space for journal back files, the visibility and usage of older materials, and more (Guthrie 1997; Guthrie 2001). As of this writing JSTOR provides access to archival and current issues of more than 1,400 scholarly journals.

Other early projects to digitize scholarly journals include DIEPER at Göttingen University (Schwartz 1999), the Australian Cooperative Digitization Project (ACDP; Burrows 1999); and NACSIS-ELS (“the Japanese JSTOR”), which was launched in 1997 (Miyazawa 2005).

Early defining digitization projects

A number of early projects demonstrated the exciting potential of digitization, especially for broadening access and opening the study of cultural heritage materials to new audiences, freed of the boundaries of time and place. Three of these projects are:

- **Perseus Digital Library** (perseus.tufts.edu). The Perseus Digital Library (see table 2.1) focuses on primary materials related to classical Greco-Roman culture. Its development began in 1987 at Harvard and the project moved to Tufts in 1993 (Crane 1996). Perseus led the way in testing what happens when libraries move online, how digital technologies would live up to their promise (or not), and how to create an infrastructure for digital libraries that others could learn from. Perseus first appeared on the web in 1995. Perseus’ culture of participation allows not only faculty, but also student researchers and citizen scholars to interact with the art and archaeology, history, language and literature, philosophy and science of the classical world (Crane et al. 2012).

- **Dunhuang Caves.** A large library of ancient Buddhist texts, tablets, prints and artifacts were discovered in 1900 in a cave near Dunhuang, China. Eventually an entire complex of hundreds of caves, containing artifacts and painted walls, was discovered in the area. Dunhuang had been a caravan stop on the Silk Road from central Asia to China. Scholars soon visited the sites and took various treasures back to their own countries. In 1993, the International Dunhuang Project (IDP) began to develop an international database of collaboratively produced and shared digitized representations of the objects. A wealth of additional information about the project (including a timeline and the database) is on the IDP site (idp.bl.uk/idp.a4d). IDP demonstrated that digitization provides a way to virtually re-gather treasures that are dispersed around the world.
- **Gutenberg Bibles.** In 1996, Keio University in Japan led an impressive project to create digital facsimiles of its own and several others' surviving Gutenberg Bibles. The project was called HUMI (Humanities Media Interface). The online site (humi.keio.ac.jp/treasures/incunabula/B42/) makes the study of Gutenberg's early printing accessible to everyone and enables side-by-side comparisons of two copies of these incredibly rare books (Keio's and Cambridge University's).
- **Greenstone** (greenstone.org). The first New Zealand digital library project was organized by computer science researchers at the University of Waikato in 1995. The Waikato team's efforts had long-term significance because their efforts produced Greenstone—open source, freely available, multilingual digital library software for use by others (Witten et al. 1999; Witten, Bainbridge and Nichols 2009). Early experiences of the Waikato researchers with United Nations and humanitarian and development organizations eventually led to Greenstone's adoption in many countries around the world, including developing countries. As of this writing, Greenstone supports digital libraries in South America, Asia, Africa, the

Middle East, Europe and North America (results from June 2013 search of opendoar.org). In New Zealand, Greenstone is the basis of the highly popular PapersPast, a project of the National Library of New Zealand (see table 2.1). The global implementation of Greenstone revealed the potential of digital libraries to address not just multilingualism but also the *digital divide* (see chapter 6).

National library programs

This subsection presents a tiny number of additional examples of early national library digitization projects. Some national library digitization projects have already been mentioned in chapter 1 (e.g., American Memory). National library projects not only produced sites that have enabled broad, online, public global access to previously hidden cultural heritage materials; the lessons learned from the projects strengthened and guided the development of the digital library field.

- **Sounds.bl.uk.** This service of the British Library goes back years; the digital library part of the story begins in 1992, when the British Library began adopting digital audio technology for the purpose of broadening access to its world-class sound recordings archive. The mission of the earliest project, “Project Digitise,” was dual—access and preservation (Copeland 1994). That first project focused on the conversion and cataloging of recordings on wax cylinders from the collection of A.L. Lloyd (an authority on folksongs) and from a collection of ethnographic field recordings. Many other projects followed, notably the Archival Sound Recordings project from 2004-2009 (JISC 2007, 5-8), funded under the extensive JISC Digitisation Programme, which continues today (jisc.ac.uk/digitisation). As of this writing, the most recent digital-library-related initiative out of the BL Sound Archive is Sounds.bl.uk, which went live in January 2012 (see table 2.1).

