4 Digital Library Collections: Repositories

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 and the next discuss digital libraries and the web through the lens of collections and collection building. This chapter begins with an exploration of the parallel but separate developments of the web, digital library repositories, and hybrid libraries. It then turns to an examination of digital library repositories. Topics include numbers, usage, and discoverability of repositories; current position and roles; systems and software; federation and dissemination of repository content; next generation repository systems; and cyberinfrastructure, data and e-research support. The next chapter moves on to the examination of hybrid libraries, then concludes with thoughts about advances, opportunities and challenges for both hybrid libraries and repositories.

The traditional library worldview

Over the course of the 19th and 20th centuries, conventional definitions of libraries and core assumptions of the general public have tended to emphasize their *collections* over their social roles. Services are sometimes mentioned, but the core assumptions are that libraries are collections of things (especially books) in fixed locations (buildings and later, online "virtual" collections or repositories), and the role of libraries is to provide access and support for these collections on behalf of the communities they serve. In keeping with this set of core assumptions, library roles and services have tended to be defined through the collections lens:

Keywords: Subject repositories; Institutional repositories; Hybrid libraries; Web search engines; Discovery systems; Open access; Scholarly communication; Google and Google Scholar; OAI-PMH; Web services and APIs; Object Reuse and Exchange (ORE)

selecting, acquiring, organizing, preserving, managing, providing for access, answering questions and providing instruction about how to use collections. The result is that collections take center stage and dominate the current library worldview, or overall perspective from which people (including librarians) define libraries.

Carl Lagoze discusses the notion of library core assumptions as a “meme” (a worldview) that has influenced digital library development. He contrasts the library meme with the web meme (2010, 48–71) and goes on to argue that over time, the library meme engendered digital library technical approaches and standards that did not play out well on the web as both the web and digital libraries evolved. Indeed, large-scale success for digital library researchers and practitioners has often depended on others being willing to adopt or accommodate digital library ways of doing things (as opposed to the ways things are typically done on the web). When digital library ways contrasted significantly with the less-constrained, low-barrier methods and simpler standards used by most web developers, digital library approaches were not widely adopted outside the digital library community.

Lagoze’s insights into the library meme provide a partial explanation of how digital library repositories and hybrid library systems evolved separately from the web. Another explanation derives from the fact that at first, the simpler web-based approaches yielded inferior results (for example, early search engines were imprecise). These practical realities, combined with mindsets and backgrounds of early designers, does appear to have led to digital repositories modeled on library collections. This had implications later for how repositories evolved and how well or poorly they have been integrated into the larger web.

In the case of hybrid libraries, as much as core assumptions or worldview perhaps, the practical realities of supporting search and retrieval for both digital and non-digital collections produced

special requirements and constrained libraries' options. Libraries manage production systems that their communities rely on daily, so it has been necessary to adapt and evolve library systems and migrate them as new possibilities emerged for production-ready environments. At the same time library leaders knew it was imperative to address dramatic shifts in their communities' requirements. Consider the following:

- Until relatively recently, publishers, professional societies, and indexing services did not allow their metadata to be harvested or crawled for inclusion in other systems. To provide for discovery and access to this content, libraries have needed to use a series of approaches that were compatible with what e-content providers and others made possible, what they could accomplish on their own, and what evolving library information systems could do.
- Most hybrid libraries' systems have been and continue to be dominated by bibliographic metadata describing non-digital (e.g., print) library collections. Libraries have understandably required systems that provide for reliable discovery and use of both online and non-digital content. This added complexity to developing and implementing systems.
- Open access repositories created new requirements for integration with library discovery environments.
- Digitization projects created new content that libraries wanted to reveal in their discovery systems.
- The need for frameworks to enable long-term preservation of content also brought special requirements.
- It became important for library systems to support not only the discovery and delivery of information for its communities, but also to be open to the exchange and disclosure of system contents to many other systems.

These factors have driven a number of related but essentially separate lines of systems development for digital repositories and hybrid libraries, described in this and the next chapter.

Repositories, libraries and the web

Repository architecture

A review of dictionary definitions of the word "repository" indicates it is a place or container where things can be deposited for storage or safekeeping. Branin (2007) has provided a conceptual introduction to repositories from the library perspective. In digital library research and development, a repository has been a fundamental component of the Kahn-Wilensky architecture (introduced in chapter 1). Kahn and Wilensky wrote their seminal paper in the early days of the web, at a time when the web was evolving in parallel with digital library research and practice. They were part of a community of computer science researchers leading digital library research initiatives at the time. Kahn and Wilensky's framework for building digital library collections grew out of their disciplines and was centered on digital objects and repositories.

Web crawling and indexing

The approach of Brin and Page (graduate students who worked on one of the DLI projects until they founded Google in 1998) was a major innovation that grew out of—and later defined—the parameters for the parallel universe in which the web was evolving. Their approach took advantage of the structure of hypertext (text with links) to crawl and index everything they could on the web. They did not begin with the notions of multiple repositories serving as collection points for particular kinds of well-defined objects; their starting point was exploiting information already present in hypertext to build a very large-scale, universal information indexing and retrieval system. The Google repository component developed by Brin and Page contained not well-described and managed digital objects, but compressed web pages that had been crawled with robots. The component they called the indexer then read and processed the repository's

compressed pages to produce a number of outputs, resulting in an inverted index and PageRanks that supported user queries (Brin and Page 1998, see figure 1 and accompanying text).

The approach, which was completely automated, enabled very large stores of documents (the web) to be automatically crawled, indexed and searched at little cost and at great scale compared to other more formal approaches to information description and retrieval. Google's success kicked off an immense wave of further innovation and development on the web.

A multiverse of research and practice

A "multiverse" is a set of parallel universes, each of which defines self-contained but co-existing realities. At the moment a multiverse of research and practice exists for organizing the world's information and making it discoverable. The evolution of the web played out in one parallel universe, which was populated with developers who were entrepreneurial, decentralized, relatively unbound by legacy systems or core assumptions, and minimally constrained by technical standards. It might be argued that this universe is now the only one that is highly visible to the information-seeking public.

