1 Emergence and definitions of digital libraries

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 traces the first decade of progress in digital libraries (1991-2001), with emphasis on the foundational innovations, vision, motivations, new technology, funding and early programs that prompted their emergence and rapid development. It next turns to the question of how to define the concept of "digital libraries" in an environment of multiple perspectives and continuous technological and societal change. The chapter's intent is to orient the reader to the field as well as to ground the rest of the book in the context of the aspirations and efforts of many diverse communities and individuals.

The emergence of digital libraries (1991-2001)

This book places the beginning of digital libraries in 1991, the year in which the National Science Foundation (NSF) in the United States sponsored a series of workshops on how to make digital libraries a reality, not just a dream. At the same time, digital libraries are an outcome of the revolution in computing, telecommunications, and information systems that began almost 40 years ago, around 1965. This section frames the emergence of digital libraries as a recognized field of endeavor in terms of four requirements for viability and growth: a compelling vision, strong motivating factors, technology and funding.

Keywords: Digital libraries—History; Definitions of digital libraries; Digital Libraries Initiative (US); eLib Programme (UK)

A compelling vision

Many authors (Arms 2000; Fox 1993a; Lesk 2004; Tedd and Large 2005) trace the vision of digital libraries to a post-World War 2 paper by Vannevar Bush called *As We May Think* (1945) and a book called *Libraries of the Future* by J.C.R. Licklider (1965). Licklider's research for the book was sponsored by the US Council on Library Resources (Clapp 1965, ix). Bush, at that time director of the US Office of Scientific Research and Development, called for a new approach to information organization and discovery based on a the visionary concept of a "memex"—a fast, flexible and efficient desktop device enabling associative indexing and instant access to both a vast library and a scientist's personal files.

The ideas and writings of Licklider, a professor of computer science at MIT, vice president of a high-technology company and imminent researcher for the Defense Advanced Research Projects Agency (DARPA), eventually led to ARPANET, a system of networked computers that preceded the internet. At the outset Licklider's *Libraries of the Future* focuses less on technology and more on solving the basic limitations of printed materials and the bricks-and-mortar libraries of the time:

> If books are intrinsically less than satisfactory for the storage, organization, retrieval, and display of information, then libraries of books are bound to be less than satisfactory also. We may seek out inefficiencies in the organization of libraries, but the fundamental problem is not to be solved solely by improving library organization at the system level. Indeed, if human interaction with the body of knowledge is conceived of as a dynamic process involving repeated examinations and intercomparisons of very many small and scattered parts, then any concept of a library that begins with books on shelves is sure to encounter trouble (Licklider 1965, 5).

Noting that "the 'libraries' of the phrase, 'libraries of the future,' may not be very much like present-day libraries," and "in the present century, we may be technically capable of processing the entire body of knowledge in almost any way we can describe," Licklider went on to create a prescient list of criteria for the future library that reflects both the progress and aspirations of 21st century libraries (1965, 1, 20, 36-39).

Key developments from 1965 to the early 1990s

Licklider laid out his challenging "libraries of the future" vision in 1965. Over the next 25 years, the technologies needed to build digital libraries became not only available but affordable—for example, digital storage, processors, connectivity, natural language processing, text formatting and scanning, optical character recognition (OCR), indexing and more (as discussed by Lesk 2004, 16-89). Perhaps most importantly, the promise of the internet (dedicated in its earliest years to research-oriented use) for public and commercial use had captured the public imagination as well as the interest of the private sector and research professionals (Weingarten 1993; Stoker 1994; Ginsparg 2011).

The computer and information sciences

Computer and information scientists made enormous progress in information retrieval theory and systems between 1965 and 1990. Computer scientists advanced the knowledge and understanding of architecture and systems, and information scientists complemented their work (Arms 2012, 581). Howard D. White and Kate McCain's renowned analysis of the structure of the information science discipline between 1972 and 1995 indicates that the discipline was principally focused on information retrieval and user-system relationships; bibliometrics; automated library systems and online catalogs; science communication; and user theory (White and McCain 1998). All of these created a solid foundation for the emergence of new research on digital libraries.

Online information industry

The online information industry predates the internet and the web. It had its start in the 1970s and by the early 1990s, it was a US$12 billion industry (1992 dollars), serving mainly the business sector (Calhoun 1994, 2). There was however a segment of the industry called "scientific, technical and diversified online services" that served primarily research and education; the market leaders in the early 1990s were Mead Data Central (NEXIS/MEDIS), Dialog and InfoPro Technologies (BRS/ORBIT) (Calhoun 1994, 4-5). Dialog dates to 1972 (O'Leary 1993).

In the early 1990s the online information industry took the form of online host services that mounted databases and software from which subscribers could retrieve information using first, dedicated terminals and later, personal computers. The firms that offered these services relied on content providers (database producers, publishers, abstracting and indexing services) and reliable, commercially-available telecommunications networks (providing dial-up services). The supply of online content was already relatively large by the early 1990s; online databases grew from around 300 in 1979 to nearly 5,200 in 1993 (Calhoun 1994). CD-ROM database vendors had also entered the market for digital information by that time.

The growing adoption of personal computers not just by businesses and other organizations but also in homes, together with the advent of internet access (which was faster and cheaper than the existing commercial networks) led to both amazing opportunities and large challenges for the most successful online services and content providers of the time. The internet has long roots, and many had been were aware of its potential for years. For example, in a 20-year retrospective piece he wrote in 2011, the well-known physicist Paul Ginsparg notes that he first used email on the original ARPANET, which preceded the internet, while a freshman at Harvard in 1973.

Online information industry services and content providers (e.g., publishers and professional societies) were faced with managing the disruptive risks and opportunities of the "information superhighway" and full-text digital content in order to maintain (or improve) their positions—or else risk extinction. This same set of new conditions encouraged the entry of many new players providing online information and services.