- **Gallica**, the digital library of France, grew out of digitization activities at the Bibliothèque nationale de France (BnF) that began in the 1990s. Gallica first launched in 1997 with digitized content of books and journals, manuscripts, many types of images, maps, and more. The BnF has exemplified an assessment-based approach to digital library development (see for example Assadi et al. 2003, which inspired a whole series of digital library usage and user studies in France and elsewhere).

In addition, Gallica has exemplified a commitment to continuous improvement in its use of digital library technologies, for example implementing OAI-PMH for publishing Gallica metadata and harvesting from other digital repositories (Delorme 2011) and experimenting with linked data and semantic web approaches (see chapter 10).

The BnF's next wave of innovative digital library leadership, announced in May 2011, is a large-scale partnership to digitize half a million copyrighted out-of-print 20th-century French books. Digitization efforts will focus on the national library's legal deposit collections. The large, five-year project will enrich Gallica and be financed by the French Centre national du livre. The Jouve Group, a digital service provider, will do the digitization (CEPIC 2011). The BnF's approach to the project, based on an agreement between the French government, the French Publishers Association and the French Society of Literary Authors, promises to avoid the traps and delays of other large-scale digitization projects that have lacked such prior agreements between key stakeholders to address the complex rights and economic issues.

- **Picture Australia**. Picture Australia first began with a image digitization project in 1999. It was the foundation project for the large-scale, highly successful Trove digital library of the National Library of Australia (Cathro 1999; Cathro 2001; Cathro and Collier 2010; Holley 2010b; see also table 2.1 in this chapter). Picture Australia was fully integrated into Trove in

2012 (trove.nla.gov.au/general/australian-pictures-in-trove). The NLA's achievement with Picture Australia was an important outcome of the first decade of digital libraries because it exemplified a project that substantially progressed the public, open availability of historic photographs. Photographs are important primary sources documenting events, people and daily life, and their digitization has been a key to the public's enthusiasm for digital libraries. Chapter 10 further discusses the importance of digital library image collections on the social web.

- **Papers Past** (paperspast.natlib.govt.nz). Digitization and digital library technologies were quickly adopted to make historic newspapers—an unparalleled primary source—widely available to the public. Digital libraries of newspapers were highly significant outcomes of first-decade work; they greatly enhanced the work of researchers who were aware of the unique value of newspapers, but who faced either crumbling pages or miles of microforms and minimal indexing. Begun in 2001, Papers Past is a highly popular digital library of newspapers maintained by the National Library of New Zealand; it runs on Greenstone software described earlier in this section (see table 2.1).
- **The British Library Online Newspaper Archive** also dates from 2001. By 2010, the British Library was providing access to around four million pages of digitized content from British national and regional newspapers from 1600 to 1900, all searchable via a single interface (Deegan, Steinvel and King 2002; King 2005; Bingham 2010).

Individual institutions

In parallel with the large-scale initiatives funded at the national level, individual institutions—principally large research libraries—were building digital libraries and investing in digitization (an average of US\$286 thousand each in 2000; Greenstein and Thorin 2002, 66). Tonta's analysis

of digitization activities in Europe (Tonta 2008) documents the considerable digital library activity in individual institutions there.

Contributions to preservation

Library digitization programs were often linked to the long-term preservation of materials. For example, in the US, three reports of the US Council on Library and Information Resources (CLIR; clir.org/pubs/reports) published between 1990 to 2000 trace digital preservation practice in the US (Aaron Brenner, personal communication to the author, 24 May 2012). Chapman and Kenney (1996) articulated early baseline standards and working principles for digital imaging projects to preserve texts; Ostrow (1998) described the issues around preservation and access to digitized images of large historical pictorial collections; and Smith (1999) identified a number of false expectations of digitization as a preservation method. Chapter 6 continues the discussion of digital preservation.