Arms, in an essay on the tenth anniversary of *DLib Magazine,* recognized the separate evolutionary paths chosen by two other universes of research and practice when he noted "computer scientists resisted the simple technology of the web; librarians disparaged the value of web search engines" (2005). Given their different starting points, histories and worldviews, the services and systems of the web, digital libraries, and libraries evolved at the same time but along largely separate paths.

While there are some trends suggesting eventual convergence, the parallel but separate realities of digital libraries and libraries persist. At least in US libraries, responsibility for digital

library efforts has been widely distributed or organizationally isolated, and remains so (Maron and Pickle 2013, 2-3). As already discussed, first-decade digital library projects in the US were most often managed out of computer or information science departments, and in the UK by librarians working on federally funded research and prototyping projects. What all digital library specialists had in common was their project-based work, which was generally undertaken outside the mainstream operations of libraries. Meanwhile, librarians had their own challenges adapting their collections, practices and systems to disruptive, constantly shifting requirements for the library's mainstream services and systems. Figure 4.1 offers a side-by-side view of three timelines that trace key events along these parallel paths. Subsequent sections of this chapter and the next describe these parallel but separate paths.

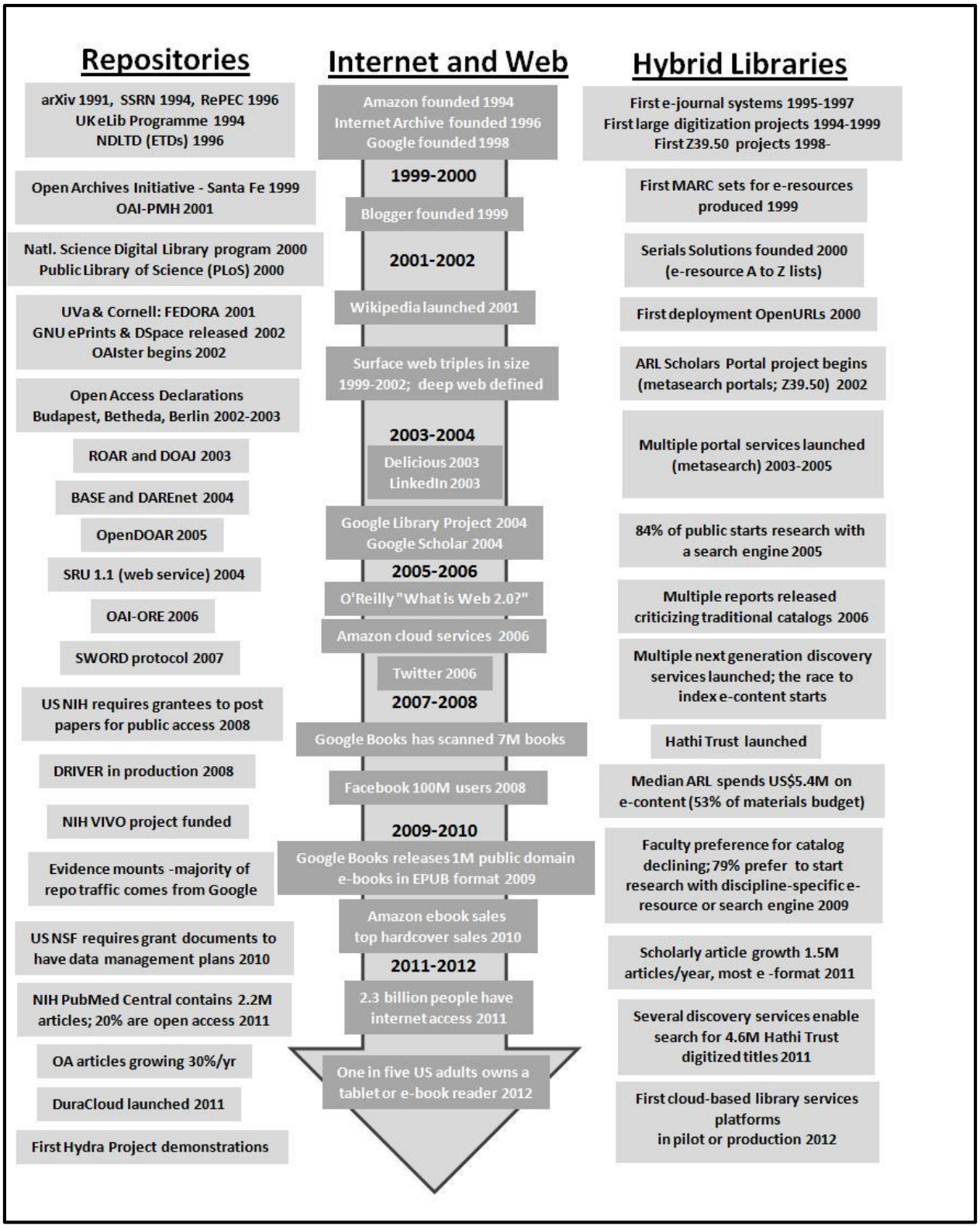

Figure 4.1 Timelines for Repositories, the Internet and Web, and Hybrid Libraries

The evolution of digital library repositories

A key theme and focus for development

In the second decade of digital libraries, the topic of repositories dominated the attention of digital library researchers and practitioners more than any other. They were a key outcome of the first decade of digital library research and practice. There are two kinds: (1) subject-based repositories and (2) institutional repositories. Subject-based repositories collect and provide access to digital objects related to a subject or group of subjects; they are sometimes called discipline-based repositories. The International Network for the Availability of Scientific Publications (INASP) defines an institutional repository as "an online locus for collecting, preserving, and disseminating, in digital form, the intellectual output of an institution." The INASP site lists a number of registries of institutional repositories (inasp.info).

Emergence, numbers, costs

Emergence of repositories

Adamick and Reznik-Zellen's analyses of subject repositories (2010a; 2010b) indicate that four of the world's five highest-ranked subject-based repositories today—PubMed Central, RePEc, arXiv, and Social Science Research Network (SSRN)—were established in the first decade of digital library work (repositories.webometrics.info/toprep.asp, July 2012). As discussed in chapter 2, the international OpCit (Open Citation) project produced GNU EPrints, which was the first open source software available for building repositories. EPrints fueled a movement to build repositories at the *institutional* level.