Libraries, standards and automation

Libraries were early adopters of online information systems, and highly trained reference librarians served as intermediaries conducting searches of the very expensive online services, which had expert, non-intuitive interfaces *not* designed for end-user searching. In addition, for library information technology and technical services operations, the first distribution of MARC (Machine-Readable Cataloging) records from the Library of Congress in 1968 (Avram 1969) was a great leap forward. Over the ensuing years MARC had a transformative influence on libraries, as did the founding in 1967 of the first shared computerized cataloging system based on MARC, the Ohio College Library Center (now OCLC Online Computer Library Center; Kilgour 1969).

The MARC record and these new systems quickly created a new plane for library technological advances. MARC made it possible to aggregate large structured data sets to underpin the conversion from printed to online catalogs of library holdings; the first generation of robust automated systems for libraries; and many new services in libraries (for examine, interlibrary lending became much easier, faster and less costly). All of these developments together put libraries in a position to be early adopters of many new information technologies, the internet and the web. Thanks to this long foreground, libraries were also ready for digital library collections, systems and services (Calhoun 2003, 282).

The Follett report

In the UK, a great deal of experience and knowledge of the latest information technologies and networks, predating the internet and the web, led up to the Follett report (1993). The UKOLN (UK Office for Library and Information Networking) had been established in 1990 (Stoker 1994, 119). Just two examples that reflect the current topics of the 1980s are from Brindley, who was writing about strategy for the "electronic campus" and the shift from print and CD-ROM to online dissemination of scholarly content (1988; 1989); and from Law, an expert on library automation since the 1970s, who among other topics was writing about projects to get the nation's academic library catalogs online (Law 1988). The logical extension of all this work was the Follett report, which placed academic libraries high on the UK national agenda for higher education and quickly generated large-scale national funding for the development of "electronic" or "virtual" libraries (the eLib Programme, discussed later in this chapter).

Archives and other professional communities

A foundational development that came out of the archives, humanities computing, linguistics and other professional communities was the Text Encoding Initiative (TEI), which produced a standard for encoding scholarly texts in machine-readable form. TEI, intended to support data interchange in humanities research, can be traced to a conference of the Association for Computers and the Humanities in 1987. The then newly available Standard Generalized Markup Language (SGML) was the needed spark to kick off the development of TEI and a new way of supporting textual research on the network (Ide and Sperberg-McQueen 1995).

Daniel Pitti (1997) describes how the advent of the internet inspired the archival community to renew its efforts to bring geographically distributed primary resources together in a way that would enable universal intellectual access. Foundational (pre-internet) work was accomplished from 1981 to 1984 when a US National Information Systems Task Force of the Society for

American Archivists paved the way to a MARC standard for the encoding of records describing archives and manuscripts. MARC records provide for the online discovery of archives and manuscript collections at the collection level, and in a library context; but machine-encoded finding aids were needed to actually lead to the materials in the collection. Archivists' next step was to develop a standard, computer-based encoding structure for finding aids. This work began in 1993 and produced the Encoded Archival Description or EAD standard. SGML is the technology underlying EADs. The development of EAD and experience with SGML were momentous developments that aligned the archival community's work with the web and the networked digital environment that was emerging.

Other developments

Given the limits of this space and my time to conduct the necessary research, this section's mini-analysis of the conditions leading to digital libraries from 1965 to the early 1990s is far from complete. I have merely touched on the work of some disciplines, organizations and communities of practice and not discussed others' contributions at all. In addition to the roles played by early research on the internet (which goes back to the 1970s) and by computer and information scientists, the online information industry, archives and libraries, the efforts of countless researchers and implementers intersected with, ran parallel, or contributed directly to the origins of digital libraries. These include the individuals and groups who developed the internet and web standards, open systems and other core aspects of networking; those who pioneered new ways of marking up and encoding text; the geospatial or informatics communities; teaching and learning communities; and more. While recognizing these many contributions, I have focused this and the next chapter chiefly on the roles of computer and information scientists; libraries and the cultural heritage sector; and scholarly communities, content providers and online services.

An ambitious agenda

Christine Borgman (2007, 21), writing of the political aspects of new, large-scale research programs, noted "visions must be grand to attract attention and the promised outcomes must be ambitious to attract money." By the start of the last decade of the 20th century, computer and information scientists, scholarly content providers and libraries were ready to embrace an ambitious agenda. They were ready for the next steps toward the systems that Bush and Licklider had envisioned in 1965. Building the first digital libraries was not just feasible: it was the logical next step for researchers and professionals in many fields. Elements of the vision of digital libraries that fueled scholarly and public interest in the first decade of digital library research and development, starting around 1991 included:

- Easy, fast, and convenient access to the world's information (regardless of where that information is stored) at any time, from anywhere in the world
- Effective storage and organization of massive amounts of text, multimedia and data beyond the bounds of what even the largest single library could provide
- Organization and access to materials in many languages
- Greatly improved searching and browsing capabilities
- Interoperability enabling the cross-searching of many diverse collections at once
- Direct, instant delivery of information and data to multiple users at the same time
- Transformative improvements in support for research and education globally; better support for interdisciplinary work and scholarly collaboration across institutions and around the world
- Significant cost savings over traditional (duplicative) methods for cataloging, storing and preserving analog materials

Strong motivating factors

As if the grand opportunities were not enough, two more powerful motivating factors converged in the early 1990s to make the time right for digital libraries. One was a sense of urgency to solve the pressing issue of an explosion of scholarly information; the other, already mentioned, was a sense of opportunity that arose in firms and communities of practice supporting scholarship. First, publishers, professional societies and indexing services seized on technological advances to improve the information storage and retrieval systems they used. Second, libraries and cultural organizations saw an opportunity to preserve and extend access to valuable collections through digitization.

