3 Key themes and challenges in 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 provides a high-level view of the key themes, current position and challenges of digital libraries and their technologies, social aspects, collections and communities. It begins by identifying the key themes of the second decade (2002-2012) of progress in the diverse, multidisciplinary, international field of digital libraries. A concept map visualizes the results of an analysis of second-decade digital library literature. The map provides new insights into this complex field by exposing thematic connections between technologies, collections, social forces and online community building. The chapter concludes with a consideration of key challenges facing digital libraries: interoperability, community engagement, intellectual property rights, and sustainability.

The key themes of digital library work

Existing research to identify core topics

Jeffrey Pomerantz and colleagues (2006) produced a curriculum for digital library education that was aligned with the “5S framework” for digital libraries discussed in chapter 1 (see also Yang et al. 2009). They validated their selection of curriculum module topics by manually classifying papers from 1996 to 2005 from two sources: (1) 543 papers in the proceedings of two renowned digital library conferences; and (2) 502 articles published in *D-Lib Magazine.* Their analysis revealed concentrations from both sources in digital library services; architecture and

Keywords: Digital libraries—Evaluation; Digital library literature; *D-Lib Magazine*; Concept map; Interoperability; Semantic web; Linked data; Digital libraries—Social aspects; Intellectual property rights; Copyright; Digital library sustainability

interoperability; and metadata. The conference papers revealed an additional concentration on the topic of digital objects. The *D-Lib* papers had additional concentrations around digital library collections, social issues and preservation.

Chern Li Liew (2009) provided a snapshot of the digital library literature from 1997 to 2007, focusing on articles about organizational and people issues. Liew was interested in digital libraries as “socio-technical systems” that support not only information seeking and discovery but also community interaction and collaboration. The analysis drew from 577 articles on socio-technical topics published in peer-reviewed library and information science journals, with some exceptions (e.g., *D-Lib Magazine* is not a refereed journal). The methodology appears to have excluded conference papers. The findings indicated first, a trend toward more articles on socio-technical topics over time and second, the dominance of topics related to digital library use and usability plus organizational, economic and legal issues. Ethical and social/cultural issues were not well represented in the Liew sample articles.

Son Hoang Nguyen and Gobinda Chowdhury (2013) identified core research topics and subtopics and created a “knowledge map” that offers a panoramic view of the digital library field over twenty years. Their work is the most up-to-date and comprehensive analysis of digital library research topics as of this writing. Their detailed analysis could serve multiple purposes: for example to develop an updated digital library curriculum for LIS education.

Nguyen and Chowdhury focused on peer-reviewed publications from 1990 to 2010. Their initial topic list came from knowledgeable experts and from calls for conference papers. They refined this list using a formal knowledge organization approach. They then searched Scopus, a large abstract and citation database of research literature (scopus.com/home.url), for digital library publications and found 7,905 records for conference papers and articles. They then used the records and the *Library and Information Science Abstracts* thesaurus to further refine and standardize the terminology for their core topics and subtopics. The result was 21 core topics and 1,105 subtopics, which they present in a large table and as a series of charts. Three of the 21 core topics—architecture/infrastructure, digital library research and development, and information organization—produced 53% of the publications in the analysis (see their figure 2).

A new concept map

This chapter builds on and extends these prior analyses by focusing on the work done in the second decade of digital library research and practice (2002-2012). My purposes in conducting the analysis included uncovering the key themes and core topics of the field in a way that would (1) suggest the nature of second-decade research and practice and (2) produce a conceptual frame for the rest of this book. The basis of the analysis was a manual evaluation of the roughly 440 full-length feature articles (articles, opinions and commentaries) published in *D-Lib Magazine* between 2002 and 2012. I considered *D-Lib* full-length features only and not its news items, conference reports or briefings.

History and impact of D-Lib

Founded early in the life of digital libraries (1995), *D-LIb Magazine* is freely available on the internet. It has tracked progress across participating disciplines, and its articles include a range of both technical and professional perspectives. The primary intent is “timely and efficient information exchange” (Wilson and Powell 2005). *D-Lib’s* founders and subsequent editors made a deliberate choice of quick turnaround from submission to publication over the long timelines generally associated with publishing peer-reviewed articles.

In a tenth-anniversary feature article on *D-Lib*, Wilson and Powell noted that *D-Lib* articles have been widely cited; in 2005, the average citation rate was 117.5 cites of *D-Lib* articles per year, comparing favorably to the citation rates of journals in the fields of computer science and library and information science. The original funding for *D-Lib Magazine* came from DARPA and the NSF and was related to the DLI initiatives. From 2006 to the time of this writing, *D-Lib* has been produced by the Corporation for National Research Initiatives (CNRI) with assistance from the D-Lib Alliance and other contributors (dlib.org, under "About D-Lib").

Other analyses of D-Lib Magazine

Others have evaluated the contents of *D-Lib Magazine*. Zhang, Mostafa and Tripathy (2002) used the contents of *D-Lib* articles from 1995 to 2002 to test their innovative information retrieval and visualization system, in the process automatically generating a set of concepts associated with these articles. Their process generated 69 concepts, which their system displayed visually in a number of ways (see their figures 1-5). Bollen and others (2005) completed an evaluation of ideas and concepts represented in *D-Lib* articles from 1995 to 2004 through the automatic detection of term co-occurrences. These two analyses used wholly quantitative methods. Park's bibliometric analysis (2010) of *D-Lib* content from 1995 to 2008 produced information about *D-Lib's* impact, authors and the number of citations per article, revealing its wide, global impact on multiple disciplines.

Methodology: evaluating the articles

The analysis of the 440 *D-Lib* articles involved both quantitative and qualitative methods. The first, qualitative steps of the analysis were to manually examine the articles, in the process assigning keywords or keyword phrases to each. Next, a quantitative analysis, using a word frequency macro, counted the occurrences of title words and keywords or keyword phrases. The frequency counts of title words, keywords, and keyword phrases revealed patterns that suggest the comparative strength and evolution of themes in the 11-year span of articles. Table 3.1 summarizes the frequently-occurring keywords or keyword phrases and their ranges of occurrences. A total of 77 keywords and keyword phrases (8.3% of all of the keywords and keyword phrases) occurred eight or more times each and accounted for a little over half (51.8%) of all occurrences of all keywords and keyword phrases in the data set.