The OAI framework

As discussed in chapter 2, the Open Archives Initiative (OAI), which emerged from a meeting in Santa Fe in October 1999 to explore cooperation among already existing repositories, has sought to facilitate the distribution, sharing and discovery of scholarly research. The OAI's

framework for supporting open scholarly communications, OAI-PMH, has facilitated scholarly collaboration and publication and a gradual shift away from increasingly less affordable models based on commercially published journals. The framers of OAI-PMH offered a fresh approach to building open systems that were capable of drawing together diverse resources in different locations into a single information service. OAI-PMH has contributed more than any other first-decade digital library innovation to the rapid growth of open access repositories around the world.

Growth of repositories

Other providers of open source software for building institutional repositories followed the release of GNU EPrints; specifics follow later in this chapter. Early adopters identified the necessary process and accompanying tools and standards for launching a repository. Once the process and infrastructure components had been identified, many organizations did launch repositories, either built locally or hosted by a third party. A number of forums and services exist to support implementers of repositories based on open source software. Some provide quantitative information, registry and tracking of repositories, guidance on management and policy matters, and other services. The Ranking Web of World Repositories (repositories.webometrics.info), a service of the Cybermetrics Lab, provides analysis and ranking of repositories.

Various directories support repositories; in their paper at IFLA, Oliver and Swain reported their research, which had identified 23 directories of open access repositories (2006). The two leading ones are:

- ROAR (Registry of Open Access Repositories; roar.eprints.org)
- OpenDOAR (Directory of Open Access Repositories; opendoar.org)

OpenDOAR and ROAR exist mainly to serve as a focal point for quantitative and statistical analysis and/or policy and standards development for the repository community. They also can be used as search portals to aggregated repository contents, although this is a secondary purpose. A report of a JISC-funded project on joint development of ROAR and OpenDOAR (Millington and Hubbard 2010), explores a number of opportunities for further cooperation between the two.

As of this writing, 3,429 repositories are registered in ROAR; 2,322 are registered in OpenDOAR. The Cybermetrics Lab's ranking site currently tracks 1,654 repositories. These registries are growing all the time, so these numbers will not be accurate for long. Using data from ROAR and OpenDOAR, Repository66 (maps.repository66.org) maps 2,311 repositories holding nearly 34 million items (Lewis 2012). BASE, another large aggregation that provides search services across open access repositories, reports access to 40 million documents from 2,400 sources that are harvested for BASE (base-search.net/about/en). These totals count both subject-based and institutional repositories plus other content indexed by BASE.

Not all repositories are registered, so the total numbers and holdings of repositories worldwide are unknown. As an indication of the difference between what the registries track and what exists, the example of DSpace may provide a guess. Currently available statistics from their web sites suggest that from two-thirds to 90% of DSpace repositories are registered in OpenDOAR and/or ROAR.

Costs

While open source software typically incurs no fees, building and running a repository is not free. A local installation incurs costs for hardware and labor among other costs. Estimates of the costs of implementing and running a repository vary widely; the OASIS site suggests labor costs

of 1.5 to 3.0 FTE for set-up and 0.5 to 2.5 FTE for ongoing operations (OASIS 2009). Grant funding supported the initial development of at least three of the top open source products—EPrints, DSpace, and Fedora—but the start-up and ongoing costs of those implementing these products at individual institutions are generally funded from an organization's operating budget. Burns and others (2013) offer the most recent and comprehensive look at institutional repository costs, usage, and value as of this writing.

Current position and roles

Open access movement

The early digital library projects that enabled new repositories were important contributors to the international open access movement. Chapter 8 discusses the open access movement with a focus on its economic and social aspects.

Improving the discoverability and accessibility of scholarly information

Open access repositories have become increasingly important discovery and delivery mechanisms for the scholarly literature. At the time they first began being implemented, many if not most repositories could be crawled and indexed by search engines; most of the data stores of publishers and libraries could not. This made the research output available in open access repositories much more visible and publicly accessible. Even after publisher content began to be indexed by search engines, open access repositories have provided alternative access to copies of articles that are otherwise available only through purchase or subscribing libraries.

Over time, some repository managers have improved their abilities to configure their sites to optimize results for crawlers (see for example Suber 2005). More recently, Arlitsch and O'Brien (2012) tested and reported on techniques for making repository contents more visible in search engines. They were able to identify and quantify ways to significantly improve repository

indexing ratios in Google Scholar. These authors' research suggests that in 2012 the current indexing ratio of repositories (the number of URLs found in a search engine's index divided by the total number of URLs in a repository) may have averaged around 30%. Their tests indicate that a repository manager can improve a repository's indexing ratio in Google Scholar by adhering to Google Scholar's "Inclusion Guidelines for Webmasters" (scholar.google.com/intl/en/scholar/inclusion.html). Following the publication of this influential article, Arltisch and O'Brien produced a book (2013) to provide additional guidance to repository managers on SEO (search engine optimization; practices by website owners to maximize the visibility of their content in search engine results). Chapters 8 and 10 further discusses SEO and other methods for extending repository reach and visibility.

Despite room for improvement, Google and Google Scholar already surface a good deal of content in open access repositories. Norris, Oppenheim and Rowland (2008) evaluated four search tools' utility for locating open access articles: Google, Google Scholar, OpenDOAR and OAIster. They searched a sample of a little more than 2,500 articles from ecology, economics and sociology journals and found open access versions of 967 of them (38%). They then searched the titles of the open access articles using the different search tools. Google and Google Scholar performed the best for finding the open access articles, locating 76.84% of them. After discussing a number of reasons for the results, they concluded that those searching for open access articles are more likely to find them with Google or Google Scholar.