Digitization specialists and best practices

Cooperative efforts to share the development of educational materials, document best practices, and deliver training are characteristic of the digital library field. For example, in the US, the Northeast Document Conservation Center (NEDEC) presented a “School for Scanning” starting in 1995 and helped get projects up and running. Best practices were documented in *A Framework of Guidance for Building Good Digital Collections* (framework.niso.org), a NISO Recommended Practice. *RLG DigiNews* (oclc.org/research/publications/newsletters/diginews.htm) also provided support and an international forum for sharing news. By the end of the first decade of digital libraries, the training programs, curricula, vehicles for information sharing, and the experiences of the projects themselves had produced a sizeable community of digitization specialists with a set of agreed best practices.

Metadata and standards

While *metadata* is often defined as “data about data,” this book uses the definition published by NISO: “structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource” (Guenther and Radebaugh 2004, 1). One of the most important outcomes of the first decade of digital libraries was a new world of metadata and standards. Arguably, the journey to this new world began in 1995 in Dublin, Ohio.

Dublin Core

As noted in chapter 1, computer and information scientists’ understanding of information retrieval had progressed enormously in the years leading up to the early 1990s. Librarians had been working on knowledge organization and cataloging theory and practice for a century, and from 1967 they had been gaining experience in encoding data (MARC) for use in and across automated library systems. A growing number of developers were working on internet and web standards. Humanities computing experts and archivists had been working on text encoding and finding aids. Fifty-two invited experts in these domains and several others convened for three days in March 1995 to collaboratively consider solutions to a problem: the web was full of valuable information resources but there was no good way to find and navigate them.

The workshop produced a proposal for a simple resource description record (the Dublin Core Metadata Element Set) and next steps for a standard, scalable, low-cost, interoperable way to describe a wide range of networked information resources (Weibel 1995). OCLC and the US National Center for Supercomputing Applications (NCSA) had convened the invitational workshop in Dublin Ohio—thus the name, Dublin Core. Another outcome of the 1995 workshop was the decision to convene an ongoing series of workshops, a series that has been going ever since (dublincore.org). The Warwick Framework, an architecture to accommodate a variety of

metadata models (Dempsey and Weibel 1996; Weibel and Lagoze 1997), came out of the second workshop. The Warwick Framework has had considerable impact on the technical development of digital libraries.

The sixth Dublin Core workshop in 1998 kicked off “a long co-evolution with the W3C’s Resource Description Framework (RDF) and the Semantic Web” (Weibel 1999; Weibel 2005). Chapter 3 continues the discussion of RDF and the semantic web.

The Dublin Core workshops have been building consensus through a dynamic process involving many stakeholder communities. OCLC provided support for the Dublin Core Metadata Initiative (DCMI) until it became an independent non-profit in 2008; in 2013 DCMI entered into a partnership with the Association for Information Science and Technology (OCLC Research 2009; ASIS&T 2013).

Metadata renaissance

Librarians and digital librarianship

By the time work on digital libraries got underway, librarians had over a century of experience producing bibliographies, catalogs, indexes and finding aids (Calhoun 2007, 174-175). They also had decades of experience with knowledge organization; for example, the first edition of the Dewey Decimal Classification System was published in 1876. By the 21st century it had been translated into many languages and was being used in over a hundred countries (Mitchell and Vizine-Goetz 2009). Even though libraries had begun to automate by the early 1990s, and the MARC format was widely deployed, library cataloging and classification methods in the early 1990s still reflected a world of information that was fairly stable and relatively small in scale, at least compared to today. Librarians generally produced one type of metadata (descriptive) and

used a few indexing vocabularies and document organizing methods (e.g., classification schemes) to manage library collections. The requirements for digital librarians were different.