Table 3.1. Summary of frequently-occurring keywords or keyword phrases

Range of occurrences	Sample of keywords or keyword phrases
Between 34 and 90 times	Repositories, digital preservation, metadata, evaluation, open access, scholarly communication, OAI PMH, aggregation (this is the complete list not a sample)
Between 27 and 33 times	Discovery, collaboration, standards, social web, federation, data sets, interoperability
Between 21 and 26 times	Education, registries, digitization, NSDL, national libraries, e-journals
Between 15 and 20 times	User studies, identifiers, data exchange, web services, portals, copyright
Between 8 and 14 times	OAIS, multimedia, automated metadata, web archives, web 2.0 and libraries, METS, mass digitization, digital curation, user-centered design, reuse, newspapers, semantic web

The next phase of the analysis was to reflect on the patterns and themes that emerged from the keyword frequency data, develop an understanding of how the themes are connected, and then group related keywords and keyword phrases together (for example, “mass digitization” and “Google Books” were grouped with “digitization”.

Methodology: constructing the map

*T*he construction of the map came next. It involved a qualitative analysis to tease out interrelated themes and decide how to group them together visually. This required choosing the map’s x- and y-axes. The choice of axes was informed by the word frequency counts but not completely determined by them. After carefully reflecting on the patterns in the keyword frequency counts, I labeled the x-axis of the concept map to organize a continuum of themes and topics ranging from “collections” to “communities.” Similarly, the y-axis organizes a continuum of themes ranging from “technology” to “social and economic aspects.”

As I constructed the map, I added a few additional keywords and keyword phrases that occurred fewer than eight times to aid the comprehensibility and completeness of the map. For example “FRBR” (5 occurrences) and “RDA” (4 occurrences) were added to the “cataloging” cluster, and “digital divide” (6 occurrences) was added as a social issue relevant to digital libraries. The last step of constructing the map was to select the themes for the two “key challenges” boxes at the top and bottom of the map.

The result of the evaluation of the articles and the construction of the map is figure 3.1.

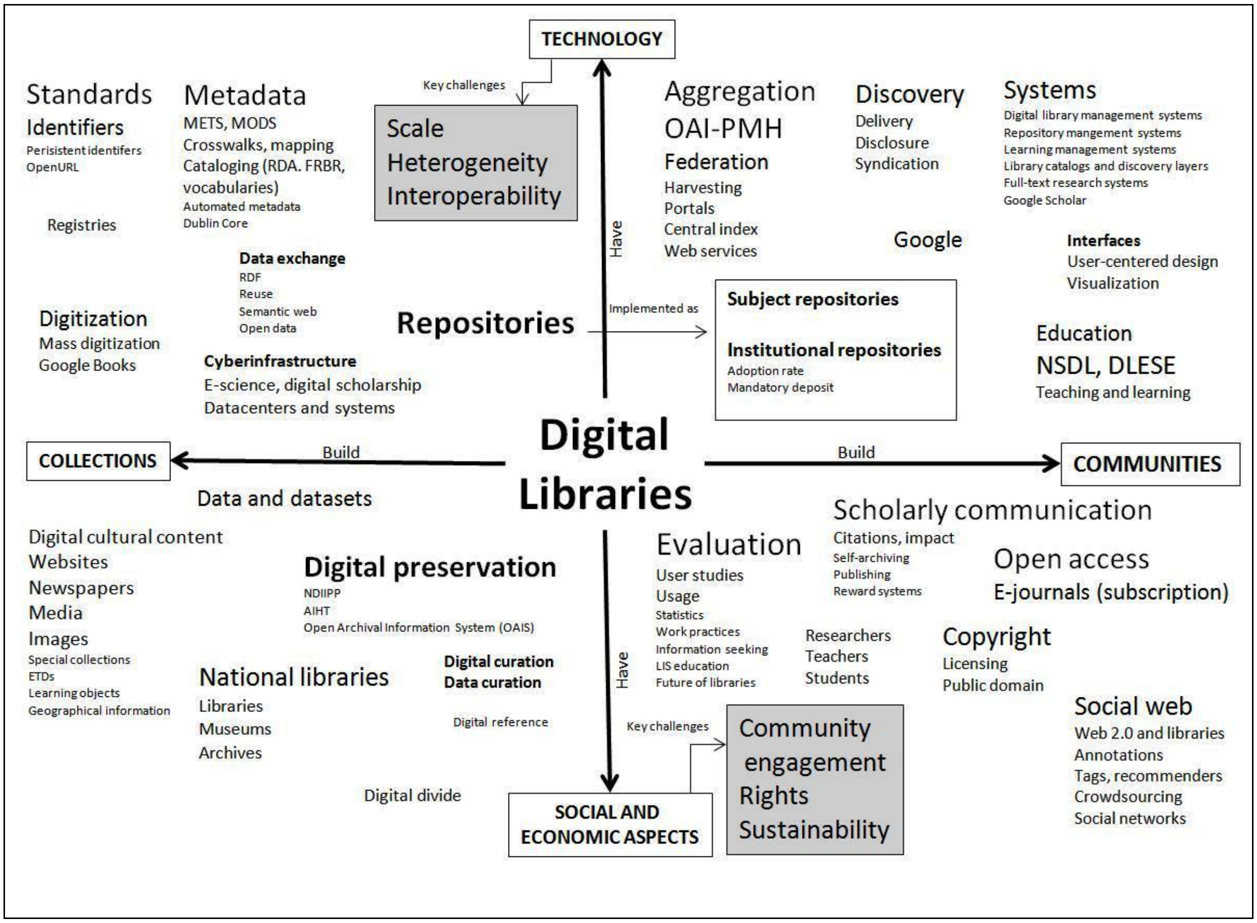

Figure 3.1. Concept Map of Major Themes of the Digital Library Literature (2002-2012)*

*As represented in feature articles in *D-Lib Magazine*

The map

Figure 3.1 places the concept “Digital libraries” at the center, then presents and clusters the results of the word frequency analysis along the x- and y-axes. The font size of the text indicates how frequently the keyword or keyword phrase occurred. This concept map can be said to reveal significant themes in the 11-year span of articles, but not all themes. The overall intent is to organize the decade’s themes and suggest one way to comprehend and explain them as a coherent conceptual whole.