Centralized, easier access to previously hard-to-find content

Repositories also create workspace and centralized access for content that had previously been more scattered and difficult to find. They typically contain not only articles but pre- and post-prints, reports, theses and dissertations, conference and working papers, teaching materials, and presentations. Figure 4.2 provides a breakdown of the types of content that the repositories

registered in OpenDOAR contained as of June 2013. The figure displays results for two subsets of OpenDOAR's registered repositories: subject-based repositories and institutional repositories. While it is clear that institutional repositories are more likely to contain theses and dissertations, and subject repositories are more likely to contain media and audiovisual materials, otherwise both kinds of repositories appear to hold similar types of content. From the perspective of the *number* of scholarly objects they contain, subject-based repositories do tend to contain more items than institutional repositories.

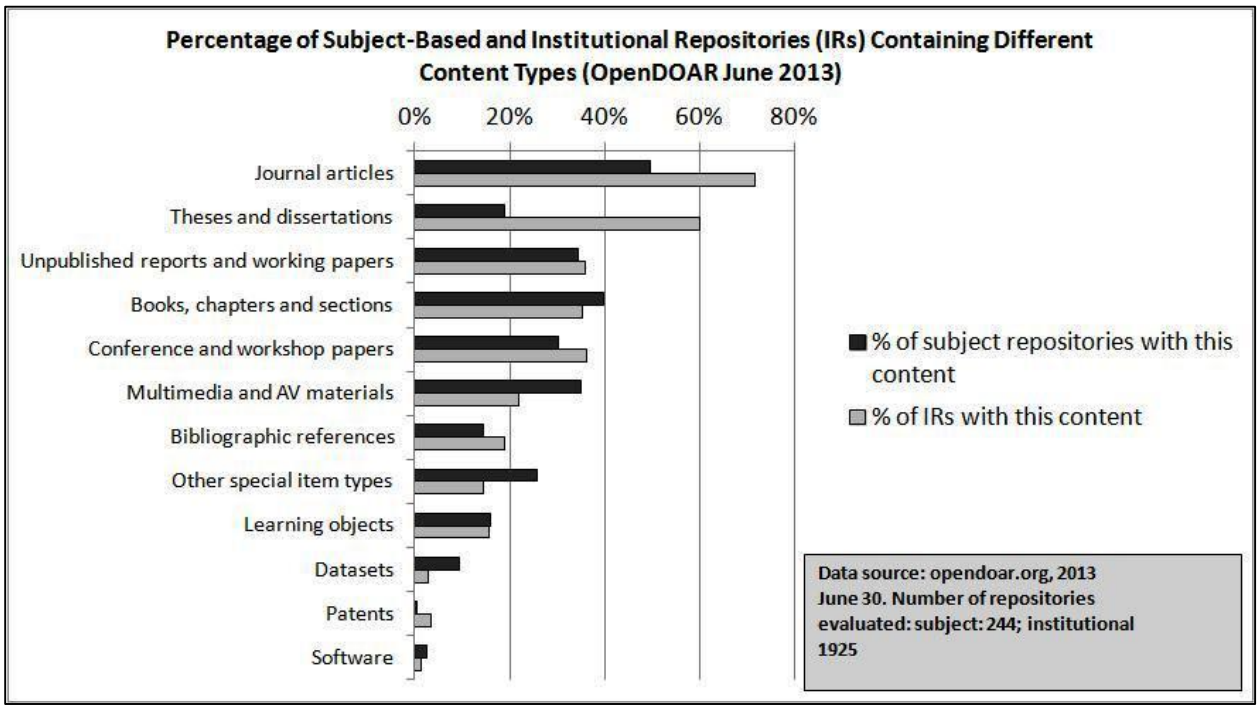

Figure 4.2 Contents of OpenDOAR Repositories, mid-2013

Reach, visibility and citation advantage

Open access repositories have had many positive impacts. They have already fostered greater discoverability and accessibility of the scholarly literature. Despite room for improvement, analyses suggest that Google and Google Scholar refer a great deal of traffic to open access repositories. Organ (2006) quantified the degree to which institutional repository content at the

University of Wollongong (Australia) was downloaded as a result of referrals from search engines. Over the six-month period they studied, Google referrals were responsible for generating 95.8% of measurable full text downloads from their repository.

Harnad and Brody's frequently cited paper (2004) makes the claim that open access articles have dramatic citation advantages and therefore greater impact on scholarship. Eysenbach's 2006 analysis suggested that open access articles "are cited earlier and are, on average, cited more often" than non-open access articles (2006). While these authors' findings about the "citation advantage" have been debated, it does seem clear that open access articles reach a broader audience and are downloaded more often (Antelman 2004; Davis 2011; Gaulé and Maystre 2011).

Proportion of annual scholarly output

For scholars in some disciplines, the subject-based repositories have succeeded in transforming the processes of scholarly communication and fostering worldwide collaboration in the disciplines they support. Deposit mandates (governmental or institutional requirements that researchers make their papers available in open access repositories) have also stimulated growth in the number of open access papers available. For example a new public access policy announced by the US National Institutes of Health (NIH) in 2008 required that papers from NIH-funded projects be submitted to PubMed Central and made publicly available within 12 months of publication (publicaccess.nih.gov). Chapter 8 discusses the benefits and challenges of deposit mandates.

The number of articles published in open access journals or available from open access repositories (both subject-based and institutional) or authors' web pages represents an increasing proportion of annual scholarly output. Björk and colleagues conducted a number of

analyses (Björk, Roos, and Lauri 2009; Björk et al. 2010; Laakso et al. 2011) and estimated that a little over 20% of the articles published in 2008 were openly available a year later. The results of the study that the team published in 2011 suggested that the number of open access journals has grown at an annual rate of 18% since 2000, and the number of open access articles has grown 30% a year.

Challenges of institutional repositories

While some of the subject-based repositories have been eagerly embraced by scholars as vehicles for facilitating collaboration and faster advances in knowledge, scholars have been comparatively slow to deposit their work in other repositories, especially institutionally-based ones. Many authors have documented these low deposit rates by scholars. A study by van Westrienen and Lynch (2005) of institutional repositories in 13 countries suggested that most repositories contained only a few hundred items. The Netherlands was an exception, with an average of around three thousand full-text items each. In a census of institutional repositories in the US, Markey and others (2007) found that respondents' repositories generally contained a few hundred documents; only a handful of reporting institutions had more than 5,000 documents. The number of items deposited in institutional repositories that responded to a survey sponsored by the SURF Foundation for DRIVER (van Eijndhoven and van der Graaf 2007) presented a slightly different picture: this survey found an average size per repository of almost 9,000 items (the average was calculated based on responses from 114 institutions in 17 countries). Even so, finding a repository with more than 10,000 items continues to be the exception rather than the rule at the time of this writing.