The needs for scale and many new classes of metadata

Many new types of metadata and knowledge organization methods became necessary as digital libraries and networked electronic resources emerged. The new methods needed to cover content on a scale previously unimagined. Prior, mostly manual approaches could not scale to meet the need; in addition the scope of the requirements for metadata and knowledge organization expanded by an order of magnitude. These new conditions resulted on the one hand in a great deal of volatility and on the other, an exciting renaissance in metadata research and practice in which I have been fortunate to participate. Lagoze, Lynch and Daniel explained the new landscape for metadata (1996, under sections 6.1-6.3). It was clear that descriptive metadata would still be needed, but new classes and characteristics of metadata would also be required to:

- Support both human and machine-to-machine uses on the network
- Encode and mark up documents
- Define and manage collections of information resources at the collection level
- Support the preservation and archiving of digital objects (digitized and “born digital”)
- Create frameworks for accommodating metadata from many different communities (e.g., publishing, geospatial, museum, teaching and learning, multimedia ...)
- Represent and encode objects and metadata using many languages and scripts
- Persistently and reliably identify digital objects and their metadata (identifiers)
- Convey and adhere to the terms and conditions for use of digital objects and their metadata (rights; authentication and authorization)

- Manage digital objects and/or their metadata, e.g., date created, date last modified (administrative metadata)
- Describe attributes of digital objects, e.g., content ratings, reviews, usage, etc. (evaluative; statistical)
- Define the sources or origins of objects (provenance) or their metadata
- Convey relationships to other objects or link to them (linking)
- Enable the syndication and exchange of digital objects
- Indicate the components of objects and how to access or manipulate them (structural, technical)
- Define document types (DTDs)
- Move beyond text-based metadata to support many new types of digital media (e.g., images, audio, video)

The preceding list is not comprehensive, but it conveys a sense of the scope of the work that needed to be done.

From a library perspective, during that first decade, an entirely new set of conditions created disruptive change, moving the library field from bibliographic control to distributed systems for metadata management (Calhoun 2012b, under “metadata management”). These new conditions also created a new, multifaceted community of metadata and knowledge organization specialists, who produced an array of new standards, protocols, reference and data models, community-specific schemas/element sets and content rules, crosswalks, application profiles and more. For a quick look at the results of these widely distributed efforts, see Riley and Becker’s “visualization of the metadata universe” (2010).

Working digital libraries

So far this chapter has reviewed first-decade digital library outcomes that built a new field of endeavor, transformed the processes of scholarly communications, or delivered key enabling technologies. Early digital library work also produced working digital libraries that continue to attract significant attention today. The final section of this chapter provides information about some of these.

A sample from the first decade

Bearman (2007, 227-30) offers a useful framework for categorizing digital libraries. Table 2.1 adapts Bearman's categories to lay out some examples of digital libraries from different countries, their histories and funding sources. The choice of examples is deliberately limited to currently existing, working digital libraries whose roots are in the first decade of digital library research and development. Numbered citations in the right-most column of the table refer to the list of statistical data sources at the end of this chapter. Other citations in the table are incorporated in the list of references at the end of the book.

Discussion of sample digital libraries

The 15 sample digital libraries in table 2.1 produced lasting, real-world collections and services that have proved highly useful to specific communities of users. Many projects in the first decade of digital library work made transformative technical advances or helpful prototypes, but did not produce working digital libraries. The digital libraries in the sample were chosen to provide a set of comparison cases and facilitate the reader's consideration of why these early digital libraries continue to thrive. This topic will be taken up again in the later chapters of this book.

Conclusion

I have provided a framework that attempts to make sense of the outcomes produced by a momentous, intensively active ten-year period. Thousands of people and hundreds of organizations contributed to these outcomes. Inevitably, and with my apologies, I have given cursory treatment or unintentionally omitted some first decade activities that are important. The framework I have presented in this chapter reflects my own professional experience, an analysis of many hundreds (but certainly not all) sources, and a resulting perspective. Others' experiences and analyses might yield other useful perspectives on key outcomes. Yet all are likely to agree that the first decade's outcomes considerably advanced the grand vision of digital libraries, as well as creating a new field of research and practice to carry that vision forward.

These outcomes "set the stage, through examples, for a renaissance in research methods and practices, scientific and cultural communication and creative representation and expression of ideas" (Griffin 2005, under "Future Directions"). The renaissance indeed began. Over the next decade of progress in digital libraries (2002-2012), amid continuing technical progress, the challenges of online community-building, long-term sustainability, and digital library integration with the web came to the fore. The remaining chapters of this book explore how digital libraries are finding their place in the larger networked information environment of the web. By the end of the second decade, what emerged as central to the value of digital libraries went beyond their collections or content, services or technologies to their efficacy for supporting their communities and their web-based, real-world practices in information seeking, learning, research, knowledge creation and dissemination, work, and play.