A sense of urgency

Runaway growth

Both scholars and librarians have considered digital libraries to solve large-scale, long-standing challenges. Chief among them is the need to make an increasingly overwhelming volume of material accessible and available. Writing at the conclusion of World War 2, Bush (1945) noted the “growing mountain of research” and the difficulty posed by an explosion of scientific publications “extended far beyond our present ability to make real use of the record.” In the UK the sense of urgency was similar, in that it was centered on the perception of runaway growth, but of a different nature. UK national attention was focused on specific problems facing UK higher education and academic libraries—increasing costs for materials coupled with a huge expansion in student populations—and the opportunity to solve them by effectively harnessing the technologies of the global information revolution (Carr 2002).

The notions of runaway growth were fueled by other early predictions as well. Although his methodology was later called into question (Molyneux 1994), librarian Fremont Rider’s conclusion in *The Scholar and the Future of the Research Library* (1944)—that research

libraries would double in size every 16 years—firmly established a sense of urgency around solving the problem of runaway library growth.

How much information?

Rider's methods may have been flawed, but he was not wrong about runaway growth in the world's information, including scholarly information. Figure 1.1 pulls estimates from a report of how much information was consumed by Americans, from what sources, in 2008 compared to 1980 (Bohn and Short).

As it turns out, there was and continues to be an information explosion, although not perhaps in the ways that Rider and digital library pioneers anticipated. Much more information does exist, and people spend more of their time consuming more of it. Bohn and Short estimate that the number of hours of information consumption per person grew 2.6% a year from 1980 to 2008 (2009, 7). The size of research library collections have not doubled every 16 years, but the amount of information available and of interest to an academic community has exceeded that growth rate, and there is more to read than ever, and reading has increased since 1980: this is because there are now so many more ways to consume words (text). The report estimates that a third of the words that people consume come through interactions with computers, and the overwhelmingly preferred way to receive words is via the internet.

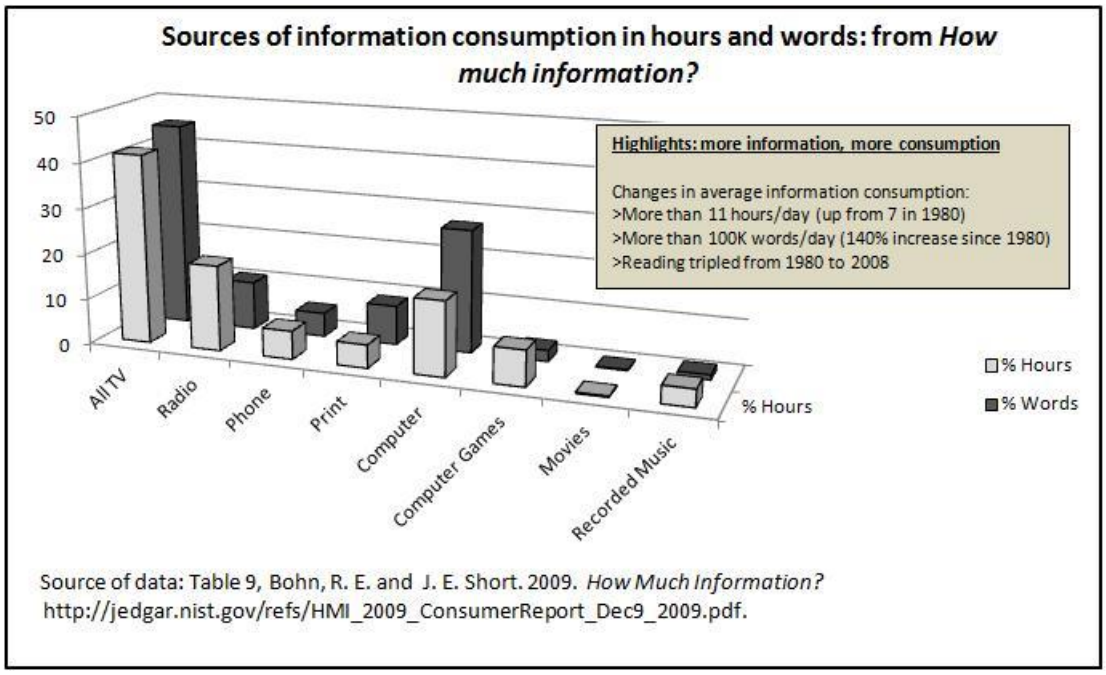

Figure 1.1. Validation of the predictions of runaway growth: more information and more consumption

A new world

Scholarly communication

"Scholarly communication" alludes to the communicative activity of scholars (people engaged in creating original scholarly works), in particular how they communicate as writers, linkers (e.g., citing others' work), submitters/disseminators (the choice of formal and informal communication channels, e.g., journals, conferences, wikis, blogs), and collaborators (Borgman and Furner 2002, 6). Quite a few groups, classes of individuals and tools contribute to the *process* of scholarly communication, but stand outside scholarly communication itself: a few of these are peer reviewers, tenure review boards, evaluation tools or systems (e.g., citation counts), editors, publishers, professional societies, online information services, libraries, and of course readers/information seekers and those who annotate or comment on scholarly work informally.

Innovation in the process of scholarly communication was already well underway by the early 1990s, and the early achievements of online information services, publishers, professional societies, and indexing services are impressive. At the time digital library work got underway, major scholarly societies and publishers saw new opportunities and were keenly interested in developing better systems for publishing full-text journals and articles.

Mercury and CORE

Two early projects, Mercury and CORE, influenced the rapid development of new kinds of networked, online retrieval systems that made papers from many independent sources appear as one integrated service. Mercury began in 1991; it was a pre-web solution for bringing together computer science articles from three different scholarly publishers. It validated new concepts for converting, storing and delivering page images from distributed sources over the Carnegie Mellon University's campus network (Arms 2012, 581). CORE was an early project and key influencer of methods for scanning document collections. It ran from 1991 to 1995, digitizing about 400,000 pages from 20 chemistry journals and demonstrating successful ways to build a full-text index for retrieval and display of digitized documents (Arms 2012, 588). Chapter 2 discusses other early projects (TULIP, Red Sage and others).