How big should they be?

One might ask if a repository at a research university that has 10,000 items in it is small. The answer is yes, provided the repository is not brand new, or highly specialized with a small

audience. Based on the rough estimates of Carr and Brody (2007), if all of the tenured academic research active staff at a UK university deposited all of their annual output (papers, presentations, learning materials, etc.) in the institutional repository, deposits would be in the range of 10,000 items per year. Even if this estimate is high it suggests that a repository of 10,000 items that has been in place five years or more is smaller than one would expect it to be, if faculty were depositing their work in it regularly.

Some repository managers responded to slow faculty adoption rates by batch-loading high-volume files into their institutional repositories, thereby growing their size substantially, as reported by Carr and Brody. Their research suggested that despite their larger size, the patterns of some of these repositories' deposits suggested they still had low faculty deposit rates and low engagement with individual faculty members.

Thomas and McDonald (2007) studied EPrints institutional repositories containing more than 500 deposits of scholarly papers; of 176 EPrints repositories registered at that time in OpenDOAR, only 11 repositories qualified by this measure, and the largest institutional repository they studied contained a little over 6,000 scholarly papers. Despite the implementation of mandates on some campuses requiring researchers to deposit their papers in institutional repositories, small repository size and low researcher deposit rates continue to be central issues for institutional repositories.

Criticism of institutional repositories

Romary and Armbruster (2010) wrote a paper highly critical of the effectiveness of institutional repositories. One criticism had to do with size: "there exist well over one thousand institutional repositories, the majority of which hold very little content." Romary and Armbruster compare central research publication repositories like arXiv and SSRN with the global network of

institutional repositories and argue the advantages of a more centralized solution (using PubMed Central as an example). Henty (2007) enumerates ten major issues for repositories in Australian universities, including engaging the community. Chapter 8 continues the discussion of the problems of recruiting content in the context of community engagement.

Repository systems and software

ROAR and OpenDOAR track repositories worldwide. They also track statistics for repositories by various categories, one of which is type of repository software or platform. The next three sections briefly discuss the two most wide-deployed types of repositories—EPrints and D-Space—plus Fedora. EPrints and DSpace are the only two types of repositories with more than 300 registered installations in ROAR and OpenDOAR. It should be mentioned that in the US, Digital Commons is another common choice of repository software; in a recent survey Mercer and others (2011) found that the Digital Commons software is a distant second to the use of DSpace in ARL libraries.

As of this writing, repositories based on EPrints and DSpace make up over half of the registered repositories in ROAR and OpenDOAR. EPrints and DSpace were early, “open source” (software for which the source code is freely available) packages that enabled building OAI-compliant repositories.

EPrints (eprints.org)

Chapter 2 and earlier parts of this chapter discuss the history of EPrints, which was a significant outcome of the eLib programme in the UK. EPrints 3, launched in 2007, is the current version, as of this writing (Millington and Nixon 2007). It is maintained at the University of Southhampton. Besides being deployed to run hundreds of repositories, EPrints supports a number of JISC projects including ROAR and SHERPA RoMEO (see chapter 8). EPrints also offers a hosting

and consulting service called EPrints Services, which generates an income stream for recovering costs. EPrints' principal contribution besides enabling the building of repositories is to provide support and advocacy for the open access movement. It is difficult to overstate the importance of EPrints' contributions in this regard.

DSpace (dspace.org) and DuraSpace (duraspace.org)

DSpace is currently the most-used repository software; its installations account for 40% of the repositories registered in DOAR and OpenDOAR at this time. Developed at MIT in a joint project with Hewlett Packard (HP), it was released as open source software in 2002, the same year as GNU EPrints. The MIT project, whose purpose was to enable the library to provide a repository for the digital research and educational material of the university, produced a system, tools and a platform that others could deploy for building repositories (Smith et al. 2003; Baudoin and Branschofsky 2003).

MIT's intent from the beginning was to make the software open source and promote it as a new service of the MIT Libraries. A business planning process that began in 2001 concluded with DSpace's being funded initially through MIT (built into the Libraries' operating budget) with supplemental funding from cost-recovered premium services for other libraries (Barton and Harford Walker 2002). A "DSpace Federation" also provided an initial collaborative framework.

In 2007, with more than 200 projects worldwide using the software, HP and MIT established the DSpace Foundation (HP 2007), a non-profit organization, to oversee DSpace using a community development model. In 2009, the DSpace Foundation and the Fedora Commons announced their intention to form a working collaboration (DuraSpace 2009). With a planning grant from the Mellon Foundation, they designed a new support framework for both organizations called DuraSpace (duraspace.org). DuraSpace, a non-profit organization that

develops and deploys open technologies for the purpose of promoting durable access to digital content, is funded through multiple sources (DuraSpace 2012, 5). Funding sources noted in the annual report at that time included grants (75%), community sponsorships (20%) and revenue generating services (6%).

Fedora (fedora-commons.org)

Chapter 2 mentions Fedora (Flexible Extensible Digital Object and Repository Architecture), which started early in the new millennium as a collaborative digital libraries project of Cornell and the University of Virginia. Fedora is more an architecture or infrastructure that uses a modular approach and web services, rather than an out-of-the-box repository solution. The intent was to provide a new framework for interoperable, web-based digital libraries that integrate well with other systems (Lagoze et al. 2006). Fedora offers an open source solution for the foundation of many different types of digital library systems, not only repositories of textual content and metadata. Fedora-based systems support access and preservation for large and complex aggregations of historic and cultural images, artifacts, text, media, datasets, and documents. As part of their article on semantic registries, Powell and others (2011) provide insight into the innovative approach to repositories and interoperability enabled by Fedora.