Table 2.1. A sample of digital libraries, their histories and funding

| Type | Examples | History, funding and notes |
|--|---|--|
| National libraries | <p>National Library of Australia (NLA) Trove (trove.nla.gov.au) “Find and get over 289,890,268 Australian and online resources: books, images, historic newspapers, maps, music, archives and more” (home page). Dates to 1999 and “Picture Australia”; aggregates eight prior discovery services that had been organized by format (Cathro and Collier 2010).</p> | <p>NLA funds Trove; some content comes from external contributors. 56% of the traffic to the National Library of Australia website goes to Trove. It is also a popular destination in the US with about 45,000 unique visitors per month (1*)</p> <p>It should be noted that a component of Trove is PANDORA—one of the first web archives created and managed by a national library (Cathro 1999; Cathro 2001; Cathro, Webb and Whiting 2001).</p> <p>Chapter 10 discusses the significance of the Trove newspaper digitization project to the social web.</p> |
| | <p>Bibliothèque nationale de France (BnF) Gallica (gallica.bnf.fr) First established 1997 One million books, manuscripts, maps, images, periodicals, sound recordings, scores (home page).</p> | <p>BnF funds Gallica and is assisted by a number of digitization partners.</p> <p>51% of traffic to the BnF website goes to Gallica (2)</p> |
| | <p>US Library of Congress American Memory (memory.loc.gov) More than nine million items in one hundred collections. Includes access to written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music. First introduced 1994</p> | <p>The 1994 launch was supported by US\$13 million in private sector donations. It was the flagship service of the National Digital Library Program. Now supported through a combination of private sponsors and the U.S. Congress (see memory.loc.gov/ammem/about/sponsors.html).</p> <p>19% of the traffic to the Library of Congress website goes to American Memory. The site attracts nearly 350,000 unique visitors a month in the US (3)</p> |
| Discipline and subject-based digital libraries | <p>arXiv.org Open access service for pre-prints of articles in physics, mathematics, computer science, quantitative biology, quantitative finance and statistics. First started 1991 at Los Alamos National Laboratory (LANL). Over 700,000 articles; 60,000 annual submissions; 30 million downloads/year.</p> | <p>Funded by Cornell University since 2001 with some support from member institutions. Was supported from 1995 to 2000 by the US National Science Foundation, Los Alamos, and the US Dept. of Energy. arXiv has been widely adopted by the physics, math and computer science communities, which it serves by providing rapid access to research findings and a platform for open peer review. arXiv ranks highly in the Cybermetrics Lab’s “Ranking Web of World Repositories” and attracts over 100,000 unique visitors a month in the US (4)</p> |

| | | |
|---|---|---|
| <p>Discipline and subject-based digital libraries, continued</p> | <p>Perseus (perseus.tufts.edu) Covers the history, literature and culture of the Greco-Roman world.</p> | <p>Hosted by Tufts University. Began with a grant of US\$2.5 million from the Annenberg/CPB Projects; DLI-2 provided US\$2.8 million in 1998 (National Science Foundation 2007). Since then Perseus has received support in the form of grants from various federal agencies, the Mellon Foundation and individuals. (perseus.tufts.edu/hopper/grants). 22% of the traffic to Tufts University goes to Perseus, which attracts about 65,000 unique visitors per month in the US (5)</p> |
| | <p>ACM Digital Library (dl.acm.org) Access limited to subscribers. Published by the Association for Computing Machinery (ACM). Comprehensive collection covering computing and information technology. The full-text database includes the complete collection of ACM's publications, including journals, conference proceedings, magazines, newsletters, and multimedia titles. First introduced in 1997; significantly upgraded and reintroduced as the ACM Portal in 2001; reintroduced with new features in 2011 as the ACM Digital Library.</p> | <p>Funded by subscription fees and payments for downloading articles. ACM invested early in the move to online journals (Arms 2000, 51) and was one of several Collaborating Publishing Partners associated with the CNRI-funded D-Lib Test Suite that followed DLI-1. The partners benefited from the transfer of technology from the Illinois testbed of the DeLiver system, which allowed for experimentation with the retrieval and display of full-text journal literature in an Internet environment (Mischo 2002).</p> <p>77% of the traffic to acm.org goes to the ACM Digital Library. It attracts about 93,000 unique visitors per month in the US (6)</p> |
| <p>Genre or format-based digital libraries</p> | <p>JSTOR (www.jstor.org) Designed to substitute for back-issue files and serve as an archive of scholarly journals. Now “an integral part of the global academic research infrastructure” (Carr 2009, 67). Close to 44 million pages of content; over 7,000 participating institutions in 156 countries; journals come from 856 publishers.</p> | <p>Supported through JSTOR participant fees. It began with a grant from the Mellon Foundation to the University of Michigan, a participant in Elsevier’s TULIP project, for software development and production costs (Kohler 2009). JSTOR was established as an independent not-for-profit in 1995. Mellon awarded additional grants through the start-up period. JSTOR went live in 1997 with 190 libraries participating (Schonfeld 2003). By the end of 1997 Mellon had invested US\$5.2 million in developing JSTOR. JSTOR has been self-sustaining since 1999 (Kohler 2009, 225-227). JSTOR is reported to have nearly 1.4 million unique visitors per month in the US (7)</p> |