The scholarly content providers and services that supported these projects brought a sense of opportunity and substantial resources to early digital library research and development, and these projects had a powerful impact. Indeed, together with many subsequent investments by scholarly societies, publishers, indexing and online service providers, the early projects eventually transformed scholarly publishing and the expectations of faculty and researchers that scholarly content will be not only online, but also interlinked.

Digitization

Librarians and other professional communities also drove the development of early digital libraries and the technologies underpinning them. Early efforts to preserve the treasures held in library collections, archives and museums through digital reformatting (digitization) took off in the early 1990s; these are also discussed in chapter 2. By 1995, there were baseline standards, working principles, and a small but growing community of digitization specialists available for digital imaging projects for texts, pictorial images and more. This field of specialization eventually grew beyond the library and cultural heritage community and spawned mass digitization projects and the public's growing demand for books in digital form (e-books).

Technology

The last barriers removed

As mentioned already, personal computers, the internet and the web were also catalysts enabling research and development of digital libraries; these technologies were firmly in place before the web's first iteration in 1989 (attributed to Tim Berners-Lee, then at CERN) and the enthusiastic take-up of the Mosaic browser starting in 1993 (the US National Center for Supercomputer Applications at the University of Illinois had developed Mosaic (Ginsparg 2011, 5). These new innovations scaled up the size of prior opportunities to build services for collections stored in digital forms, retrievable over networks.

Arms reported a series of technical developments in the early 1990s that "removed the last fundamental barriers to building digital libraries" (2000, 10):

- Storing information on computers became significantly less costly
- Major advances had been made in the quality of personal computer displays
- Receiving information over the internet became fast, affordable and reliable

- Portable personal computers had become affordable and powerful

The Kahn Wilensky architecture

The general principles for the design of a digital library that is "open in its architecture and which supports a large and extensible class of distributed digital information services" were laid out by Kahn and Wilensky (1995; Arms 1995). Kahn and Wilensky strongly influenced how early digital libraries were built by technologists. Micah Altman (2008) characterizes the "Kahn-Wilensky architecture" as having four main types of components:

- *Repositories*, ranging from file systems to distributed storage systems for content
- Mechanisms to support *search* (indexing or metadata)
- *Identifier systems* for identifying and locating digital objects
- *User interfaces* to perform user services (for example searching, browsing, visualization, delivery)

This is not an exhaustive list. Other components of digital library architecture include, for example, security systems for authenticating users, services to aggregate search results from multiple sources, and tools for supporting collaboration and other types of interaction.

Interoperable, web-based digital libraries

By 1991, computer scientists already had extensive experience with the development of information retrieval (IR) systems. Fox and Sornil (2003) wrote "DLs can be regarded as extended IR systems with multiple media and federation." As web search and retrieval tools improved and gained acceptance over the course of the decade, digital library researchers and professionals sought new approaches to integrate their methods with web technology and the network. Lorcan Dempsey, then at UKOLN (1994), offered a prescient analysis of how the internet and web would generate entirely new kinds of systems and move information creation,

publication and discovery to the network. This is indeed what has happened. He also foresaw the immense challenges that libraries would face aligning and integrating their traditional knowledge organization practices, metadata silos, and fragmented information systems with the new networked environment—things he has continued to write and speak about today.

Computer scientists and librarians began to respond to the challenges. For example, early in the new millennium, Cornell and the University of Virginia began work on Fedora (Flexible Extensible Digital Object and Repository Architecture), a new system for digital library architecture. The intent was to provide a new framework for interoperable, web-based digital libraries (Payette and Staples 2002). As will be discussed in chapter 3, interoperability (the provision of uniform access to diverse information stored on different computer systems in different locations) has proved to be an ongoing challenge facing digital library developers.

Hybrid libraries

Interoperability became increasingly important as digital library projects, publishers, professional societies, indexing and online services brought content online and demand grew for unified access to content locked up in separate systems with separate interfaces. Furthermore, libraries began looking for ways to integrate the digital content with their predominantly non-digital collections (printed books and journals, prints, slides, maps, analog sound recordings and films, government documents, etc.). Rusbridge (1998) usefully described library collections in four categories: *legacy* (non-digital), *transitional* (legacy resources that have been or will be digitized), *new* digital resources (those expressly created as digital), and *future* digital resources. This book refers to the third and fourth categories as born digital resources. Rusbridge called for the development of technologies, systems and services for the "hybrid library*,"* which would integrate all four categories of resources. As discussed in chapter 5, from

early days to the present, the necessity of accommodating the requirements of hybrid libraries has been a key driver in the field of digital libraries and the profession of librarianship.

Funding

This section provides a high-level summary of key national and international funding sources and programs in the first decade of digital libraries. A subsequent section, which contains a review of early large-scale digital library programs, also contains information about funding. The next chapter also incorporates information about funding from foundations, membership organizations, individuals, commercial or non-profit entities, universities and national libraries.

Funding streams

Federal and international agencies, national libraries, higher education institutions, public and private sector organizations, even individuals—all provided streams of funding for the early development of digital libraries. First-decade digital library funding tended to gravitate to national or local institutional levels, or it was invested as a result of the strategic capital budgeting decisions of commercial firms. The variety of streams has resulted in many technical advances, diverse digital libraries, and a complex landscape.

Large-scale efforts

Large-scale efforts tended to be funded by international bodies, government agencies, foundations, and non-profit organizations. Some libraries invested heavily in digital library programs in keeping with their missions to support historical and cultural studies, provide a national research information infrastructure, and preserve their nations' creative output (examples from Australia, France, the Netherlands, New Zealand, the UK, US and elsewhere are represented in the next chapter). As noted previously, the investments of scholarly societies, publishers, indexing and online services also considerably advanced the early efforts to put

scholarly content online; the amounts invested are unknown but collectively it must have been substantial.