Subject-based repositories

As previously noted three of the most successful subject-based repositories were launched in the first decade of digital library research and practice (arXiv.org, RePEC, SSRN). Already mentioned in this chapter are Adamick and Reznik-Zellen's two analyses of subject repositories, which contain considerable detail about the state of subject repositories in 2010. They estimate there are 150-400 subject repositories worldwide. To supplement their information, Table 4.1 provides brief information compiled from OpenDOAR, ROAR, DRIVER, and other sources about some highly ranked subject-based repositories.

Table 4.1. Summary information – A sample of subject-based repositories worldwide

Name	Number of items (year reported)	Coverage	Software if known
Social Science Research Network (SSRN)	796,541 (2012)	Social sciences	
arXiv.org	776,811 (2012)	Physics, math, astronomy, computer science, quantitative biology	arXiv (See table 2.2)
Research Papers in Economics (RePEc)	302,882 (2012)	Economics	
UK PubMed Central	2.87 million (2012)	Biology, biochemistry, health, medicine. A mirror site for NIH PubMed Central; in addition contains output of UK researchers	PMC
SAO/NASA Astrophysics Data System (ADS)	Index entries for 9.7M items in 3 databases (including arXiv.org)	Astronomy, astrophysics, physics, geophysics, and the contents of arXiv.org. Abstracts submitted by authors; access to full text when possible.	
CiteSeerX	716,772 (2006)	Computer and information science	
AgEcon Search	55,268 (2012)	Agriculture and applied economics	DSpace
E-LIS	13,564 (2012)	Library and information science	DSpace
Perseus Digital Library	1591 (2012)	Classics	Perseus (see table 2.2)
Organic ePrints	12,887 (2012)	Organic agriculture	EPrints

Institutional repositories

Table 4.2 provides brief information compiled from OpenDOAR, ROAR, DRIVER, and other sources about institutional repositories with ten thousand or more items that were highly ranked by the Cybermetrics Lab in July 2012.

Table 4.2. Summary information – A sample of institutional repositories worldwide

Name	Number of records (year reported)	Software if known
CERN Document Server	1.2 million (mostly metadata) (2012)	Invenio
Universidade de São Paulo Biblioteca Digital de Teses e Dissertações	34,865 (2012)	
HAL Institut National de Recherche en Informatique et en Automatique Archive Ouverte (HAL-INRIA)	219,706 (2012)	HAL
Universitat Autònoma de Barcelona Dipòsit Digital de Documents	81,163 (2012)	Invenio
NASA Technical Reports Server	1,064,884 (2012)	
MIT DSpace	56,528 (2012)	DSpace
Universiteit Utrecht Igitur Archive	32,090 (2012)	DSpace
SIM University DSpace (Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences)	10,728 (2011)	DSpace
National Taiwan University Repository	184,572 (2012)	DSpace
HAL Sciences de l'Homme et de la Société (HAL-SHS)	39,192 (2012)	HAL

Common search services across distributed repositories

OAI-PMH harvesting and service providers

The 1999 meeting in Santa Fe (described in chapter 2) was convened to work on facilitating interoperability across early e-print archives. The meeting generated several significant outcomes—the Open Archives Initiative and OAI-PMH—and led to the growth of OAI-compliant open access repositories. With the ensuing rapid growth in the number of repositories and digital collections generally, digital library researchers and practitioners were soon motivated to establish common search services across distributed repositories (the initial motivation for the 1999 Santa Fe meeting).

OAIster (oclc.org/oaister)

One early project to federate multiple repositories was OAIster, a project undertaken at the University of Michigan in 2001. The project was one of seven funded by the Mellon Foundation that year to study the use of the new OAI harvesting protocol for making catalogs and other valuable content of the deep web more accessible (Andrew W. Mellon Foundation 2001, 28). Hagedorn (2003) provides a detailed description of the history, objectives, methodology and results of the OAIster project, which has been highly influential.

Bitter harvest

Implementers learned that metadata harvesting using OAI-PMH sometimes fell short of being the low-barrier, low-cost framework for interoperability that its creators envisioned. Nevertheless it was better than other approaches available within the digital library community at that time. Tennant, then at the California Digital Library (2004a), reported his shock at his “bitter harvest” of metadata and provided an overview of the problems with OAI-PMH harvesting, most of which stemmed from a lack of common practices for creating harvest sets and applying metadata. Many problems were later alleviated through community work to define and adopt best

practices, so that OAI-PMH is commonly used as a protocol for federating metadata from distributed repositories.

DAREnet and other projects

BASE (discussed in chapter 5), early projects of the US National Science Digital Library (discussed in chapter 8), DAREnet (see for example Dijk et al. 2006; Hogenaar and Steinhoff 2008) and the China Digital Museum Project (Tansley 2006) are or were among a number of services that brought distributed repository content together, using OAI-PMH by itself or as part of a larger framework for achieving interoperability.

New approaches to discovery

DRIVER (driver-repository.eu)

Peters and Lossau (2011) describe DRIVER, a project funded by the European Commission to build a sustainable global digital infrastructure for the networking of European open access repositories. With inspiration from the DARE project, the project team began work in 2005 and completed the first phase of DRIVER in 2007. DRIVER-II extends and builds upon its results (EC Framework Programme 7 2007).

DRIVER has succeeded in delivering a common infrastructure and establishing a confederation of European digital repositories, called COAR, which has advocacy, policy and coordination roles (coar-repositories.org). As of June 2011 the DRIVER search portal provided access to 3.5 million publications from 295 repositories in 38 European countries (driver-repository.eu/PublicDocs/FACT_SHEET_I3_driver_ii.pdf). Manghi and others (2010) discuss why DRIVER represents an important breakthrough in the evolution of repositories.

aDORe

The developers of aDORe were among many repository managers who, as the scale of digital content grew, faced "the harsh reality that their solutions need to handle an amount of artifacts that is orders of magnitude higher than originally intended, and are reaching an understanding that approaches that work at the originally intended scale do not necessarily work at that next level" (Van de Sompel, Chute, and Hochstenbach 2008). A team at Los Alamos developed the aDORe federation architecture starting in 2003 to address three problems with their current design:

- Their approach was "metadata-centric, treating descriptive metadata records as first class citizens and the actual digital assets as auxiliary items"
- They stored tens of millions of digital assets as files in a file system, resulting in "a system administrator's nightmare"
- Their design at that time tightly integrated content and the discovery application, preventing other applications from re-using the content

Solutions were readily identified: use a compound object view instead of a metadata-centric view; put assets in storage containers instead of files; separate the repository from applications; and provide the needed machine interfaces. Implementing the solutions however, required multiple years of work. aDORe is one of several projects that signaled a trend to design repositories and interoperability mechanisms using approaches more closely aligned with those of web developers and the architecture of the web.