| | | |
|--|--|--|
| <p>Genre or format-based digital libraries, continued</p> | <p>ScienceDirect (www.sciencedirect.com) Provided by Elsevier since 1997. Access limited to subscribers. Offers more than ten million articles primarily from e-journals; also includes some book chapters. Elsevier journals are known for including the leading research in the physical, life and social sciences. Half a million additions per year; backfiles reported to go back as far as 1823.</p> | <p>Funding comes from subscription fees, which many libraries consider too high (Van Orsdel and Born 2009). Elsevier invested substantially in the early development of e-journals and online delivery systems. They organized the TULIP project in 1991 with nine US universities to test the networked desktop delivery of e-journals (Elsevier 2012, Kluiters 1997, Bonn et al. 1999). Concurrently they conducted an experiment with Tilburg University in the Netherlands (Collier 2004). In 1995 Elsevier introduced EES (locally-delivered e-journals) and also began developing the Web-based service that became ScienceDirect, whose beta release was in 1997. The Koninlijke Bibliotheek (KB), national library of the Netherlands, archives all Elsevier journals. ScienceDirect is reported to have over one million unique visitors per month (US only). (8)</p> |
| | <p>Papers Past (paperspast.natlib.govt.nz) Began in 2001. Contains more than two million pages of digitized New Zealand newspapers and periodicals from 1839 to 1945 and includes 70 publications from all regions of New Zealand.</p> | <p>Hosted and supported by the National Library of New Zealand. Began in 2001; relaunched in 2007 using Greenstone, a suite of open source, multilingual software for building digital libraries (NLNZ 2007, Boddie et al. 2008, Thompson Bainbridge and Suleman 2011). Greenstone, an early and well-known player in the digital library arena, developed its system as part of an international cooperative effort. 49% of the traffic to the website of the National Library of New Zealand goes to Papers Past (9)</p> |
| | <p>NDLTD (ndltd.org) The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization that began in 1996 at Virginia Tech. Participating institutions grew from 20 in 1997 (Fox et al. 1997) to over 200 today. In 2010 the NDLTD Union Catalog contained one million electronic theses and dissertations (ETDs) from contributing institutions from over 25 countries on all continents.</p> | <p>Supported by membership fees from about 200 NDLTD members. The NDLTD Union Catalog runs on systems provided by Scirus and VTLS. With origins dating back to 1987, the initial 1996 funding from Virginia Tech for developing a working system was supplemented by a three-year grant from the US Dept. of Education). Additional support came from public and private sector partners over the years. NDLTD was incorporated as a non-profit in 2003 (Hagen, Dobratz and Schirmbacher 2003). It has become an important ETD program for developing nations. NDLTD, its annual ETD conference and its director Edward Fox have been key influencers in the development of digital libraries as a field of endeavor. The project and the conferences have given a major boost to the adoption of ETDs at universities worldwide.</p> |