Universities and institutions

It should also be mentioned that the funding from many universities and institutions supporting individual library projects, when taken as a whole, probably surpasses the financing provided by the centrally organized programs. Daniel Greenstein and Suzanne Thorin's report (2002) of a survey conducted by the Digital Library Foundation indicated that in 2000, responding libraries spent an average of over US$1 million each on digital conversion and digital library personnel (see their table 3.1). University library projects at that time focused predominantly on digitization of cultural heritage materials.

Other sources

In a few cases the vision, commitment and financial resources of single individuals produced lasting and influential digital libraries. Brewster Kahle, for example, founded the Internet Archive in 1995, providing the capital himself. In 2003 one journalist reported that "the ten million-dollar annual budget [of the Internet Archive] continues to come primarily out of Kahle's pocket" (Womack 2003). Chapter 2 continues the discussion of how a variety of types of organizations supported the emergence of digital libraries as a new field of endeavor.

The one universal digital library

National agendas have contributed to the sense of urgency that spurred the eventual development of digital libraries in many countries. The dream of one universal, global digital library has been relevant everywhere to some degree, and it still is (see Arms 2005 for a discussion of the user's viewpoint). However, while digital libraries are relevant globally, with few exceptions they have been funded at the regional, national or local level.

Writing for the 2007 issue of the *Annual Review of Information Science and Technology*, David Bearman (2007, 223–4) stated "although the vision of a singular 'Digital Library" was what captured the popular and political imagination, and was promoted especially by Vice President Al Gore in the 1992 election campaign, through the 1990s the United States government supported 'digital libraries' in the plural." Bearman's perspective is supported by a review of the *Source Book on Digital Libraries* (Fox 1993b), which reports on a series of NSF invitational workshops that preceded the NSF's call for the DLI-1 proposals. That work in the foreground of funded projects had two long-lasting outcomes: a preference for "digital library" over "electronic library" and a shift from the goal to "develop a prototype national digital library" (singular) toward funding opportunities for the development of digital librar*ies* (plural). The last chapter of this book returns to a consideration of the dilemma created for digital library implementers as a result of the disunion between who funds digital libraries and who benefits from them.

Early digital-library projects

UK, US and multinational programs had considerable influence on digital library development and they produced significant outcomes that defined the way forward as digital libraries continued to evolve. The key projects included:

- **UK eLib Programme (eLib).** The driving force for the commissioning of eLib was the Follett report (1993), which reviewed the system of UK academic libraries in light of the problems of huge expansion of undergraduate populations; rising costs for library materials; and the opportunities of new forms of information storage, access and retrieval over networks. Recommending that the problems be addressed through the use of information technology, the Follett report was highly influential and released the funding for eLib (Dempsey 2006b). Managed by the Joint Information Systems Committee (JISC), eLib ran for seven years

(1995-2001) and involved 70 projects. For more information see the first feature article in the first issue of *Ariadne*, which itself grew out of eLib (Kirriemuir 1996). Pinfield (2004) offers a detailed review of eLIb's influential outcomes.

- **DLI-1.** The first large-scale funding for digital libraries in the US began in 1994 with an initial four-year Digital Library Initiative (DLI-1) sponsored by NSF, the National Aeronautical and Space Agency (NASA) and DARPA (Defense Advanced Research Projects Agency) (Arms 2000, 62-63; National Science Foundation 1993). The projects emphasized mainly technical aspects of digital libraries (Mischo 2004, 6) and were led for the most part by computer scientists. Behavioral, social and economic issues got little attention during the first round of NSF funding.

- **DLI-2.** In 1998 NSF issued a second call for proposals (National Science Foundation 1998a; Griffin 1999; Mischo 2004). DLI-2 began with more concern for the social, behavioral and economic aspects of digital libraries and attracted funding from multiple agencies including national libraries and the Institute of Museum and Library Services (IMLS).

- **Other US national programs.** Arms (2012) reports on American Memory, a digital library that started in 1995 as a result of the Librarian of Congress' establishment of a project to digitize five million items and make them available on the web within five years. Arms, Bianchi and Overly (1997) discussed the technical building blocks, which came from the National Digital Library Project (NDLP) at the Library of Congress. American Memory is one of the 15 working digital libraries described in table 2.2. The US National Institutes of Health (NIH) engaged early with digital library efforts. In February 2000 they launched the digital library PubMed Central, which as of this writing contains 2.7 million articles. The US

PubMed Central was developed and is managed by the US National Center for Biotechnology Information (NCBI) (Humphreys 2000).

- **Joint NSF/JISC international projects.** In 1998 NSF called for proposals for multi-country, multi-team projects. In the UK, JISC issued a matching call. Six projects were funded jointly by NSF and JISC to explore cross-domain resource discovery, digital archiving, search and retrieval for musical information, reference linking, subject gateways, and metadata for multimedia digital objects (Chowdhury and Chowdhury 2003, 56-7).

- **European Commission (EC).** Even before the first decade of digital library research and practice, the European Commission devoted substantial attention and funding to library-related programs. As Dempsey (2006b) notes, "the first EU call for proposals in the libraries area was as far back as July 1991. The motivating framework for this and later calls was established in the Libraries Action Plan, a document first circulated in 1988." Digital library programs were funded under the European Union's Framework Programmes, beginning with the Third. Funding for digital library research has continued at generous levels (cordis.europa.eu/ist; see also Collier, Ramsden and Zhao 1995; Dempsey 1995; 2006b).

- **Projects in China and India.** A considerable body of digital library research and development has occurred in China and India. Zhou (2005) and Shen and others (2008) describe a number of large-scale digital library projects in China, starting with the introduction of CALIS (Chinese Academic Library Information System) in 1998, followed by CADLIS (Chinese Academic Digital Library) completed in 2005. Kumar (2010) and Das, Sen and Dutta (2010) describe digital library research and development in India, which began early in the new millennium and now includes open repositories, a number of cultural heritage digital libraries and the Digital Library of India.