Building on and for the web: web services

Web services and web APIs are the currently dominant enabling technologies for supporting Interoperability, in particular the exchange and re-use of content between sites. At this point a few sentences about web services and web APIs will be helpful for understanding how

repositories and library systems have been changing and will continue to change as they adopt web-based approaches. This brief discussion begins with XML (Extensible Markup Language).

XML

XML was an important development on the web because it created a basis to exchange information in new, useful ways. XML was designed to support the encoding of structured data in a relatively simple and standardized way, to be usable over the web, thus facilitating the development of tools for automatic processing of this content and its use in applications and syndication (web feeds such as alerts).

Web services

Using XML to code and decode information and a protocol to transport the data over HTTP (Hypertext Transfer Protocol), web services take applications that run on the web and publish those applications in a machine-readable way. A directory or registry enables the registration and location of web service applications. These components, taken together, allow servers of different types and in different places to find and use each others' web services. Web services provide applications that can be called at any time, like a weather report or a converter that changes yards to meters, or they be used to exchange data.

Web APIs

Web APIs (Application Programming Interfaces) are like web services except they have moved toward simpler communications methods. Web API requests and responses between systems are expressed in XML or a simpler alternative called JSON. Web APIs are what enable combining multiple services into new applications (called "mashups").

It is important to remember that web services, Web APIs and mashups provide software-to-software interfaces and are not user interfaces. As discussed in chapter 3, the semantic web and linked data may offer new approaches to interoperability and the exchange of digital content between sites.

Data reuse, disclosure and dissemination

Object Reuse and Exchange (ORE)

A seminal 2004 article about rethinking scholarly communications and "building the system that scholars deserve" (Van de Sompel et al. 2004) brought to the fore changes in the nature of scholarly research, which by that time took place on the internet and was highly collaborative, international and data intensive. Citing opportunities to facilitate network-based collaboration and increase the speed of scientific discovery, the authors proposed a fundamental redesign to replace the increasingly problematic existing system. They made a case for a natively digital and interconnected set of services for capturing, making accessible, and preserving the scholarly record. Their envisioned system focused on "units of communication" (text, datasets, simulations, software and more) moving through "pathways" associated with the scholarly communications "value chain": *registering* a new unit of communication; *certifying* it through peer review; generating *awareness* of the new unit; *archiving* it; and enabling *"rewards"* (e.g., being cited by others). The proposed system would bring many distributed services and players together and require the easy re-use and exchange of units of communication as well interconnections to support the flow of units through the system,

Two years later, Van de Sompel and his colleagues (2006) reported on work supported through the NSF Pathways project to investigate and prototype a natively digital, interconnected and interoperable scholarly communications system based on distributed repositories. Their prototype had three levels: repositories with internal data models and services for machine

interaction; an interoperability layer; and a top layer consisting of registries for supporting shared infrastructure and loosely federating autonomous participating systems. In the prototype they experimented with building pathways for objects from arXiv, aDORe, DSpace and Fedora. They received funding from the Mellon Foundation to continue their work on a new OAI project called Object Reuse and Exchange, or OAI-ORE.

The ORE project has been influential. The team reported the results of the project—a set of specifications and user guides—in a paper that combined the characteristics of a technical report and a call to action (Lagoze et al. 2008). In keeping with the earlier work reported, the objective of the specifications and guides was to make it possible for many systems to use scholarly digital objects. The paper begins and ends with the reasons for the digital library community to embrace methods that are more integrated with web architecture and more accessible to web applications.

The principles of web architecture (www.w3.org/standards/webarch), the semantic web and linked data are the basis of the OAI-ORE specifications (Van de Sompel et al. 2009). A new publication by Lagoze et al. (2012) updates and expands on the earlier 2008 paper and discusses related work.

Lagoze and his colleagues, together with Tarrant and others (2009) contrast the approach of ORE with the interoperability mechanisms of OAI-PMH, which harvests metadata from repositories. OAI-PMH was a first step toward repository interoperability; OAI-ORE, while not an extension or replacement for OAI-PMH, provides a model for expressing digital objects for exchange and re-use in many contexts.

Repository implementations of ORE

ORE offers a way to make the scholarly content in repositories easier to exchange and re-use across systems and services. It makes it possible to identify and interlink many types of scholarly resources—pre-prints and their corresponding refereed publications, software, e-research data, visualization, one or more presentations, etc. The ORE approach has numerous use cases in the domain of repositories. A couple of implementations described in the literature are:

- Tarrant and others (2009) describe their award-winning demonstration of an application using ORE at the Open Repositories Conference 2008. Their demonstration combined OAI-ORE with the Fedora and EPrints repository platforms and transferred two live archives from one software to the other. Their work was completed under the aegis of the JISC-funded Preserv 2 project (preserv.eprints.org), which sought a way to replicate an entire repository across any repository platform. The ORE import and export plug-ins are available with the EPrints software.
- Maslov and others (2010) report on their OAI-ORE project for the Texas Digital Library to add OAI-ORE support to the DSpace repository platform, enabling better data exchange between repositories.
- Foresite (foresite.cheshire3.org), a JISC-funded demonstration project that uses ORE to describe the compound digital objects that make up JSTOR (journals, issues, articles, pages), which can then be referenced by repositories (Witt 2010). Foresite uses ORE with SWORD, described in the next section.
- In work related to OAI-ORE, Haslhofer and Schandl (2008 and 2010) describe the work they did to create the OAI2LOD Server, which exposes OAI-PMH metadata as linked data. Haslhofer was later active in work on the Europeana Data Model, discussed in chapter 10.