The map suggests the nature and thematic structure of the past decade's digital library research and practice. It represents the principal themes and the relationships between key topics using the map's four cardinal directions and quadrants. The northern hemisphere represents a body of work focused on the enabling technologies of digital libraries and on addressing the field's key technological challenge: interoperability. The southern hemisphere clusters the body of work devoted to the social and economic aspects of digital libraries and to addressing the key challenges of community engagement, intellectual property rights and sustainability. The northwest and northeast quadrants of the map cluster work on the technological aspects of collection and community building, respectively. The southwest quadrant clusters the body of work on the social and economic aspects of digital library collections, while the southeast represents work on the social and economic aspects of building communities around digital libraries.

The remainder of this chapter focuses on the "key challenges" identified at the top and bottom of the map. Before continuing to those sections, however, it is important to write a few words about the limitations of this analysis of the second decade of digital library literature.

Limitations of the analysis

The analysis and concept map provide a snapshot but not a definitive evaluation of the digital library literature from 2002-2012. It focuses on what appeared in *D-Lib* alone, leading to the potential overrepresentation of some topics and the underrepresentation (or omission) of others. A more comprehensive analysis would examine the second-decade literature as represented in other forums, especially peer-reviewed journals and conference proceedings, and in languages other than English.

Key challenges

The next chapters of this book define and expand on the themes from the concept map in the context of the map's x-axis: *building collections* and *building communities.* Prior to those chapters' detailed discussions, this chapter describes and evaluates four key challenges to building collections and communities for digital libraries: (1) interoperability and its facets; (2) community engagement; (3) intellectual property rights; and (4) sustainability.

Key challenge 1: Interoperability

The information landscape can be said to be a highly distributed, heterogeneous one containing many islands of content. Interoperability became increasingly important as more and more content moved online and demand for unified access grew.

Many people perceive that Google—which grew out of Brin and Page's work on PageRank (discussed in chapter 2)—and other general-purpose search engines have now solved the problem of interoperability. For many, whose needs are met by what search engines can achieve through the associative indexing of web documents, this is true. For the communities for whom digital libraries are or would be useful, it is not true. The next section discusses two of the reasons.

Hybrid libraries and the deep web

As described in chapter 1, librarians have sought interoperability for hybrid library content. For the foreseeable future libraries will require ways to bring digital representations of their non-digital collections (printed books and journals, archives, primary sources, images, slides, maps, analog sound recordings, historical government documents, and more) into digital libraries.

Second, there is a part of the web called the “deep,” “hidden” or “invisible” web that is not indexed by Google, Yahoo, etc. This means that searchers who rely only on results from search engines do not see and cannot reach deep web content. Bergman (2001) estimated the deep web is much larger than the “surface” web that search engines crawl—in fact from 400 to 500 times larger. Bergman’s 2001 analysis further suggested that more than half of the deep web content resided in topic-specific databases (most publicly accessible) that weren’t being crawled or indexed by the main search engines. In 2003, Lyman and Varian, using Bergman’s results, estimated that the deep web contained up to 92 thousand terabytes (more than 45 times the estimated contents of all US academic libraries at that time).

Research into deep web extraction has made progress (as described for example by Liu et al. 2010), and search engines are doing a better job finding and indexing deep web content than they were in 2003. However Zillman (2013) estimates that the deep web still contains “in the vicinity of one trillion plus pages of information located through the world wide web in various files and formats that the current search engines on the Internet either cannot find or have difficulty accessing.” Obviously, not all deep web content is of interest to the users of digital libraries, but a substantial portion of it is. Onaifo and Rasmussen (2013) investigated deep web indexing issues from a library perspective and found that (because it is in databases), a good deal of libraries’ content is part of the deep web. They report strategies that libraries and subscription database vendors are taking to structure content in formats that common search engines can index (see chapter 8’s discussion of SEO and ASEO).

The problem of digital library interoperability is not solved

Carl Lagoze defines interoperability in terms of the user’s experience: “providing the user with a seamless experience as they use heterogeneous, distributed information services (discovery, access, browse, etc.)” (2010, 102). Search engines provide a degree of interoperability across the web of documents; a great deal of content can be discovered via the surface web or through the centralized indexes underlying tools like Google Scholar. But not all of it. And so, after twenty years of progress in the field of digital libraries, the challenge of interoperability remains.

The vision of researchers and digital library pioneers was to integrate “tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system” (Lynch and Garcia-Molina 1995, under “Executive Summary section III). They could probably not have anticipated the scale and complexity of the ocean of content to be coherently integrated today. Digital libraries need to scale to a large amount of content; in addition they must be scalable in terms of efficiency and performance. This is made particularly difficult because the content of interest to the communities that digital libraries serve is heterogeneous*,* and so are the systems, software, and formats associated with that content.

Heterogeneity

Dictionary definitions of “heterogeneity” suggest it describes a condition or quality “lacking in uniformity,” “diverse,” and “composed of unrelated or differing elements.” In the field of digital libraries, heterogeneity refers on the one hand to diverse systems, interfaces, and networks; and on the other to the greatly distributed, complex content that digital libraries seek to bring together for easy discovery and use. Besides being widely distributed on the web, content of interest is managed by many different organizations, and the formats of the digital objects are diverse: text, images, audio and video, other multimedia, geographical information, data and so on.

Many digital library experts’ writings devote attention to the topic of interoperability, from the field’s earliest days (see for example Lynch and Garcia-Molina 1995; Paepcke et al. 1998; Arms

2000, chapter 11; Borgman 2000, 212–213; Miller 2000; Tedd and Large 2005, chapter 4; and Lagoze 2010, under "Technologies for interoperability"). Borgman (2000, 213) described interoperability in three dimensions:

- Getting disparate systems to work together in real time
- Enabling software to work on different systems
- Supporting the exchange of content between systems

A list of the aspects of interoperability ranges across user interfaces, naming and identification, formats for content and metadata, network protocols, search and retrieval protocols, authentication and security, and more (Arms 2000, 70–72).

