CHAPTER

DIGITAL PRESERVATION

9

INTRODUCTION

Digital preservation represents an emergent area of digital library research and practice. It focuses on the policies, technologies, and strategies to ensure that digital library objects and collections are available and usable now and in the future. Digital preservation encompasses materials born in the digital format as well as those converted from the analog format through the digitization process. Concerns about preserving digital content are not unique to digital libraries. All resources in the digital format are fragile and susceptible to information loss. Multiple risks stem from the unstable nature of digital formats, degradation of storage media, and technological obsolescence. As the members of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access note, digital preservation is a universal and "urgent societal problem" (Berman et al., 2010, p. 9).

In the context of digital libraries, the challenge of digital preservation is compounded by the structural complexity of digital objects and the interrelatedness of objects, collections, and repositories (Ross, 2012). The models and solutions for preserving content in present-day digital libraries will have an impact on future access to cultural heritage and scientific information. As Seadle (2008) emphasizes, "the digital libraries in 100 years will face problems that stem from the choices that we as librarians make today" (p. 5). Long-term preservation of digital content is recognized as a core function of digital libraries in the DELOS Manifesto (Candela et al., 2007), but it does not feature prominently in other definitions and frameworks. The practice of preserving digital objects has evolved since the first digital libraries were developed in the mid-1990s, but research in the area of digital preservation is relatively new (Chowdhury, 2010; Ross, 2012).

The concept and principles of preserving analog materials are well established in the library and archival community, although, as Cloonan (2007) points out, the institutional, custodial model is somewhat paradoxical in the modern world. Preservation is understood as an act of responsible custody aimed at preventing the deterioration of cultural heritage materials and restoring their usefulness and information value (Conway, 2007, 2010). Conway argues that the fundamental principles of digital preservation are the same as those of the analog world and define the priorities for ensuring the longevity and the useful life of information resources. The core concepts of preserving analog materials in regard to longevity, choice, quality, integrity, and access carry over to the digital environment, but the methods and practice are fundamentally transformed.

A comparison of analog and digital preservation approaches points to a continuum in principles but also highlights the distinct nature of preservation activities in the digital realm. In contrast to traditional practices, digital preservation is an urgent and ubiquitous issue. All digital objects, rather than selected items, are subject to preservation, although the level of activities performed on the object can differ. Digital objects are inherently more vulnerable than analog materials and require immediate attention

Discover Digital Libraries. http://dx.doi.org/10.1016/B978-0-12-417112-1.00009-0

from the point of creation. The standards and formats selected for encoding have implications for the quality and long-term maintenance of digital content. As Walters and Skinner (2010) stress, "the ways that objects are created, curated, and stored matter immensely in how preservation-ready they ultimately are" (p. 264).

Digital preservation needs to be ongoing with activities integrated into all phases of creating, managing, and storing information. Cloonan (2015) emphasizes the dynamic nature of the digital preservation cycle—that it is not linear and requires multiple actions. Lavoie and Dempsey (2004) point out that digital preservation is not an isolated activity but rather a set of practices diffused throughout the information lifecycle. In the analog world, conservation activities tend to occur toward the end of a resource's lifecycle. Once physical items receive conservation treatment and are stored properly, no additional attention may be required. In contrast, this type of "benign neglect" can be catastrophic for digital materials (Corrado and Moulaison, 2014; Ross, 2012; Walters and Skinner, 2010). Ross (2012) notes, "as a result of the constant evolution of technology, the degradation of storage media and the ever-increasing pace of 'semantic drift,' digital objects do not, in contrast to many of their analog counterparts, respond well to benign neglect" (p. 46). Digital preservation involves not only an active and continuous management of digital content but also monitoring of the evolving technological environment and preservation methods.

Digital preservation is a complex technical, social, economic, and organizational issue. Its complexity in digital libraries stems from the fact that it is interwoven into the process of creating, using, and maintaining a wide array of digital materials and collections. The sustainability of digital content depends on the careful management of preservation risks, organizational policies, institutional commitment, and technical infrastructure (Bradley, 2007; Corrado and Moulaison, 2014). Technical aspects have received a considerable amount of attention in the preservation community because of the immediate need of keeping intact files and protecting them from storage media failure and obsolescence. Increasingly, the researchers in the digital library field recognize that contextual information needs to be preserved along with the bitstream to render the bits as useful and meaningful objects (Beaudoin, 2012a; Chowdhury, 2010; Ross, 2012). Lesk (2014) captures the broader aspects of digital preservation by observing, "the greatest danger to digital materials is that we forget the meaning of them. Preservation depends on our knowledge: we may have bits but be unable to interpret them" (p. xvi).

The field of digital preservation is still evolving, but significant progress has been made in building technological infrastructure and in developing policies, recommendations, and standards. The Task Force on Archiving Digital Information was established in 1994. The work of the Task Force resulted in a foundational report, which not only identified the critical challenges to preserving digital content but also provided a set of far-reaching recommendations (Waters and Garrett, 1996). The National Digital Information Infrastructure Preservation Program (NDIIPP) was formed by the Library of Congress in 2000. The National Digital Stewardship Alliance continues the work of NDIIPP, setting the agenda for national digital preservation and contributing to the development of standards and tools. Similar collaborative initiatives have been established in other countries with exemplary programs in the Netherlands and New Zealand (Library of Congress, n.d.). A number of research projects undertaken in Europe, including DELOS, Open Planets Foundation (currently Open Preservation Foundation), and DigitalPreservationEurope, have had a significant impact on advancing the field of digital preservation (Ashenfelder, 2011; Brown, 2013; Library of Congress, n.d.). The last two decades of preservation research and practice resulted in developing more stable formats, preservation metadata standards, and trusted repositories.