- **Other projects.** A number of large scale, ambitious projects were inspired by democratic ideals and attracted multiple sources of funding and voluntary support.
    - **Project Gutenberg** (gutenberg.org) is the first and oldest digital library. It began in 1971 as an idea from Michael Hart, who, given free computer time at the University of Illinois, decided to type in the US *Declaration of Independence* and then tried (unsuccessfully) to send it to everyone on the campus network (Hart 1992). Gutenberg's goal has been to provide public domain e-texts a short time after they enter the public domain, for free, using only volunteers and donations to get the work done.

    - **Internet Archive.** Brewster Kahle started the Internet Archive in 1995. The Internet Archive (IA) has numerous components, but the Wayback Machine, which provides access to archived versions of an estimated 220+ million websites, may be the best known. The IA is an advocate for universal and free access to knowledge and it founded a cooperative project called the Open Content Alliance to build and preserve a massive digital library of multilingual digitized text and multimedia content (Dougan 2010).

    - The **Million Books project** (ulib.org; the first project of the Universal Digital Library) began with some preliminary test projects that led to an initial grant from NSF in 2000 (Linke 2003; St. Clair 2008). Raj Reddy, an award-winning computer science professor at Carnegie Mellon University, continues to inspire and direct it. The Universal Digital Library's mission is to foster creativity and free access to all human knowledge; its purpose is to make digital texts freely available to anyone who can read and has access to the network (ulib.org/ULIBAboutUs.htm). Partners came from China, Egypt, India and

the US. It reached and exceeded its goal of a million scanned books in 2007. Collections are represented on mirror sites in China and India.

Definitions of digital libraries

The definition used in this book

The definition of "digital libraries" that underpins this book has two parts. Digital libraries are:

1. A field of research and practice with participants from many disciplines and professions, chiefly the computer, information and library sciences; publishing; the cultural heritage sector; and education.
2. Systems and services, often openly available, that (a) support the advancement of knowledge and culture; (b) contain managed collections of digital content (objects or links to objects, annotations and metadata) intended to serve the needs of defined communities; (c) often use an architecture that first emerged in the computer and information science/library domain and that typically features a repository, mechanisms supporting search and other services, resource identifiers, and user interfaces (human and machine).

My intention is to provide a practical definition that reflects the current situation, but can evolve as digital libraries evolve in the context of the web. Lagoze (2010, 25-31) has persuasively discussed the trend of digital libraries toward the resource-centered architecture of the web (mentioned again in subsequent chapters of this book). The definition used in this book refers for the most part to the traditional repository-centered architecture, because this model remains characteristic of most digital libraries today. Through the chapters of this book, I attempt to make the case that the important characteristics of digital libraries are (in this order) the social roles they play; the communities they serve; the collections they gather for those communities; and the enabling technologies that support them. Social roles and communities are more likely to abide over time; collections and enabling technologies are more likely to shift.

Other definitions of digital libraries

Different perspectives

At the start of digital libraries' first decade, what came to be called a digital library had a number of names—electronic library, virtual library, library without walls. The first decade's explosion of activity and funding for digital library research and practice engendered many diffuse definitions of the phrases *digital library* or *digital libraries.* Some of the principal authors during this first decade paid little heed to definitions; others' discussions of definitions are lengthy. Considered as a whole, the digital library literature contains an enormous amount about how to define digital libraries. Fox and others (1995, 24) suggest an explanation: "the phrase 'digital library' evokes a different impression in each reader."

The public on the one hand, and those involved in building digital libraries naturally had a variety of perspectives on the nature of digital libraries, when they were first conceived. The following list represents a few of these initial perspectives:

- A computerization of traditional libraries (people in general)
- A framework for carrying out the functions of libraries in a new way with new types of information resources (librarians)
- A new set of methods to innovate and improve fee- or membership-based indexing, full-text repositories and hyperlinking systems (publishers, online information services, professional societies, indexing services)
- A distributed text-based information system (computer and information scientists)
- A collection of distributed information services (computer and information scientists)
- A distributed space of interlinked information (computer and information scientists)
- A networked multimedia information system (computer and information scientists)

- A space in which people can collaborate to share and produce new knowledge (those working on collaboration technologies)
- Support for formal and informal teaching and learning (educators)

Arguably the most comprehensive and thoughtful discussion of first decade digital library definitions is by Borgman (1999 and 2000, 35-52), who notes that the many definitions arise because “research and practice in digital libraries are being conducted concurrently” and by individuals and teams from different fields. Borgman made sense of the definitions that had emerged by 2000 by grouping and discussing them in variety of ways including:

- Orientation (research-oriented versus practice-oriented definitions)
- Concept of a library—narrow: library as a collection of content supporting information retrieval—versus broad: library as a continuous and trusted social entity.
- Emphasis (definitions emphasizing collections, a particular type of content or communities versus those with an emphasis on institutions or services)

A sample of definitions

Table 1.1 builds out from the core of definitions considered by Borgman in 1999 and 2000. It offers a sample of definitions, considers their principal facets, and cites their sources. The sample is far from comprehensive but attempts to show the progression of definitions from those emphasizing the enabling technologies of digital libraries (text analysis, distributed retrieval systems, metadata, indexing and knowledge representation, data communication networks, intelligent agents, interface design, multimedia storage, etc.) toward a new generation of definitions that place more emphasis on the communities and social roles of digital libraries. A number of authors have made the point that early research engendered definitions that focused

more on technical issues and less on the broader social context of digital libraries (for example Lagoze 2010, 6).