Early work on interoperability: Z39.50

This section extends the discussion of interoperability in chapter 2 to discuss the contribution of an information retrieval protocol and International Standards Organization standard, ISO 23950, known as Z39.50. Z39.50, which pre-dates the web and has been used mainly by libraries, was the basis of early digital library efforts to achieve interoperability of distributed digital content stores. It performs broadcast searching in real time across a range of different information sources stored in different systems. Organizations can also set up their online resources (e.g., catalogs, databases, indexes) as Z39.50 targets—in other words, Z39.50 search services can gather records *from* them.

In some early digital library initiatives, Z39.50 was important for cross-searching and federating results in hybrid libraries; for example, from multiple catalogs, abstracting and indexing databases, and other kinds of resources of interest to libraries. Dempsey, Russell and Kirriemuir (1996) discussed its potential for building distributed information systems in Europe. NDLTD, which is one of the digital libraries described in table 2.2, used it (Fox et al. 1997). Some UK projects funded under eLib or by JISC also relied on Z39.50 (Stubley 1999; Gilby and Dunsire 2004; Gilby 2005). The European Library (theeuropeanlibrary.org; TEL), a portal and cooperative framework for 48 European national libraries and some research libraries, has used Z39.50 in particularly innovative ways and experimented with its descendants SRU/SRW (Woldering 2004; Van Veen and Oldroyd 2004). Chapter 5 contains an extensive discussion of portal projects, like TEL, and the use of Z39.50 as a protocol for metasearch.

Early digital library developers' experience using Z39.50 suggested the protocol's limitations for solving information retrieval and exchange problems outside the library domain. Many systems outside the library space are not Z39.50 compliant, and Z39.50 is a complex protocol, perceived by some developers as costly to implement. As Moen points out in his excellent conference paper on Z39.50 as a resource discovery tool, "It is a standard that addresses important interoperability challenges but does so in a way, perceived as a library way, that may keep it a niche solution rather than as a broader solution to critical problems of networked information retrieval" (Moen 2000). Paepcke and others (2000) noted that there has been a "culture clash between the comprehensive, often complex approach of Z39.50, and the generally light-weight approaches typical in the design of Web related protocols."

Other early work on interoperability

Chapter 2 discusses the Open Archives Initiative, OAI-PMH, reference linking, and the importance of persistent identifiers. All are important outcomes of the first decade of digital library research and practice that continue to support digital library interoperability today.

Syntactic and semantic interoperability

In the same workshop discussed earlier in this chapter, Lynch and Garcia-Molina (1995), identified a continuum of interoperability with "deep semantic interoperability" at one end, “syntactic interoperability” in the middle, and “superficial uniformity” at the other end. The word *semantic* relates to meaning in language or logic; the word “syntactic” relates to the proper arrangements of elements according to a structure and set of rules. Lynch and Garcia-Molina noted that syntactic interoperability can achieve common navigation, query and viewing interfaces as well as other functionality to support a degree of interoperability for digital library users. They saw deep semantic interoperability as holding the promise of enabling searchers to “consistently and coherently” find and use autonomously managed, distributed information objects and services without being troubled by differences in the underlying systems and content.

Syntactic interoperability achieves coherence across systems based on common protocols, metadata formats, and digital object exchange standards. Tedd and Large (2005, chapter 4) may provide the most comprehensive discussion of various aspects of standards and interoperability up to 2005. The best overview of the digital library community’s development of syntactic interoperability may be that of Lagoze (2010, 102–114).

Interoperability and standards

The digital library community’s approach to achieving interoperability has been to define, agree on, and implement standards that ensure open systems and exchangeable data. This chapter applies the term *“*standards” broadly to refer to established technical norms, requirements, specifications, processes or practices that can be officially ratified, proposed or draft, accepted, or recommended by international, regional or national organizations. Alternatively, there are *de facto* standards that are generally accepted and dominant in their communities. The purpose of standards is to support predictable, consistent results and when widely implemented, they benefit the communities that use them and make cooperation and sharing easy and affordable across organizations.

The family of digital library standards

A large array of standards exists for the digital library field. Tedd and Large state that "if standards and interoperability traditionally have been important in libraries, this importance is further emphasized in digital libraries" (2005, 85).

Relatively early on, Bill Arms recognized the potential challenges of an approach to interoperability based on large-scale community adoption of standards. He wrote "an ideal approach would be to develop a comprehensive set of standards that all digital libraries would adopt" (2000, 207–209), but like other implementers, he quickly questioned the practicality of the ideal approach. He proposed a tempered approach to achieving interoperability—one that balances the costs (sometimes quite high) that organizations are willing to incur to implement standards against the degree of interoperability that adopting the standard will achieve. He later wrote "if the cost of adopting a standard is high, it will be adopted only by those organizations that truly value the functionality provided. Conversely, when the cost is low, more organizations will be willing to adopt it, even if the functionality is limited." Chapter 4 continues the discussion of the tension between standards and approaches developed or preferred by digital libraries and the less-constrained, low-barrier methods and simpler standards typically used by the larger web community.

Semantic interoperability

Digital library researchers and practitioners have been quite successful in advancing syntactic interoperability, but until recently, semantic interoperability has seemed to be a "holy grail" (Lagoze et al. 2005). The web environment has considerably matured since the field of digital libraries emerged, and the principles of web architecture are clearer and better defined (see www.w3.org/standards/webarch and the documents linked from it). A number of new developments have made it possible to renew the pursuit of the elusive “holy grail of semantic interoperability.” How to deploy the semantic web and linked data to advance digital libraries is a new grand challenge for the digital library field. The following section, which provides a brief introduction to both, provides the background for this book’s examination of the prospects of the semantic web in digital libraries (chapters 9 and 10).