DEFINING DIGITAL PRESERVATION

Several definitions and conceptual models of digital preservation emerged in the cultural heritage community. They tend to focus on general policies, strategies, and activities to ensure access to any cultural or scientific information encoded in digital form. The context of digital libraries and their complex structures are rarely addressed as a separate issue. Conway (2007) notes, "preservation remains an ill-defined concept when applied to the development of digital library projects and collections" (para. 1). digital libraries not only share goals and approaches for preserving digital content with other scientific and cultural heritage domains, but also face unique challenges in regard to maintaining the relationships between digital objects and collections. The evolving terminology and theoretical models in digital preservation and associated disciplines impact the understanding of preservation in the context of digital libraries.

TERMINOLOGY: DIGITAL CURATION, DIGITAL STEWARDSHIP, AND DIGITAL PRESERVATION

The plethora of terms that are used to describe activities associated with managing and maintaining digital assets complicates an attempt to understand and define digital preservation. Many authors note that digital preservation is a young discipline and precise vocabulary has yet to mature (Brown, 2013; Jones and Beagrie, 2008). A number of alternative terms have been used for the same or similar concepts, reflecting different origins or the evolving understanding of the concept. Terms, such as digital curation, digital stewardship, and digital preservation, are often used interchangeably. Caplan (2008) points out the differences in usage between the United States and the United Kingdom (UK). In the United States, the use of the term digital preservation tends to be broader and encompasses all activities in managing digital assets from the point of creation. In the UK, the term digital curation is used for lifecycle management, while digital preservation is reserved for those activities specifically geared toward future accessibility (Caplan, 2008). Lazorchak (2011) attempts to discern "what's in (some) names" by looking at their different origins and context of use. He notes that preservation has a long-standing tradition in the cultural heritage community and is a core component of broader concepts, such as digital curation and digital stewardship.

Digital curation originated in the scientific and e-science community with a focus on research data and the entire information lifecycle. The term is relatively new. The Digital Curation Centre (DCC), a UK-based consortium that was launched in 2004, has contributed to promote the concept (Higgins, 2011). Digital curation, as defined by DCC, involves maintaining, preserving, and adding value to digital research data throughout its lifecycle (DCC, 2004–15). The digital curation cycle includes the whole range of actions from creation, through access and use, to transformation. Preservation is one of the actions undertaken throughout the curation lifecycle to ensure the long-term maintenance and retention of digital objects (Higgins, 2008). Harvey (2010) expands on this definition by emphasizing active management of data and the goals of digital curation in "supporting reproducibility, reuse of, and adding value to that data, managing it from its point of creation until it is determined not be useful, and ensuring its long-term accessibility, preservation, authenticity, and integrity" (p. 8).

Digital stewardship has its roots in the cultural heritage community. It is promoted as a broader concept, encompassing both cultural heritage materials and research data. Lazorchak (2011), in the Library of Congress blog, notes, "digital stewardship satisfyingly brings preservation and curation together in one big, happy package, pulling in the lifecycle approach of curation along with research in digital libraries and electronic records archiving, broadening the emphasis from the e-science community on scientific data to address all digital materials” (para. 11). The term has been adopted by the National Digital Stewardship Alliance (NDSA), a consortium of US research, government, and cultural heritage institutions committed to the long-term preservation of digital information. The definition of digital stewardship included in the *2014 National Agenda for Digital Stewardship* (NDSA, 2014) echoes earlier definitions of digital preservation, but the shift in the vocabulary indicates a broader approach. The adoption of the term, however, is still limited.

Digital preservation is an integral part of digital curation and digital stewardship frameworks and thus applies to activities focused on managing and preserving a wide range of materials from scientific data to cultural heritage resources. The term digital preservation has been used the longest (Cloonan, 2015). It makes a connection to the principles of analog preservation and places the new activities of curating digital content in the long tradition of preserving cultural heritage materials. Digital preservation is at the center of several definitions adopted in practice and is used in the context of digital libraries.

DEFINITIONS OF DIGITAL PRESERVATION IN THE PRACTICE COMMUNITY

The early definitions and models of digital preservation were developed in the practice community, often as part of training efforts to prepare library professionals for the emerging discipline. The Digital Preservation Management (DPM) Tutorial, launched by the digital preservation team at Cornell University Library in 2003 and maintained since by the DPM workshop faculty, offers a working definition of digital preservation as “a broad range of activities designed to extend the usable life of machine-readable computer files and protect them from media failure, physical loss, and obsolescence” (DPM Tutorial, 2003–15, para. 1). The DPM Tutorial makes an important distinction between preservation activities that promote the long-term maintenance of bitstream and those that provide continued and meaningful access to its content. The Tutorial also cites the DPM model, an early example, showing the multiple dimensions of digital preservation with three core components:

- Organizational infrastructure with policies, procedures, practices, and people
- Technological infrastructure consisting of the required equipment, software, hardware, a secure environment, and skills
- Resources framework that addresses the necessary funding for starting, continuing, and sustaining the digital preservation program

The authors use the metaphor of a three-legged stool (see Fig. 9.1) to demonstrate that digital preservation is not just a technical issue. Fully implemented and viable preservation programs require balancing of technological infrastructure, organizational aspects, and funding resources.