Table 1.1 A Progression of Digital Library Definitions

Definition	Facets	Source and comments
"The library of the future will be based on electronic data … contain both text and graphics and be widely available via electronic networks. It is likely to be decentralized …"	• Digital data (collections) • Multimedia • Services (widely accessible) • Networked • Distributed • Enabling technologies	Lesk, Fox and McGill 1993, 12, 19-24 This was a white paper for NSF created in 1991. It led to the series of NSF workshops and the first NSF call for proposals. The focus of the definition and white paper was on enabling technologies and maintaining US national competitiveness.
"A service; an architecture … a set of information resources, databases of text, numbers, graphics, sound, video, etc.; a set of tools and capabilities … [with] users … [and] contributors …" Another key assumption: For use on the network	• Services (networked; with tools/capabilities) • Architecture (enabling technologies) • Digital data (collections) • Multimedia • Community-based (users/contributors)	Borgman 1993, 122
"Systems providing a community of users with coherent access to a large, organized repository of information and knowledge… enriched by the capabilities of digital technology … span[ning] both print and digital materials … provid[ing] a coherent view of a very large collection of information …integrat[ing] materials in digital formats … such as multimedia, geospatial data, or numerical datasets … [characterized by] continuity [with] traditional library roles and missions… [and with] many digital repositories … appear[ing] to be a single digital library system …"	• Systems • Community-based • Services (coherence; collected and organized) • Enabling technologies • Distributed, interoperable • Digital and non-digital data (hybrid) • Multimedia • Extension of existing libraries	Lynch and Garcia-Molina 1995

Definition	Facets	Source and comments
"A large collection of the full contents of high use materials including books, journals, course materials, and multimedia learning packages, which can be directly accessed by students and staff" [with personal computers]	• Collection • Digital data (digitized) • Multimedia • Terms and conditions (licensed content)	Zhao and Ramsden 1995 ELINOR project; concerned with digital library development for teaching and learning. Led to insights and progress on copyright and publisher content licensing issues (see Collier et al. 1995).
"Organized collections of digital information. They combine the structuring and gathering of information, which libraries have always done, with the digital representation of information that computers have made possible."	• Services (organized, structured and gathered) • Digital data (collections) • Extension of existing libraries • Computers (enabling technologies)	Lesk 1997, xx, xxii Lesk also stressed the importance of the economics of digital libraries: "We know how to build a digital library … we do not know how to make it economically supportable."
"The definition of the *digital library* will require an understanding of the role and nature of public institutions in a postindustrial society." "A realm of free speech and association as well as an information market place."	• Extensions of existing libraries (but not as collections; rather in their societal roles) • Social (emphasis on social aspects)	Lyman 1996 Emphasizes the social role of libraries offering free and equal access to knowledge and ponders the question of how digital libraries might support the traditional role of the library as a "marketplace of ideas" and the public interest in education and democratic participation.
"Organizations [i.e., institutions] that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities."	• Organizations (institutions) • Digital data (collections) • Community-based • Services (selecting, collecting, organizing, providing access, delivering, *preserving*)	Waters 1998 The definition developed by the Digital Library Federation. Services encompass a curatorial role. See also Deegan and Tanner (2002, 22)

Definition	Facets	Source and comments
1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. 2. Digital libraries are constructed—collected and organized—by [and for] a community of users, and their functional capabilities support the information needs and uses of that community.	• Digital data (collections) • Enabling technologies • Services (collecting, organizing, searching, using information) • Community-based • Use- and user-centered • Emphasis on social aspects (life cycle of information)	Shortened version of Borgman 2000, 42. This definition has been very influential in the digital library field. From the beginning, Borgman has stressed the importance of the social aspects of digital libraries.
"*Sociotechnical systems*—networks of technology, information, documents, people, and practices."	• Systems • Networked • Community-based • Use- and user-centered (work practices and people) • Emphasis on social aspects • Systems • Enabling technologies • Collections	Bishop, Van House and Buttenfield 2003 Emphasis on balancing the needs of people with the requirements for collections and enabling technologies.
"A tool at the center of intellectual activity having no logical, conceptual, physical, temporal, or personal borders or barriers to information. Generally accepted conceptions have shifted from a content-centric system that merely supports the organization and provision of access to particular collections of data and information, to a person-centric system that delivers innovative, evolving, and personalized services to users. Conceptions of the role of Digital Libraries have shifted from static storage and retrieval of information to facilitation of communication, collaboration, and other forms of dynamic interaction … [and] the capabilities of Digital Libraries have evolved from handling mostly centrally located text to synthesizing distributed multimedia document collections, sensor data, mobile information, and pervasive computing services."	• Service (Tool) • Systems • Use- and user-centered • Community-based • Social (communication, collaboration, dynamic interaction) • Multimedia • Mobile • Terms and conditions (policies)	Candela et al. 2007 A conceptual definition from the DELOS Digital Library Manifesto (Candela et al. 2006). Defines six core components of digital libraries: content,, users (both humans and machines), functionality, quality, policy (e.g., rights) and architecture. The Manifesto contains a useful discussion of digital library definitions.

Discussion

The DELOS definition offers a framework for understanding, planning and evaluating for digital libraries. Another model is the "5S" framework (Streams, Structures, Spaces, Scenarios, and Societies) introduced in the dissertation of Marcos André Gonçcalves (2004), which has been used to inform the development of a curriculum for digital library education and for other purposes. Another influential definitional model—one that pushes beyond read-only digital library repositories—is the one proposed by Lagoze and others in 2005. This paper introduced a more flexible, richer information model for digital libraries based on an "information network overlay" for modeling resources, their descriptions and relationships. It represented breakthrough thinking that led to new possibilities for digital libraries that facilitate "the creation of collaborative and contextual knowledge environments."