ORE and e-research data

Research related to cyberinfrastructure (discussed in chapter 8) has led to more attention for e-research data (Van de Sompel et al. 2009). An example of a project deploying ORE is the US National Virtual Observatory, which has the goal of enabling new ways of doing astronomy by combining astronomical data from telescopes worldwide. Librarians at Johns Hopkins, working with the American Astronomical Society, have contributed to the project by using ORE and SWORD to enable automatic capture of data related to an article as part of the article submission process (DiLauro 2009).

Another use case for OAI-ORE is for modeling the many interrelated objects described in archives (Ferro and Silvello 2013), in the process making them easier to find, link to and reuse on the larger web. Witt (2010) provides a readable and useful set of descriptions of ORE implementations in about a half a dozen projects in the US, UK, Belgium and Australia—some but not all related to repositories. Several of the projects use SWORD.

Simple publishing interface (SPI) and SWORD

The Simple Publishing Interface (SPI) was developed under the auspices of the European Committee for Standardization (see Ternier et al. 2010). Like ORE it was another development that made it easier to disseminate and re-use content and metadata in multiple systems and applications. SPI grew out of the e-learning community; it is a protocol used in combination with AtomPub, a format for web feeds commonly used by web developers. SPI is important because content and metadata can be created once and consumed directly in multiple applications. Ternier's paper provides a number of scenarios for using SPI and explains how it differs from SWORD.

SWORD, the Simple Web-service Offering Repository Deposit, was “unapologetically built on and for the world wide web: in this it differs from many information exchange protocols arising out of the library/repository domain” (Duranceau and Rodgers 2010). A UK working group, supported as part of JISC’s Digital Repositories Programme, initially developed SWORD (ukoln.ac.uk/repositories/digirep/index/SWORD_Project). The working group was seeking a lightweight method for facilitating deposit in institutional repositories. Version 2 of SWORD was released in 2010. Duranceau and Rodgers describe an experiment in which MIT successfully used SWORD to enable automatic deposit of papers published by BioMed Central into DSpace@MIT, the institutional repository. Lewis and others (2012) describe nine different scenarios to demonstrate the many ways in which SWORD can make it easier for faculty and repository managers to deposit new scholarly output in multiple locations. Some of the scenarios considered are:

- publisher to repository
- research information system to repository
- desktop to repository
- repository to repository

As of this writing, arXiv, DSpace, EPrints, Fedora and a few other repositories are SWORD-compliant.

The semantic web and semantic interoperability

As discussed in chapter 3, the semantic web and linked data have inspired excitement and much discussion in the field of digital libraries. Fedora and ORE, described earlier, use semantic web methods, as does VIVO, a researcher profiling system (chapter 9).

Next generation repositories

Islandora and Hydra, described in the two sections that follow, are the outcomes of new thinking about repositories, their architectures and objectives and new approaches to achieving interoperability. They are not themselves repositories; rather they provided a layer on top of a repository that supports specific interactions with repository content: deposit, discovery, display, etc.

Islandora and Drupal

Islandora (islandora.ca) is an open source framework that combines Drupal (a web content manager) and Fedora Commons repository software in a digital asset management system. It was developed in 2006 at the University of Prince Edward Island (UPEI). Mark Leggott of the UPEI development team noted the choice of Fedora for its data models and ability to support diverse content types (2009). Islandora supports creating, editing, discovering, viewing and managing digital assets in a Fedora repository. It has been used to create "Virtual Research Environments" or VREs. For a broader perspective and more background on VREs, De Roure, Goble, and Stevens' highly-cited paper (2009) makes the case for systems enabling shared scholarly workflows in a virtual research environment.

Scholar's Workbench and the Hydra Project

Green and Awre (2009) describe two JISC-funded projects undertaken at the University of Hull from 2005 to 2009—RepoMMan and REMAP—that led to the Hydra Project. RepoMMan provides a browser-based interface and web services to support scholarly authoring and deposit processes. A second part of the project focused on publishing the scholar's content to a public-facing repository, in the process automatically generating metadata for the object. REMAP publishes and preserves the content. Green and Awre carried this work forward into a multi-

institutional collaboration called the Hydra Project, whose initial stage ran through 2009 and involved three institutional partners. The project developed a Scholar's Workbench that provided a search and discovery interface and also enabled interactive workflows for pre- and post-publication. The word "hydra" was chosen deliberately to convey the concept of one body and many heads, that is, one common Fedora repository with many purposes or applications. The point was to create a framework for integration and content re-use whose pieces could be deployed at multiple institutions.

The next stage of the Hydra Project ran from 2008 to 2011, included more partners, and produced a framework that uses web architecture and web tools to support a range of uses and workflows. Awre and others (2009) defined a number of use cases that a Hydra implementation might support, including accessioning digital content, managing a personal or institutional repository, and integrating content across systems or services. Awre and Cramer (2012) report on the project's most recent progress and new partners; the project wiki provides further information ((wiki.duraspace.org/display/hydra).

Repositories in the cloud

Chapter 5 discusses cloud services in hybrid libraries. The potential for cloud services for repositories is a new area of investigation. In 2010 JISC and Eduserv organized an event to discuss "Repositories and the Cloud." a new area of investigation and experimentation. A JISC-sponsored project called "Fedorazon" had looked into setting up a repository using Amazon cloud services (Flanders 2008). Following a pilot program with the Library of Congress and a number of partners, in late 2011 DuraSpace began offering a service based on DuraCloud, a cloud-based platform for backing up, archiving and preserving repository content (duracloud.org). The 2011 DuraSpace annual report (2012) discusses the strategy and possible uses for DuraCloud.

Conclusion

Subject-based repositories appear to be on a firm footing. All repositories are contributing to the broader diffusion of knowledge to the public—an important social role of digital libraries, as discussed in chapter 6. As for institutional repositories, a number of developments indicate they could get past current barriers, move to a new level, and take on broader roles in libraries, research institutions, on the web and in society. There is also the possibility that institutional repositories will evolve beyond their current forms. Chapters 8 and 9 further discuss the opportunities and challenges for repositories.