The Association for Library Collections and Technical Services, a division of the American Library Association (ALA), developed a range of definitions to promote an understanding of digital preservation within the library community. The core concepts are presented in short, medium, and long versions to accommodate a variety of needs. The medium definition states:

> Digital preservation combines policies, strategies, and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time (ALA ALCTS, 2007, para. 8).

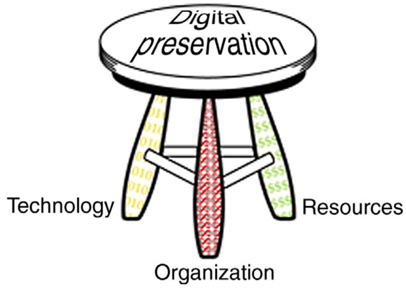

FIGURE 9.1 Three-Legged Stool Representing Three Aspects of Digital Preservation (Kenney and McGovern, 2003)

This definition has been widely adopted and incorporated into more recent statements. The definition of digital stewardship articulated in the *2014 National Agenda for Digital Stewardship* emphasizes "policies, strategies, and actions that ensure that digital content of vital importance to the nation is acquired, managed, organized, preserved, and accessible for as long as necessary" (NDSA, 2014, p. 6).

Digital Preservation Coalition (DPC), a UK-based organization, defined digital preservation broadly as the series of managed activities necessary to ensure continued access to digital materials, but also distinguished different levels of preservation activities:

- Long-term preservation that ensures continued access to digital materials, or at least to the information contained in them, indefinitely
- Medium-term preservation that provides continued access to digital materials beyond changes in technology for a defined period of time but duration is not indefinite
- Short-term preservation that provides access to digital materials either for a defined period of time, but it doesn't extend beyond changes in technology (Jones and Beagrie, 2008, p. 25)

The levels of preservation help to establish institutional goals for preservation programs and specify the requirements for repositories. This approach also reflects a realistic assessment of the digital landscape where not all resources will or should be preserved indefinitely.

Recent discussions of digital preservation are built on the early foundational concepts but shift the attention from the challenges of technology to management issues. Corrado and Moulaison (2014) analyze the key aspects of the DPC definition in terms of policy implications. Those key aspects include managed activities; necessary, continued access; and digital materials. The authors acknowledge the importance of technological infrastructure and technical skills but stress that digital preservation is mostly a management issue. Corrado and Moulaison (2014) propose a modified model of the three core components of digital preservation. The Digital Preservation Triad is presented in a form of Celtic knot with three interconnected parts: technology, content, and management. The emphasis of this model is on planning and policy development.

UNDERSTANDING DIGITAL PRESERVATION IN THE CONTEXT OF DIGITAL LIBRARIES

Research on theoretical models of digital preservation in the context of digital libraries has been limited. The fields of digital preservation and digital libraries have existed side by side for over two decades, but the emphasis of digital library research has been primarily on expanding access to cultural heritage and scientific resources rather than long-term sustainability (Chowdhury, 2010; Ross, 2012). The early discussions of digital preservation make a distinction between digital archives and repositories intended for long-term preservation and digital libraries, which provide access to digital information, but may not be committed to its long-term preservation (DPM Tutorial, 2003–15; Waters and Garrett, 1996). In practice, this distinction has prevailed with content management systems specifically for building digital collections for access and a separate suite of repository systems. In the research domain, however, this situation has been changing gradually, with more attention given to the issues of digital preservation in digital libraries and calls for conceptual models and frameworks for capturing contextual information. Both technological and semantic aspects of digital preservation are discussed in research literature.

In practice, analog and digital preservation goals coexist in many digitization projects and are a source of some confusion between the concepts. Digital libraries contain a significant number of resources converted from analog formats, including fragile and deteriorating materials digitized for preservation purposes. In this context, and in light of broader discussions about preserving cultural heritage with digital technologies, it is necessary to make a clear distinction between digital preservation and digitization for preservation (Caplan, 2008; Conway, 2010). Digitization as a preservation strategy undertakes the conversion of deteriorating analog materials to create high-quality copies for preservation purposes, while digital preservation activities focus on the preservation of the digital assets created as a result of digitization as well as born digital materials. As Conway (2010) points out, "digitization for preservation creates valuable new digital products, whereas digital preservation protects the value of those products, regardless of whether the original source is a tangible artifact or data that were born and live digitally" (p. 64). Digital preservation represents an important component of all digitization projects, as digital master files created as a result of the conversion process require long-term preservation activities. Digitization as a preservation method and a debate surrounding this approach are discussed in more detail in Chapters 3 and 4.