The semantic web and linked data

The web of data

The first realization of the web has been called a “web of documents” (w3.org/standards/semanticweb). Even though documents are marked up for use on the web, they are mainly intended for people to read, and it has been difficult to extract pieces of information from them in an automated, consistently generalizable way. A new vision of the web, the “semantic web,” has been called a “web of data.” The intent of the semantic web is to automatically bring together and disclose meaningful relationships between related resources stored in different places, as described by Hagedorn and Sattler (2013):

> The problem of the inability of machines to interpret and process information published on web pages caused the development of a web of data, next to the web of documents. The idea is known as the Semantic Web, where links between information are established in a way that machines can understand and interpret.

The semantic web does not replace the web of documents, but it has the potential to enable interoperability at a significantly higher and more useful level. The semantic web is important for many reasons. In the context of the progress of digital libraries:

1. The semantic web revives (and reshapes) the notion of a singular, “universal digital library” that inspired the first digital library builders in the early 1990s. Bizer (2010), who manages DBPedia (dbpedia.org), has envisioned the semantic web as a “single global information

space," with hyperlinks connecting everything. The end goal is being able to query the web as if it was one global database and get back useful results.

2. Semantic web applications offer new functionality and benefits for particular online communities (see chapters 9 and 10).

From an individual's point of view, the semantic web has the potential to greatly facilitate information seeking. Instead of having to search and examine multiple web sites and assemble needed information manually, many questions can be answered in one step. In addition, semantic web applications can disambiguate (identify separate meanings for) names that are the same, like Jerome the saint and Jerome the town in Arizona.

Computer and information scientists and librarians tend to articulate the benefits of semantic web approaches in different ways, but with equal enthusiasm. For example Keller (2011), university librarian at Stanford, explains why semantic web approaches are superior to current approaches to information discovery and access, which lock up pieces of information in silos and fail to comprehensively surface relevant information. Leading computer scientists Bizer, Heath and Berners-Lee (2009, 14) have made the point that semantic web approaches are superior to classic data integration systems as well as newer approaches using machine-to-machine data exchange based on web services, APIs and mashups (discussed in chapter 4).

History of the semantic web

The idea of a semantic web is traced to Tim Berners-Lee, director of the World Wide Web Consortium (WC3), and the person widely acknowledged as the inventor of the web in 1989. In a set of slides from his plenary talk at the first International Conference on the World Wide Web in Geneva (Berners-Lee 1994), he makes the point:

> To a user, [the World Wide Web] has become an exciting world, but there is very little machine-readable information there. The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the anchor. To a computer, then, the web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them.

At that conference, Berners-Lee proposed “adding semantics to the web,” and he and colleagues further elaborated on the idea in a book published in 1999. In 2001, *Scientific American* published Berners-Lee, Hendler and Lassila’s article entitled “the semantic web,” thus bringing the phrase into mainstream usage.

The Resource Description Framework (RDF)

RDF is a standard data model that supports data interchange and reuse on the web; it “allows structured and semi-structured data to be mixed, exposed, and shared across different applications” (w3.org/RDF). RDF uses URIs (discussed next) as unique global identifiers to make simple statements about resources (entities) in terms of their properties and values. To make these statements machine readable and interpretable, RDF uses an XML syntax. There is much to know about RDF; the *RDF Primer* (w3.org/TR/rdf-primer/2004), a W3C recommendation, is a good place to start. Chapter 2 mentions the history of RDF, which dates to 1998 and is associated with the sixth Dublin Core Workshop and, more broadly, the digital library field. It should be noted that not all advocates for the semantic web are advocates of RDF (Miličić 2011 discusses this).

URIs, HTTP and RDF

The semantic web rests on a small number of web standards that serve as its foundation: URIs, HTTP and RDF.

- URIs (Uniform Resource Identifiers) are the technology for naming and addressing resources on the web (w3.org/Addressing); they consist of short strings identifying many types of resources including documents, images, services, etc. A URI can identify an abstract or physical resource (Masinter et al. 2005), and URIs are therefore more generic than URLs (Uniform Resource Locators), which point to web pages. URIs can point to any entity, including real-world objects like people. In the RDF, they both identify a resource and provide the means to express relationships to other resources.
- HTTP (Hypertext Transfer Protocol) is the data transfer protocol used on the web; it is a foundational standard of the web that has been in use since 1990. Expressing URIs using *http://* allows them to be looked up and retrieved on the web.

Linked data

Linked data provides the framework for publishing and consuming semantic web content. Heath and Bizer (2011) have written a highly readable book on linked data. Baker and colleagues (2011) make an important point about it: “linked Data is not about creating a different web, but rather about enhancing the web through the addition of structured data.” Briefly, linked data provides a “set of best practices for publishing and connecting structured data on the web” (Bizer, Heath and Berners-Lee 2009, 1). Linked data relies on structured data (URIs) in RDF format, and the notion of a uniquely identifiable resource that can to be pointed to and retrieved on the web is fundamental. Berners-Lee (2006) introduced four principles for publishing linked data in accordance with the general architecture of the web:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF; SPARQL)
4. Include links to other URIs, so that they can discover more things

Building and using the web of data

Realizing the semantic web has required the creation of two things: a web of data and applications that make use of it. In 2007 the W3C and a number of partners began the Linking Open Data Project to encourage the building, publication and interlinking of open linked data sets (w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData). This active, community-based project began with nine interlinked data sets from DBPedia, GeoNames, MusicBrainz, the US Census, and others (Bizer et al. 2007). The links between datasets demonstrated the enormous potential for applications using semantic web data; for example, a single search result could combine information about a computer scientist represented in DBPedia (an extraction of structured data from Wikipedia) and her publications in the DBLP database (dblp.org; a computer science bibliography).

The number of open, interlinked data sets has grown many times over since then. In September 2011 (the latest count available at the time of this writing), there were 295 data sets containing billions of pieces of information and millions of links. The Linked Open Data (LOD) community has held annual workshops and periodically released a Linked Open Data Cloud diagram (lod-cloud.net/state), which shows continual rapid progress in the production of linked open data by many types of organizations. Currently, linked data sets can be found using the Data Hub (datahub.io), a community-run catalog. In June 2013, the Data Hub contained over 6,500 open data sets; of these 339 were identified as linked data sets. As will be discussed in chapter 9, several national libraries, Europeana, and many other organizations related to the digital library field are taking part in the realization of the semantic web.