| | | |
|--|---|--|
| <p>Genre or format-based digital libraries, continued</p> | <p>British Library Sounds (Sounds.bl.uk) Began in 1992 with “Project Digitise.” Other projects followed, notably the Archival Sound Recordings Project from 2004-2009. A new version went live in January 2012 containing two levels of online access to 50,000 selected recordings of music, spoken word, and human and natural environments.</p> | <p>Supported out of the British Library Sound Archive, one of the world’s largest collections of sounds, and from 2004-2009 through the JISC Digitisation Programme, a set of large-scale projects with multiple phases. The British Library has set up innovative terms and conditions for online access to the recordings on the Sounds website; some of the content is freely available to all, and all 50,000 recordings are open to users from UK higher education institutions. Sounds.bl.uk is a popular destination on the British Library website.</p> |
| <p>Mission and audience-directed digital libraries</p> | <p>Project Gutenberg (gutenberg.org) Begun in 1971 by founder Michael Hart with the goal of providing free access to literary works in the public domain. The first producer of ebooks and the oldest digital library. Offers over 40,000 free ebooks. More ebooks are available through affiliates.</p> <p>Internet Archive (archive.org) Founded in 1995 by Brewster Kahle (11), the Internet Archive (IA) is a mission-oriented digital library and archive of internet sites, texts, music, moving images, recordings and software. The IA is an active advocate for open, universal and free access to knowledge. The Wayback Machine provides access to archived versions of an estimated 220+ million websites. Three other popular digital library projects from IA are the Open Library (openlibrary.org), Archive-It (archive-it.org) and publicly-available digital images from NASA (nasaimages.org).</p> | <p>Supported by volunteers and donations to the Project Gutenberg Literary Archive Foundation, a non-profit organization. After starting at the University of Illinois and transferring for a time to Carnegie Mellon, the Gutenberg system is now hosted by ibiblio, an online, public “collection of collections” supported by the University of North Carolina at Chapel Hill. Gutenberg is estimated to have over 500,000 unique visitors a month in the US (10)</p> <p>IA is a non-profit organization. Funding for projects and services comes from the Kahle/Austin Foundation with support from other partners over the course of developing particular projects. IA also solicits donations. IA is reported to attract around <i>three million</i> unique visitors a month in the US alone; other web traffic analysis services place it among the top few hundred busiest sites worldwide. Traffic to the Wayback service is reported to account for over 75% of IA traffic (12).</p> <p>The Open Library is reported to attract nearly 400,000 unique visitors a month (13), while Archive-It attracts about 18,000 unique visitors a month (14) and NASA Images attracts about 12,000 unique visitors a month (15). All estimates are for the US only.</p> |

| | | |
|---|---|---|
| <p>Mission- and audience-directed digital libraries, continued</p> | <p>SciELO (scielo.org) SciELO (Scientific Electronic Library Online) began in 1997 in Brazil (scielo.br) with the mission of enabling cooperative e-publishing in developing countries. The SciELO network (scielo.org) expanded and now includes eight national collections and two thematic collections in public health and the social sciences. Includes more than 500 Latin American open access journals and 191,000 articles.</p> | <p>Publicly funded by the State of São Paulo Research Foundation and BIREME (the Latin America and Caribbean Center on Health Sciences Information), an organization belonging to PAHO (the PanAmerican Health Organization) and to WHO (the World Health Organization) (Marcondes and Sayão 2003). It was one of the first collections of open access journals in the world. SciELO has brought considerably greater impact to Brazilian and Latin American journals (Packer et al. 2010). Currently (2012) SciELO.br is ranked by the Cybermetrics Lab as the top portal in the world. Of the Cybermetrics Lab’s top fifteen rankings of portals, six are SciELO sites (16).</p> |
| | <p>ICDL International Children’s Digital Library (www.childrenslibrary.org) Available since 2002. A mobile application for iPhone and iPad has been available since 2008; a second mobile application (StoryKit) was released in 2009 (Bederson, Quinn and Druin 2009; Quinn et al. 2009). ICDL’s mission is to support the world’s children by building a digital library of freely available, multilingual, online and outstanding children’s books from around the world. Contains over 4,500 books in 61 languages. Visitors come from 228 countries (17).</p> | <p>Administered by the International Children’s Digital Library Foundation, a non-profit founded in 2006, with continuing support from NLF, IMLS, and the Library of Congress. Initial funding came from the US National Science Foundation and other publicly funded agencies; ICDL was one of the six-year projects funded under the DLI-2 initiative. Research and development started in 1999 at the University of Maryland with an interdisciplinary team led by the Human Computer Interaction Lab and the College of Information Studies. Initially the Internet Archive hosted the ICDL site. A high-impact result—beyond the creation of the digital library itself—was validating the importance of working with the primary user group (in this case, children) to design digital libraries and services (Druin et al. 2003 and Druin 2005).</p> <p>The ICDL attracts about 24,000 unique visitors a month from the US. No data is available for non-US visits to the site but it is an important destination outside the US. (18)</p> |