A distinction between bitstream preservation and semantic aspects of preserving digital content is particularly relevant in the digital library context. Bitstream preservation focuses on the technical aspects to ensure that bits are accessible and usable, while semantic preservation is concerned with maintaining the means of interpreting informational content of preserved bits. Technical level of bitstream maintenance is critical, but as Ross (2012) stresses, digital preservation is more than preserving the streams of 0s and 1s. Digital preservation is about "maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its 'interrelatedness,' and about securing information about the context of its creation and use" (Ross, 2012, p. 45). Contextual information needs to be preserved along with the bitstream to render the bits as useful and meaningful information objects. Furthermore, Ross (2012) notes that the approach to preservation in the digital library environment needs to vary with different levels of preserving the content and context, including:

- Retaining the content of some materials held in digital libraries
- Retaining the environment and context of creation and use for other materials
- Reproducing the experience of use for other materials to ensure the right semantic representation and information is passed to the future

In his paper, Ross (2012) also argues for more research on preservation issues in digital libraries and for the development of theoretical models and a knowledge base. He proposes adopting the framework of archival science with the core principles of authenticity, uniqueness, provenance, arrangement, and description as a theoretical foundation for digital preservation in digital libraries.

The contextual dimension of digital preservation emerges as a new area of research. Chowdhury (2010) examines the research on digital preservation from the user perspective and concludes that context, especially information about the time and place where digital objects were created, is particularly important to facilitate understanding, interpreting, and future use of digital content. Time and place also appear as important categories in a framework for contextual information in digital collections (Lee, 2011). The framework proposed by Lee focuses on the contextual categories required for the comprehensive documentation of the "life history" of digital objects. The proposed contextual information framework identifies nine classes, including object, agent, occurrence, purpose, time, place, form of expression, concept, and relationship. Lee (2011) notes, "relationships to other digital objects can dramatically affect the ways in which digital objects have been perceived and experienced" (p. 6). The relationship class allows for capturing contextual information about the relations of the object as well as the collection level. Although the framework does not focus specifically on preservation, the proposed classes of contextual information are intended to support the curation of digital objects and collections.

Beaudoin (2012a) emphasize the role of context in the digital preservation of cultural objects and argues that knowledge about the context of digital objects is critical to making sense of them—their use, care, and preservation. The current approaches to recording contextual information in descriptive metadata are insufficient to ensure meaningful preservation and use of digital objects. Beaudoin (2012a) stresses that "context is especially important in discussions of digital preservation since in most instances the digital materials have been separated from their original format and context in the process of digitization and preservation" (p. 3). She proposes a framework for the use of contextual metadata in the digital preservation of cultural objects with a focus on multiple dimensions, including technical, intangible, utilization, curatorial, authentication, authorization, and intellectual (Beaudoin, 2012a, b).

A comprehensive approach to digital preservation that includes both technological and semantic aspects is particularly important in the context of digital libraries. Digital objects that represent the content of deteriorating analog materials serve as preservation copies, and contextual information is critical to their understanding, future use, and long-term management. The recent research on the contextual dimensions of digital preservation provides a foundation for the undertaking of broad preservation activities on digitally born as well as digitized objects. Digital preservation in the context of digital libraries concentrates on policies and technologies to ensure the long-term maintenance and rendering of digital files as well as the retention of the contextual information that enables interpretation of digital objects and collections.

PRESERVING DIGITAL CONTENT

Digital preservation is challenging due to the vulnerable nature of digital objects and the constantly changing technological environment. Information encoded in digital form can be easily altered and corrupted, which is a source of distrust and concerns about the authenticity and integrity of digital objects. In addition, turning digital objects into meaningful and usable information requires multiple layers of technology. As Lynch (2000) states, "bits are not directly apprehended by the human sensory apparatus—they are never truly artifacts. Instead, they are rendered, executed, performed, and presented to people by hardware and software systems that interpret them" (p.4). The reliance on software and hardware poses risks to access because this technology is susceptible to technological failure and obsolescence. A number of strategies as well as practical guidelines in the digital library field have been developed to address the challenges in maintaining integrity of digital objects and ensuring long-term accessibility. Authenticity remains a challenge because of the very nature of digital information and the connection to broader issues of trust and organizational management.

DIGITAL PRESERVATION CHALLENGES

> Digital objects break. Digital materials occur in a rich array of types and representations. They are bound to varying degrees to the specific application packages (or hardware) that were used to create or manage them. They are prone to corruption. They are easily misidentified. They are generally poorly described or annotated (Ross, 2012, p. 44).

The list of challenges to preserving digital content can go on, but, as the above quote indicates, it starts with the fragile and complex nature of information in digital form. Unlike resources in analog form where content and carrier are inseparable, digital objects are not affixed to any permanent medium. Instead, informational content encoded as streams of 0s and 1s is copied from one storage medium to another and transmitted over networks. On one hand, this separation from physical carriers offers tremendous benefits for access and even for preservation as multiple digital copies can be stored in several different locations. On the other hand, the lack of permanence poses risks to the authenticity and integrity of information encoded in digital form. Information unattached to a permanent medium can be easily altered, damaged, or even destroyed. The lack of fixity and the separation of descriptive metadata from content files also make it harder to determine the authorship and provenance.

In addition, digital objects may exist in multiple copies, in several manifestations, and may have associated representation information. This phenomenon is evident in digitization projects where multiple copies of master files are created for preservation purposes. A master file is then a source for several derivatives, which tend to be smaller and have a different configuration of bits. Yet, all these different objects are representations of the same informational content, ideally described by consistent and linked metadata. A useful distinction between data objects and information objects is made in the Open Archival Information System (OAIS) reference model, described later in this chapter in the Digital Preservation Technology section. Brown (2013) expands on this concept and notes, "each conceptual information object can be manifested through one or more different data objects, which can in turn exist in multiple identical copies" (p. 201). He recognizes two primary threats to preservation in light of this distinction:

- Loss of data objects, referring to the physical loss of 1s and 0s that encode information
- Loss of the information object, referring to the loss of means to interpret those 1s and 0s as meaningful and authentic information (Brown, 2013, p. 202)

Another preservation challenge lies in determining what objects need to be preserved long-term, distinguishing between multiple manifestations and their copies, and maintaining the relationships between them.