Applications for linked data

Linked data is increasingly being used to build innovative applications. Articles by Mendes, Jakob and Bizer (2012), Raimond and others (2012) and Suchanek and Weikum (2013?) discuss the uses that can be made of linked data, which include:

- Providing a knowledgebase for disambiguating names
- Automatically answering natural language questions (e.g., where was John Lennon born?)
- Identifying related entities (e.g., titles that occur in multiple languages)
- Expanding queries (e.g., suggesting other related searches)
- Repurposing and representing web content created for one site in another context

Heath and Bizer's book (2011) provides a section on deployed linked data applications (browsers, search engines and domain-specific applications) at the time of the book's publication. The W3C also hosts a list of use cases (w3.org/2001/sw/sweo/public/UseCases). Two highly visible linked data applications are the Google Knowledge Graph and the linked data applications of the BBC (Programmes, Music, Nature).

BBC applications of linked data

Since 2007 the BBC (British Broadcasting Corporation) has been using a semantic web approach to structuring information about its programs (bbc.co.uk/programmes) so that the data can be easily used in other contexts within the BBC. In effect, the BBC is using its own and other linked data on the web as its web content management system (Raimond et al. 2012). Since implementing BBC Programmes, the BBC has launched BBC Music (bbc.co.uk/music) and BBC Wildlife Finder (bbc.co.uk/nature/wildlife). The BBC creates a web identifier (and an RDF representation of it) for each entity of interest (artist, species, habitat, etc.). BBC Music and Wildlife Finder are underpinned by other linked data sets (Musicbrainz and Wikipedia) plus data from other sources such as the World Wildlife Fund, Zoological Society of London and Animal Diversity Web. The linked data application repurposes the data and places it in a BBC context. For example, the BBC Music page for Bob Dylan brings together BBC features, a biography from Wikipedia, a list of similar artists from the Echo Nest, links to a number of Dylan tracks, information about Dylan's personal relationships, news, blog posts and other links. In addition the BBC makes data feeds available to others, and its editors contribute to MusicBrainz and Wikipedia, rather than internal systems.

The Google Knowledge Graph

The Google Knowledge Graph, introduced in 2012 (Singhal 2012), is based predominantly on Freebase, a large, open linked data set (Bollacker et al. 2008) that was acquired by Google in 2010 (Freebase 2013). Google's Knowledge Graph also uses information from Wikipedia, the CIA World Fact Book, and other non-public sources that it has gathered based on its research into what people search for. Some have criticized Google because (as of this writing) access to the entire Knowledge Graph data set is restricted. Some of the data is from closed data sets with terms that limit public redistribution (see Stewart 2012; Torzec 2012), and the linked data community highly values openness. In May 2013 at Google's annual I/O Conference for developers, Google developer Shawn Simister presented on open APIs for the core of Google's Knowledge Graph, Freebase (Simister 2013).

Progress and prospects for semantic interoperability

Chapters 9 and 10 further discuss the semantic web and linked data with respect to their early adoption in digital libraries. Chapter 10 also discusses schema.org, a collection of schemas (HTML tags) that enable webmasters to provide structured metadata within web pages. The schema.org approach achieves greater visibility in search engines.

While the semantic web is growing rapidly and applications built upon it are beginning to appear, its evolution is still in early stages. Some have pointed out shortcomings and disappointment with progress; Robert Sanderson (2013) describes some of these in his recent presentation to the spring meeting of the Coalition for Networked Information. It must be admitted that at the time of this writing, the prospects for success remain unclear and unpredictable. Implementers have found that publishing and consuming linked data can be complex, and many practical and research challenges remain to be addressed (such as those described by Bizer, Heath and Berners-Lee 2009, 15-22; Miličić 2011).

Key challenge 2: Community engagement

In many ways, this book is about community engagement with digital libraries and the opportunities that digital libraries have to play stronger social roles. The digital library field has made enormous technical progress, and digital libraries have substantially improved the discoverability and accessibility of scholarly and cultural heritage content. But deep engagement with the communities that digital libraries are meant to serve has been uneven. A key challenge of the field is increasing digital libraries' value and engagement with the communities they serve. This book explores various aspects of increasing digital libraries' engagement with their communities, including:

- What sets thriving, long-lived digital libraries apart from those that attract only modest attention or have faded into memory; why some digital libraries have a distinctive impact on their communities, while others are more or less ignored
- The opportunity to embed digital libraries much more effectively in the web's discovery environments
- The opportunities afforded by the social web to greatly increase digital libraries' community engagement

- The opportunity to participate more fully in revitalizing the processes of scholarly communication in ways that will improve network-based scholarly collaboration, the speed of scientific discovery, and the progress and accessibility of knowledge
- The risk of continuing to emphasize the collections and information processes surrounding digital libraries over their societal or community-based roles
- The risk of continuing to conceive of digital libraries as ends in themselves, and not in the context of the enormous amount of online content that is useful to, and trusted by, the communities that digital libraries serve
- The opportunity to re-cast digital libraries in terms of their social roles and in ways that are better aligned with individuals' information needs, preferences and practices

Key challenge 3: Intellectual property rights

The legal framework protecting intellectual property rights has been a key challenge for digital libraries. The digital world is here. With a massive amount of diverse content now online (text, images, audio and video and more), radical changes have occurred in how people and systems communicate, create, interact with, exchange and re-use, and link to content. This high-speed, dynamic, participatory online information environment benefits greatly from open systems and the easy, low-barrier sharing and exchange of digital content.

It remains essential to balance the values of openness with carefully protecting intellectual property rights. At the same time, the more open digital libraries are, the greater potential they have to play important social roles—for example helping people everywhere gain access to high-quality content, benefit from the free flow of ideas, learn and make new discoveries, and advance knowledge and culture. The social roles of libraries have historically been supported by the legal framework. In many countries this takes the form of the principles of "fair dealing", and in the US by the historic concept of "fair use" and related exceptions and limitations of copyright. The following sections elaborate on these concepts.