*Numbered references in this table refer to the list of statistical data sources at the end of the chapter.

References to websites in Table 2.1

This list cites the data sources for the statistics reported about the sample of working digital libraries. Most of the statistics came from Alexa (alex.com) and compete.com, which are well-known providers of global or US web metrics for websites, as they were reported in April 2012. The following references are numbered in Table 2.1.

1. Site and web traffic information for the National Library of Australia, including Trove.
www.alex.com/siteinfo/nla.gov.au. Unique visitors from the US:
<http://siteanalytics.compete.com/trove.nla.gov.au/>
2. Site and web traffic information for the BnF, including Gallica. www.alex.com/siteinfo/bnf.fr
3. Site and web traffic information for the Library of Congress, including American Memory.
www.alex.com/siteinfo/loc.gov# and <http://siteanalytics.compete.com/memory.loc.gov/>
4. Site and web traffic information for arXiv.org: Cybermetrics Lab ranking
<http://repositories.webometrics.info/toprep.asp> and monthly traffic
<http://siteanalytics.compete.com/arxiv.org/>
5. Site and web traffic information for Tufts University, including Perseus.
www.alex.com/siteinfo/tufts.edu#. Perseus traffic analysis:
<http://siteanalytics.compete.com/perseus.tufts.edu/>
6. Site and web traffic information for the Association for Computing Machinery, including the ACM Digital Library. www.alex.com/siteinfo/acm.org#. ACM Digital Library traffic analysis:
<http://siteanalytics.compete.com/dl.acm.org/>
7. Site and web traffic information for JSTOR: <http://siteanalytics.compete.com/jstor.org/>
8. Web traffic and ranking information for ScienceDirect:
<http://siteanalytics.compete.com/sciencedirect.com/>
9. Web traffic and ranking information for Papers Past: www.alex.com/siteinfo/natlib.govt.nz#
10. Web traffic and ranking information for Gutenberg.org:
<http://siteanalytics.compete.com/gutenberg.org/>
11. From the "About" pages on the Internet Archive website: "Since the mid-1980s, Kahle has focused on developing technologies for information discovery and digital libraries. In 1989 Kahle invented the

internet's first publishing system, WAIS (Wide Area Information Server) system and in 1989, founded WAIS Inc., a pioneering electronic publishing company that was sold to America Online in 1995. In 1996, Kahle founded the Internet Archive which may be the largest digital library. At the same time, he co-founded Alexa Internet, which helps catalog the Web. Alexa was sold to Amazon.com in 1999.”
<http://archive.org/about/bios.php>

12. Web traffic and ranking information for Internet Archive: <http://siteanalytics.compete.com/archive.org/>
and www.alexa.com/siteinfo/archive.org#
13. Web traffic and ranking information for the Open Library:
<http://siteanalytics.compete.com/openlibrary.org/>
14. Web traffic and ranking information for Archive-It:
<http://siteanalytics.compete.com/archive-it.org/>
15. Web traffic and ranking information for NASA Images:
<http://siteanalytics.compete.com/nasaimages.org/>
16. Information for SciELO: <http://repositories.webometrics.info/topportals.asp>
17. Information in the “About” and “FastFacts” sections of the ICDL website: <http://en.childrenslibrary.org/>
18. Web traffic and ranking information for ICDL: <http://siteanalytics.compete.com/childrenslibrary.org/>