Thibodeau (2012) recognizes the complexity and fluidity of digital information as a major challenge in preserving digital memory, depicting it as "a shape shifter that takes on very different forms" (p. 15). He relates the challenges in preserving cultural heritage in digital form to the characteristics of digital information itself and the rapid rate of technological change. Digital information is fluid, variable, and, as Thibodeau describes it, "polymorphous." This polymorphism results from several key factors, including:

- The transmission from one storage medium to another.
- The difficulty of determining the boundaries of digital objects, which in many cases are dynamic and dependent on external applications; the dynamic and transitory nature of those objects makes it difficult to define what content needs to be captured and preserved.
- The complex relationship between data objects stored in computer systems and objects presented to users through the online delivery systems.
- The necessity to process data objects with computer technology in order to be used; the process of transmission and rendition can involve changes in the object structure and even lead to alteration or corruption.

Thibodeau (2012) views the lack of permanence of digital objects as a source of tension between fluid digital information and digital preservation, which seeks to keep things unchanged.

The unstable and mutable nature of digital information poses risks to its authenticity and integrity. Ross (2002) indicates, "digital objects that lack authenticity and integrity have limited value as evidence or as information resource" (p. 7). The concepts of authenticity and integrity have been debated in the field of digital preservation in an attempt to determine the essential properties of digital objects that need to be preserved. They relate to the basic questions in digital preservation: (1) How do we know that digital objects are complete and have not been altered or corrupted? and (2) Are preserved digital objects reliable and genuine representations of what they claim to be? In an exploratory paper, Lynch (2000) provides working definitions of these fundamental, yet elusive concepts:

- *Integrity* means that a digital object has not been corrupted over time or in transit; in other words, that we have in hand the same set of sequences of bits that came into existence when the object was created (p. 5).
- *Authenticity* entails verifying claims that are associated with an object—in effect, verifying that an object is indeed what it claims to be (p. 6).

The integrity of files can be checked through technical measures such as checksums or digital signatures, but as Lynch (2000) comments, verifying authenticity is more challenging, as it requires judgment and an inquiry into an object's nature, provenance, and chain of custody. He relates the process of verifying authenticity and integrity to the broader concepts of trust and identity. One of the factors users employ to determine the authenticity or integrity of digital information is the level of trust attributed to the infrastructure or the organization responsible for preservation activities. Ross (2002) expands on the concepts underpinning authenticity and integrity. In addition to the trust, he lists fixity, stabilization, and the requirements of custodians and users. Furthermore, user needs and requirements can vary and may depend on the types of objects.

In a more recent article, Seadle (2011) reexamines the concepts of authenticity and integrity in light of the criteria used in the evaluation of analog materials. He acknowledges the difficulty of defining and assessing digital authenticity and states, "in the digital world, there are not originals, only copies, and the mutability of digital objects makes authenticity especially challenging" (Seadle, 2011, p. 548). While digital integrity can be measured through checksums, there are no clear measures for authenticity. In his comparative analysis, however, he notices that the concepts of authenticity and integrity in the digital environment are more closely related. He suggests using some of the technical measures to verify authenticity. Seadle (2011) concludes that analog and digital environments are different, and new means of assessing authenticity need to be developed for digital content.

In addition to the inherent properties of digital information, the technological environment in which digital objects are created and maintained represents a second area of challenge. This environment encompasses equipment and formats for generating and encoding digital information as well as any hardware and software platforms necessary for processing, storing, rendering, and transmitting it. There is a wide range of risks associated with digital technology, from hardware failures where a loss is sudden and catastrophic, to the obsolescence of formats or software, which can go unnoticed until a file cannot be rendered. Overall, technological threats can be grouped into those that result in:

- *Physical loss*, damage, or decay of digital objects
- *Inability to access* digital objects due to technological obsolescence

Brown (2013) outlines a number of threats and the ways they endanger the integrity, reliability, and usability of preserved objects (pp. 202–206). A long list of risks to integrity ranges from accidental deletion to software failures.

Technological failures as preservation challenges in the context of digital libraries are presented in a case study of Chronicling America (Littman, 2007). Chronicling America, part of the National Digital Newspaper Program (NDNP), was a collaborative initiative aimed at digitizing American historic newspapers. Littman (2007) reports on the actual preservation threats encountered in the process of constructing a repository to provide for the long-term preservation of the digital objects generated as a result of this large-scale digitization project. The project team, building a repository for both access and preservation, experienced a number of challenges, including:

- Media failures, particularly problems with portable hard drives that were used to transfer files between the partner institutions; fixity checks performed on the files readily caught the problems.
- Hardware failures—a number of hard drive failures were encountered in the storage system; this issue was addressed by having an array of hard drives in the storage system.
- Software failures were experienced at different stages of ingesting digital objects into the repository.
- Operator errors, which represented the most serious threat as it resulted in the deletion of some files from the repository system.