In many countries, the legal framework protecting intellectual property is out of step with the new conditions of the digital world. The effect of not upgrading the copyright laws to reflect these new conditions has led to a poor climate for innovation, diminished public access and the limitation of some former provisions permitting the use, dissemination or long-term preservation of content by libraries. This section provides some basic information about the current situation and briefly lays out some ways in which digital libraries are responding.

Definition of intellectual property (IP)

The UK Intellectual Property Office (ipo.gov.uk) introduces the concept of intellectual property in this way:

> **"**Intellectual Property (IP) results from the expression of an idea. So IP might be a brand, an invention, a design, a song or another intellectual creation. IP can be owned, bought and sold."

The UK IPO office goes on to define four main methods for legally protecting intellectual property: patents, trademarks, designs and copyright. The IP protection most relevant in the digital library domain is copyright, which is generally "an automatic right which applies when the work is fixed, that is written or recorded in some way." Most copyright systems require both this aspect ("fixity") in addition to "originality" (an original work fixed in a tangible medium of expression).

Copyright

Copyright in the US is based on the Constitution, article 1, section 8. The pertinent part of the section, which lays out the powers of Congress to enact laws, reads:

> "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."

Copyright prevents unauthorized copying of creators' work, performing it publicly, or developing derivative works of it. In the US, the objective is to grant creators exclusive rights for a period of time, after which their work enters the public domain (that is, copyright ends, though some rights may still apply). The purpose is to stimulate the creation and distribution of new original works, thereby benefiting the public. The incentive for creating new works is achieved by enabling creators to economically benefit from their works for a period of time. The purpose of copyright may differ in other countries; for example, to protect the moral rights of creators to control the use of their works. Hirtle, Hudson and Kenyon (2009) offer a comprehensive guide to US copyright law and its application to libraries, archives and museums. Cornish (2009) has provided guidance for UK libraries, archives and information services.

Adapting to the digital age

Copyright periods have been getting longer, and the copyright laws are not at this time well-adapted to the digital era. There are deep concerns that the public domain is shrinking (Lessig 2013). In the US, libraries have traditionally had certain exemptions as well as provisions for the "fair use" of non-digital copyrighted works, including copying them without permission of the copyright owner and without the payment of a licensing fee (Hirtle, Hudson and Kenyon 2009, chapters 5 and 6). Attempts to adapt these permissions for the digital age have been only partially successful (for a US perspective, current as of this writing, see Brown 2013).

Recent developments in the UK are encouraging. In 2010 the government commissioned the Hargreaves Review of IP and Growth, which produced recommendations for an IP framework better suited to "supporting innovation and promoting economic growth in the digital age"

(Hargreaves 2011, 7). The report called out copyright law as being particularly in need of revision, especially to enable appropriate copying activities and certain types of text and data mining. The government broadly accepted the Hargreaves report's recommendations in August 2011 (ipo.gov.uk/types/hargreaves.htm) and efforts to implement them have begun.

Barriers for digital libraries

Digital libraries are by their nature open and they make the display, use and exchange of content easy. Furthermore, once content is digital and networked, many new innovations and applications become possible. The current state of copyright law—and its misalignment with the realities and opportunities of the digital era—has had significant impact on how digital libraries have developed. The influence of the current legal framework on digital library work is pervasive, affecting for example how digital content and metadata can be re-used and exchanged; what is lawful to digitize or preserve; national legal deposit programs; the process and costs of licensing scholarly digital libraries, e-journals and e-books; and the complexity and costs of digital library development and implementation generally. The difficulties have driven digital library development in particular directions; table 3.2 provides a brief guide to a number of the issues and some sources for further information.

Table 3.2 Issues associated with digital libraries and copyright

Issue	Description	Suggested Sources
The public domain and orphan works	*Orphan works* are copyrighted works whose owner is unknown or cannot be located. Certain categories of works whose copyrights have not been renewed are in the public domain. The problem is, it can be difficult and costly to determine whether copyright exists or has been renewed. The number of orphan works in vast. Orphan works are a problem for digitization and digital preservation projects because it is not clear what is lawful to do with them, and the rightsholder cannot be located to ask permission. Most large-scale projects do not have adequate funds to thoroughly investigate the copyright status of orphan works.	JISC 2009, *In From the Cold.* Hirtle, Hudson and Kenyon 2009, chapter 7. European Commission 2012. Directive 2012/28/EU.
Mass digitization	Mass digitization generally alludes to the digitization of very large, whole collections of content, with no or minimal selection. Google Books is a mass digitization project of books, both copyrighted and public domain. Mass digitization projects can be done for other types of material under copyright besides texts (e.g., photographs). Library, museum and archive collections contain a massive number of orphan works. Both the Google Books project and Hathi Trust have encountered legal difficulties. The European Commission announced a multi-national mass digitization initiative in 2005, which has since been supported by research, development (e.g., the ARROW project), and implementation, as well as legal changes.	• Aaron 2012 (Hathi Trust) • Baksik 2006 (Google Books lawsuit) • De La Durantaye 2010 (European Union initiatives and Google Books) • Hahn 2008 (preservation & Google Books) • Jockers, Sag and Schultz 2012 (Hathi Trust) • Ricolfi et al. 2008 (EU i2010 Digital Libraries) • Stratton 2011 (ARROW) • Travis 2010 (Google Books) • Van Houweling 2012 (photos) • Chapter 5 of this book

Issue	Description	Suggested Sources
Digital preservation	Digital preservation, which is important to economic growth and cultural memory, is the active management of digital information to ensure it remains accessible over time. Both digitized and born-digital content can be preserved. The amount of digital content that could be preserved is staggering. Traditionally, libraries have had the responsibility and legal rights to preserve the physical collections they own. National libraries have had the responsibility to collect and preserve publications through legal deposit programs. With the rise of massive online networked information, the responsibility and rights to preserve digital content have become diffuse and unclear.	• Saarti and Vattulainen 2013 (legal deposit) • Waters 2007 (preserving the scholarly record) • Preservation section of chapter 6 of this book
Scholarly communications and open access	The open access movement is having an impact. The number of publishers requiring authors to transfer copyright is declining. Open access advocates have created “author addenda” to publisher agreements and other means to help authors retain rights to their work, for example to make the work freely available under certain conditions. Scholars themselves exhibit a range of concern, confusion or indifference about copyright. Around the world, and especially in Europe, there is a policy shift toward open access to scholarly content produced as a result of public funding.	Policy and legal frameworks section of chapter 8 of this book Hirtle 2006 (author addenda) ALPSP 2004 (recommended publisher practices)