Other authors have also contributed insightful commentary on how to define digital libraries, rather than specific or formal definitions or frameworks (two examples are Chowdhury and Chowdhury 2003, 4-9; Chowdhury and Foo 2012, 2-4). Bill Arms offers an informal definition ("a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network," but at the same time Arms has consistently emphasized that digital libraries must be understood as an "interplay of people, organizations and technology" (2000, ix, 2). The already cited article by Peter Lyman offers another, quite different perspective; I recommend it to anyone with an interest in libraries' (and digital libraries') roles supporting the public good. The IFLA/UNESCO manifesto on digital libraries (ifla.org/digital-libraries/manifesto), which contains a definition of a digital library, also emphasizes the role of digital libraries in bridging the "digital divide" (discussed in chapter 6).

Levy and Marshall's article (1995, 78, 80, 82-83) is particularly important because it applies a work-oriented (ethnographic) perspective, noting that the emergence of digital libraries

challenges the assumptions but not the basic character of libraries as an interplay of collections, enabling technologies, and services supporting the work that communities of users want to do. Noting "an infrastructure by itself does not constitute a library" and "the highest priority of a library, digital or otherwise, is to service the research needs of its constituents," their article presaged the ensuing shift away from enabling technologies and digital collections as ends in themselves and toward user-centered design and networked services supporting collaborative work.

A few definitional issues

There are many challenges associated with attempting to define digital libraries. Some of the issues discussed by various authors include:

- **Distributed digital libraries:** Some digital libraries are central archives that provide digital content storage and deliver services from a single system; others' content and services are distributed in multiple locations on the network. Still others aggregate the content of many digital libraries (repositories of repositories). Suleman (2012; 17-21) discusses centralized and distributed digital libraries.

  It should be noted here that digital libraries that are crawled and indexed by common or academically-oriented search engines are discoverable in search engine results as if they were aggregated. The definition of digital libraries in this book covers some of these but excludes the virtual aggregations offered by common search engines like Google or Bing. The academic search engine Google Scholar, however, has characteristics of a digital library (it has a social role and it is intended for scholars' use). Beel, Gipp and Wilde (2010, under section 2.1) further discuss academic search engines, including Google Scholar, PubMed and IEEE Explore.

- **Hybrid libraries:** As already noted, Rusbridge coined the term "hybrid library" to refer to combinations of traditional collections, licensed e-resources, and openly available digital collections produced in-house or elsewhere. Some digital content is directly accessible; other content can be linked to; still other content is represented only by citations (metadata). Schwartz (2000) writes "the hybrid library is the context within which most academic digital libraries are found—the ecosystem of the digital library, as it were." Chowdhury and Chowdhury (2003, 6-7) confirm this view; in their book they use the phrase "digital libraries" to denote both digital-only and digital-plus-analog (hybrid) libraries. As Bearman (2007, 223) remarks, an assumption that a digital library contains only digital works is overly limiting; it is necessary to include within the scope of digital libraries those that "service some physical items in addition to digital content." The definition of digital libraries in this book includes hybrid libraries provided that the amount of digital content directly available or accessible through links exceeds the content represented by metadata only. Databases of metadata only fall outside the definition.

- **"Library" or "digital library"?** As library collections are increasingly dominated by online content, the concepts of "digital libraries" and "libraries" are less distinguishable than they were in the 1990s, when digital libraries began to emerge. Chapter 5 discusses the possible convergence of strategic agendas for digital libraries and traditional libraries. However, the definition provided in this book does not conflate digital libraries and libraries.

- **Preservation:** Deegan and Tanner's definition (2002, 22) is a set of principles emphasizing the curatorial role of digital libraries as managed collections, requiring that digital objects be selected, made accessible, and preserved as *long-term, stable resources*. The definition of digital libraries used in this book does not explicitly require a preservation mission.

- **Open or restricted content?** As discussed in chapter 4 of this book, digital library innovations have led to rapid growth in the availability of open, freely available digital content and a culture of open data interchange. The definition used in this book notes that digital libraries are often open. However, the definition does not exclude fee-based or restricted-access digital libraries such as those produced by publishers and other e-resource providers, provided they are intended to serve defined communities. Borgman (2000, 46-47) and Chowdhury and Chowdhury (2003, 8-9) also discuss this issue. The definition in this book includes, for example, open or fee-based digital libraries from scholarly publishers, professional societies, aggregators like JSTOR or the Directory of Open Access Journals. It also includes library or consortially-provided, cloud-based library discovery layers that provide access to a substantial amount of open, licensed and/or fee-based digital content.

- **Global digital library:** Borgman proposed a working definition of a "global digital library" as "a useful construct that encompasses all the digital libraries that are connected to and accessible through a global information infrastructure" (2000, 48). Such a construct does not exist as of this writing. The world wide web, in and of itself, or its representation in a search engine like Google, falls outside the definition of a digital library that is used in this book.

Conclusion

This chapter has traced the antecedents of digital libraries to 1965 and J.C.R. Licklider's challenging vision for "libraries of the future," which, he noted, "may not be very much like present-day libraries." Key developments from 1965 to 1990 in computer and information science, telecommunications and networks, online publishing, personal computer ownership, libraries, archives and other professional communities—not to mention the internet and web—prepared the ground for an ambitious digital library research and development agenda. The

vision for digital libraries was grand, and it attracted top research and professional talent and generous funding.

Early projects in the US and UK, programs funded by the European Commission, scholarly publishing projects, a number of projects inspired by democratic ideals, and many other initiatives led to groundbreaking innovations and the emergence of a new field of endeavor. Multifaceted and surrounded by dynamic technological and societal conditions, digital libraries are challenging to define, because they evoke diffuse impressions and continually evolve. The chapter concludes with a practical definition that underpins the use of the phrase “digital libraries” in this book.

The next chapter examines the outcomes of digital libraries’ exhilarating first decade: a new field of endeavor; transformative change in the processes of scholarly communication and in how (and where) people look for information; new ways of organizing, interlinking, and aggregating digital content; large-scale digitization; digital preservation; the open access movement; and working digital libraries.