This case study demonstrates that technological failures and human errors represent very real and serious threats to preserving digital content. The risks can be mitigated by careful planning and the implementation of a number of preservation strategies, such as fixity checking and using multiple storage media.

Technological obsolescence has been recognized as a major challenge since the early days of digital preservation (DPM Tutorial, 2003–15; Jones and Beagrie, 2008; Rothenberg, 1995; Waters and Garrett, 1996). Physical storage media, data formats, hardware, and software all become obsolete over time, posing significant technical challenges to preserving digital objects. Technological progress introduces innovations and improvements in formats and computer platforms but simultaneously deems older versions obsolete. Digital objects created with the older generation of technology may be intact but inaccessible because of the lack of functioning software to render them. Technological obsolescence poses threats to long-term access to digital objects. It impacts several components of the technological environment:

- *Storage media obsolescence* is a serious concern because of rapid changes in the storage technology. As Brown (2013) comments, "no computer storage medium can be considered archival, irrespective of its physical longevity—technological obsolescence is inevitable" (p, 222). Physical degradation of hard disks and magnetic tape used for storage poses a threat as well, but obsolescence often occurs long before deterioration of media becomes a problem (Jones and Beagrie, 2008). Newer technology offer increases in storage capacity but at the same time requires copying digital objects into new media.
- *Hardware and software obsolescence* is an inevitable result of technological advancement, part of the environment of ongoing change (Thibodeau, 2012). Constant upgrades to hardware, operating systems, and software applications bring improved speed and functionality, but also cause incompatibility and the inability to render objects created with older platforms. Software and hardware obsolescence are interrelated. New computers may not support older versions of the software necessary for executing files. New software may not run on legacy hardware.
- *File format obsolescence* has been recognized as one of the major threats to preservation (Abrams, 2004; DPM Tutorial, 2003–2015; Jones and Beagrie, 2008). Formats provide structures for encoding and decoding bistreams but can be superseded by newer specifications. The threat of format obsolescence has been addressed by the development of tools for format identification (DROID) and validation (JHOVE). Format registries, such as PRONOM, identify access tools and migration pathways for converting legacy formats. Rosenthal (2010) argues that format obsolescence is not as a significant threat as it was previously assumed because of more mature technology and the availability of open source formats. As discussed in Chapters 3 and 4, the guidelines in the digital library field recommend selecting nonproprietary, well-documented, and standardized formats. Wide adoption of the guidelines in digital library practice contributes to minimizing the risks of format obsolescence. The Library of Congress (2013) maintains a list of formats and provides a review of sustainability factors.

DIGITAL PRESERVATION GOALS

The ultimate goal of digital preservation is to ensure long-term access to digital content, but it needs to be considered in light of the challenges related to the unique characteristics of digital objects as well as the risks of technological failures and obsolescence. Furthermore, as the research in the digital library field indicates, the objectives of digital preservation should also encompass contextual and semantic aspects related to provenance, context of creation, and use in order to render preserved objects as authentic, meaningful, and useful information (Beaudoin, 2012a; Ross, 2012). Thus, the goals for digital preservation are twofold: (1) focusing on bit preservation and ensuring the integrity of digital objects, and (2) maintaining sources of representation information and ensuring the authenticity of preserved objects. Brown (2013) recognizes the dual nature of preservation activities and identifies a number goals related to maintaining the integrity of data objects as well as those that focus on dimensions of authenticity, specifically the reliability and usability of preserved information. This list is expanded by the goals identified in the digital library contextual research and includes:

- Maintaining integrity of digital objects by protecting them from alteration and corruption
- Protecting digital objects from media failure, physical loss, and technological obsolescence
- Ensuring that digital content can be uniquely and persistently identified
- Maintaining the semantic meaning of objects and the relationships between objects and collections

- Maintaining documentation on provenance and curatorial process
- Providing context of creation and use

Meeting preservation goals requires careful planning, developing a preservation policy, establishing a robust technological infrastructure, and practicing active management of digital assets.

DIGITAL PRESERVATION STRATEGIES

The active approach to digital preservation means not only an ongoing attention to the integrity and authenticity of digital information, but also constant monitoring of the technological environment to ensure that the digital objects can be accessed and reused in the future. Digital preservation practitioners should emphasize prevention rather than recovery. Although techniques to recover inaccessible files are being developed in digital archaeology and digital forensics, those approaches, as Brown (2013) points out, should be used as a last resort or an emergency measure. Digital forensics tools and techniques are useful in an archival practice that deals with the acquisition of born digital legacy materials. In digital library practice, digital forensics can assist with the recovery of born digital or digitized objects stored on removable media, and can help with the process of transferring them into more sustainable preservation environments (Lee et al., 2012). Again, forensics techniques should be used as an exception rather than a norm. Digital forensics tools are also very useful in determining and establishing authenticity and provenance.