Issue	Description	Suggested Sources
A new library specialization, new enabling technologies	Instead of purchasing scholarly content as they did in the past, libraries now license access to its digital forms. Publishers restrict the rights to access, display and export most online scholarly content. These conditions have engendered a new field of library specialization and new enabling technologies to support licensing, renewals and payments; authenticating and authorizing access to licensed scholarly content; e-resource metadata management in catalogs and knowledge bases and on library web sites; new systems and services for indexing, end-user discovery and use; and more.	Chapter 5 of this book (hybrid libraries)
New ways to lawfully distribute digital content and data	The web, online information services and digital libraries are driving a shift to discovery and access models that rely on exchanging and linking digital content and metadata. Digital library implementers and others began searching for new ways to incorporate freer copying, distribution and re-use of content, while minimizing the potential for negative outcomes. Models supporting new lawful ways to distribute and exchange content and data have appeared. These new models (e.g., Creative Commons licensing) reserve a range of rights ("some rights reserved") or explicitly dedicate the content or data to the public domain ("no rights reserved").	Creative Commons (creativecommons.org) GNU Free Documentation License (gnu.org/licenses/licenses.html#FDL) Open Data Commons (opendatacommons.org)

Key challenge 4: Sustainability

The digital library field's knowledge of how to build digital libraries outpaces its understanding of how to sustain them. While the digital library builders aspire to offer free access to all, digital libraries are not free for their builders to create and maintain. These costs must be recovered somehow. Financial sustainability is critical.

Digital library sustainability has several aspects. Setting aside the technological aspects of sustainability for the moment, sustainability in digital libraries has economic, social and ethical characteristics (see figure 3.2). Consider the following brief overview of these:

- **Economic**: A sustainable digital library has ongoing funding and a workable business model for recovering its costs; its managers engage in ongoing business planning; it regularly gauges community needs, awareness and satisfaction with its services; it has clear accountability and evidence-based metrics to underpin strategic plans and investments in ongoing development
- **Social:** A sustainable digital library is considered essential by the communities it serves (Hamilton 2004, 393); it maintains its visibility and community awareness; it provides ongoing access to content and services that are highly valued by the communities it serves
- **Ethical:** A sustainable digital library provides the broadest possible access to its content, and it supports open inquiry and the free flow of ideas while respecting the rights of content creators and producers

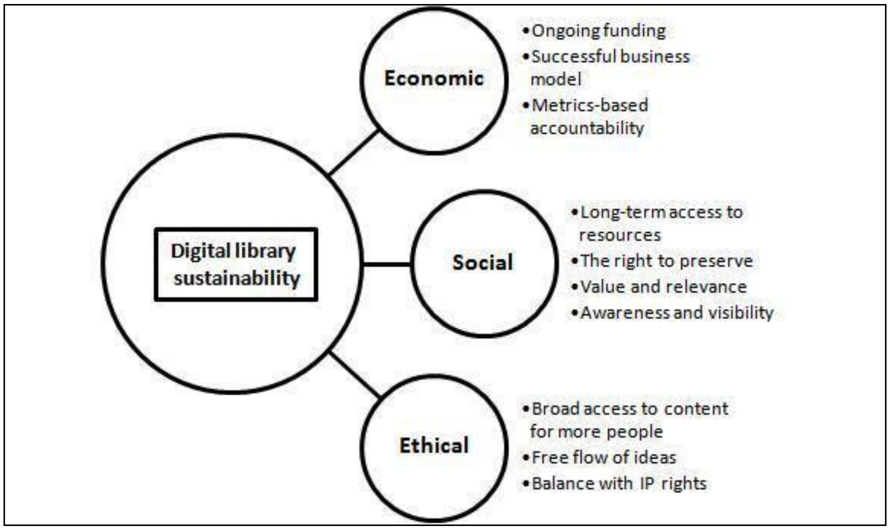
Digital library sustainability
Economic
•Ongoing funding
•Successful business model
•Metrics-based accountability
Social
•Long-term access to resources
•The right to preserve
•Value and relevance
•Awareness and visibility
Ethical
•Broad access to content for more people
•Free flow of ideas
•Balance with IP rights

Figure 3.2 Aspects of digital library sustainability

Achieving and maintaining digital library sustainability is a challenge in a world dominated by the web, dynamic information-seeking expectations and practices, and transformed processes for scholarly communication. From an economic perspective, while funding is sometimes available for getting a digital library started, ongoing financial support can be difficult to find. Chapter 7 of this book provides further analysis of the current situation and the characteristics of sustainable digital libraries.

Conclusion

This chapter offers a new concept map derived from a qualitative and quantitative analysis of the digital library literature from 2002 to 2012. The map provides one framework for comprehending a large and diverse body of work as a set of interrelated key topics and challenges.

The second decade of digital library research and practice carried forward the first decade's emphasis on enabling technologies and on building collections. Three main areas of focus were building and aggregating repositories; technologies and models for digital preservation; and metadata. The key technological challenges continued to be scale, heterogeneity and interoperability; but over the course of the decade, the standards, processes and methods for achieving interoperability changed. Interest in the semantic web and linked data increased strongly from about 2007 forward. While the longer-term prospects for semantic web approaches remain unclear, impressive applications using these approaches have begun to demonstrate their potential. The long section on interoperability is intended to give readers a basic foundation for understanding subsequent chapters and other digital library literature.

Digital library research and practice evolved over the second decade, resulting in greater attention to social and economic issues, especially with respect to evaluating the use and users of digital libraries; advancing education and the processes of scholarly communication; and broadening access to high-quality digital content through open access. The continuing focus on digital collections is now paired with a new body of work that focuses on digital library communities. The second decade began to address key challenges related to engaging communities around digital libraries, coping with the barriers associated with a restrictive legal framework, and identifying success factors for sustaining digital libraries.