A number of preservation strategies have been developed as preemptive measures to mitigate the risks and challenges associated with digital technology. There is no single approach that would provide a universal solution; rather, an appropriate strategy or a combination of them need to be selected depending on the changes in the technological environment and the types of objects that need to be preserved. The preservation strategies that address risks associated with technological failures and obsolescence include:

- *Bitstream copying*, known as “backing up your data,” refers to making multiple exact copies of digital objects. Bitstream copying is not a long-term preservation strategy; rather it serves as a preventive measure, protecting data from media failures and physical loss (DPM Tutorial, 2003–15).
- *Refreshing* mitigates the risk of media storage obsolescence. It involves copying files from one storage device to another. There should be no detectable change in the bitstream configuration during the refreshing process.
- *Migration* is undertaken in response to technological obsolescence, either format or hardware and software obsolescence. It involves the periodic transfer of files from one hardware/software configuration to a newer platform. Objects encoded in formats that are at risk of becoming obsolete need to be transformed into new target specifications. Brown (2013) emphasizes that the migration process might result in some information loss, as not all properties of the original object may be transformed or supported by the target format.
- *Normalization* is a form of format migration undertaken at the point of capturing or ingesting into a repository. The goal of normalization is to transform data into open and consistent formats or to minimize the number of managed formats in a repository. In the context of digital collections, normalization often takes place when images captured with digital cameras are transformed from proprietary raw formats into a standard TIFF format.

- *Emulation* represents a different strategy to combat technological obsolescence. Rather than transforming digital objects into new formats, emulation keeps digital objects in their original form, but reconstructs the functionality of an obsolete platform, usually through the use of emulation software. Emulation is often used in the preservation of games but can also be applied to the preservation of complex multimedia objects in digital libraries.

DIGITAL PRESERVATION IN DIGITAL LIBRARY PRACTICE

Digital libraries share access and preservation goals. Digital libraries collect, manage, and provide access to cultural and scientific resources to the current community of users, but an equally important part of their mission is the preservation of valuable digital content for the long-term (Candela et al., 2007). In practice, the balance is often tipped toward access and building online collections, with digital preservation activities often delegated to an IT department or outsourced to a service provider. Many stakeholders, including library administrators, do understand and support building digital collections as a form of expanding access, meeting user expectations, and increasing the online visibility of their institutions. Gaining full support for digital preservation as an equal component of digital library programs, however, is more difficult, especially at institutions with limited resources.

Brown (2013) argues that awareness of digital preservation is growing in the professional community. A lack of financial support and a lack of technical expertise are cited as major obstacles for establishing and implementing digital preservation programs. A study of 72 research libraries, members of the Association of Research Libraries (ARL), found that 39.3% of the surveyed institutions had no digital preservation system in place (Banach and Li, 2011). A case study conducted at CUNY Queens College demonstrates that small cultural heritage institutions not only lack an infrastructure for preserving digital content created as a result of digital projects, but don't even have preservation plans (Dolan-Mescal et al., 2014). However, it needs to be noted here that the state of digital preservation practice in the digital library field varies. While smaller institutions still struggle with developing preservation plans, national libraries and large research libraries have well-established and robust programs.

A sustainable digital library program requires developing an institutional approach to digital preservation and establishing a policy of commitment to the long-term maintenance of digital objects and collections. Developing a long-term preservation policy is a broader organizational issue, as a policy needs to encompass not only digital content created and collected as part of digital library programs, but also born digital archival acquisitions and other institutional digital assets. Preservation policies define how to manage digital assets to prevent the risk of content loss or damage. The process of developing an institutional digital preservation policy involves multiple steps, which includes stating the objectives, appraising and selecting content, assessing risks, outlining the scope of preservation actions, identifying resources and responsibilities, and establishing the requirements for building a technical infrastructure. A number of recent publications on digital preservation offer useful guidance on developing a policy (Brown, 2013; Corrado and Moulaison, 2014; Harvey, 2010). *Digital Preservation Policies Study* provides a model for developing institutional policies and a range of exemplars from the cultural heritage and research institutions (Beagrie et al., 2008). A recent case study presents the process of developing an organizational policy for digital preservation at the Ohio State University Libraries (Noonan, 2014).

A digital preservation plan in the context of digital libraries needs to be prepared at the very beginning of a project, ideally as part of a broader institutional policy. The plan can rely on the institutional infrastructure but may also call for new resources and solutions. Preservation of digital library content requires:

- Creating an institutional digital repository, or participating in a shared preservation repository, or using a reliable preservation service
- Conducting an inventory of the existing digital assets and verifying their integrity
- Defining the levels of preservation (long-term, medium-term, and short-term)
- Establishing a policy for data transfer, backup, refreshment, and migration
- Recording preservation metadata using standards, such as PREMIS and METS
- Maintaining project documentation and preservation metadata to support identification, access, and preservation process
- Maintaining the relationships between archival master files, access files, and metadata

Digital projects, if undertaken according to best practices and guidelines, produce valuable digital assets. The *Framework of Guidance for Building Good Digital Collections* states as one of its principles: "a good object is preservable" and provides a series of guidelines for creating digital objects with digital preservation in mind (NISO, 2007, p. 48). In addition to digital master files that are the primary focus of digital preservation actions, digital projects also generate a considerable amount of descriptive metadata and contextual information. In practice, the relationships between those various components are not always well maintained and preserved. The current digital library environment does not integrate access and preservation requirements. Most content management systems used for building digital collections present access files and metadata, but do not provide preservation functions. In a best-case scenario, digital preservation is managed by a separate repository system or outsourced to service providers. A study of open source repository systems conducted in 2012 found that the support for metadata standards and preservation functions varies between systems and is particularly limited for digital library multimedia objects (Madalli et al., 2012). Maintaining the relationships between multiple versions of digital objects, metadata, and curatorial documentation in this complex environment remains one of the major challenges for preserving digital content.

PRACTICAL GUIDELINES

This section provides practical recommendations for preserving master files on bit-level, but does not address the issues of maintaining associated metadata and preserving the relationships between the various entities in the digital library environment. This is an area that requires more attention and research, and hopefully useful guidelines and best practices will emerge in the near future. Bit preservation is a useful starting point but represents only a subset of preservation activities (Johnston, 2010). The technical guidelines for creating high-quality digital assets are discussed in more detail in Chapters 3 and 4. In light of discussions on technological obsolescence, it is worth reiterating that master files should be created with a use-neutral approach and saved in open, standard, and widely accepted formats. They should be saved uncompressed and follow an established file-naming convention.

Master files need to be stored in a reliable and secure preservation system. A decision about selecting a dedicated repository system needs to be made prior to undertaking a digital project in light of project goals and the established preservation plan. A number of factors have to be considered in selecting an appropriate solution, including the size of the project, the infrastructure available in-house, staff technical expertise and skills, and cost. The solutions range from institutional or shared repositories to hosted preservation services. The section on Digital Preservation Technology provides an overview of the options currently available. In practice, an organization may select a repository and a combination of a networked drive and/or removable media for backup. If a repository system is implemented, some tasks, such as integrity checking or format validation and normalization, can be automated.

The following guidelines for bit-level preservation of digital master files are based on the recommendations included in the digital preservation handbooks and practical guides (Brown, 2013; DPM Tutorial, 2003–15; Harvey, 2010; Jones and Beagrie, 2008).

- Create multiple copies (minimum three) of master files
- Calculate checksums on the file level and use them for integrity checking
- Store checksum data separate from master files
- Ensure that sufficient identification and representation information is stored with files
- Store at least one copy in a separate geographical location
- Ensure that all objects are stored on a minimum of two different reliable storage media
- Create an inventory of all files before moving them into a repository/storage system
- Copy or migrate digital objects to new media at regular intervals
- Practice active preservation by checking and verifying archival files regularly.

DIGITAL PRESERVATION TECHNOLOGY: STANDARDS AND REPOSITORIES

"Technology advances, while sure to present new challenges will also provide new solutions for preserving digital content" (Arms, 2000, para. 36). Arms discusses the challenges in preserving digital content at the National Digital Library Program (NDLP) of the Library of Congress in the early phase of digital preservation. Her quote indicates a great amount of uncertainty about the future of digital preservation but also some hope that technology will offer new solutions. Fifteen years later, we can definitely talk about some progress in preserving digital resources with more stable digital formats, preservation metadata standards, and trusted repositories in place. McGovern (2007) analyzed the first decade of digital preservation activities in terms of balancing the fundamental components of digital preservation (organization, resources, and technology), represented in the three-legged stool metaphor (see Fig. 9.1). She notices considerable progress in all three areas, but especially in organization, with a strong development of preservation policies.

The first decade, marked by the publication of the influential report, *Preserving Digital Information* (Waters and Garrett, 1996), established a basis for building standardized preservation systems and services. The concept of trusted repositories and two fundamental standards, OAIS reference model and PREMIS preservation metadata, were all introduced in the early 2000s. The first repository platforms, DSpace and Fedora, were also developed around that time and have been widely adopted. Format registries, such as PRONOM, and a range of tools for integrity checking, an automatic file format identification tool (DROID—Digital Object Record Identification), and an object characterization tool (JHOVE—JSTOR/Harvard Object Validation Environment) were also introduced during the first decade.

The most recent period is characterized by the widespread development of operational digital preservation services. The new generation repository software incorporates the relevant standards and tools, enabling the building of more reliable preservation systems. In addition, the current environment offers a number of options for selecting preservation approaches, from institutional repositories to shared or hosted solutions. The models with cloud-based services, such as DuraCloud or Preservica, make digital preservation more affordable and accessible to smaller cultural heritage institutions.

STANDARDS

The development and adoption of open standards proved to be critical to progress in digital preservation. Conceptual frameworks and metadata standards provide a theoretical foundation for developing reliable preservation systems and services. Two standards that have been recognized as particularly influential are the OAIS reference model and PREMIS metadata standard.

OAIS reference model is a high-level standard that provides a conceptual framework and consistent terminology for developing and maintaining archival information systems (Lee, 2010). The major purpose of the model is "to facilitate a much wider understanding of what is required to preserve and access information for the long term" (CCSDS, 2012, p. 2.1). It was developed by the researchers at the Consultative Committee for Space Data Systems (CCSDS) in 2001 and became an ISO standard in 2002. The model identifies the key players in the information environment, including Producers, Managers, and Consumers. In defining Information Object, it makes a distinction between Data Object (sequence of bits) and Representation Information. Data Object is interpreted with the associated Representation Information, yielding a useful and meaningful Information Object. This distinction is important in the context of archival information systems that need to support preservation of bits as well as the maintenance of Representation Information.

In addition to defining informational concepts, the Reference Model provides a functional layout of an archival system, identifying six main entities (Preservation Planning, Administration, Ingest, Data Management, Archival Storage, and Access), and the way that information flows between them (see Fig. 9.2). It addresses both the access and preservation aspects of ingesting digital objects and associated descriptive information into a repository for long-term storage. Lee (2010) notes that many aspects of the model rest on the distinction between the Submission Information Packages (SIP) received from Producers, the Archival Information Package (AIP) generated from SIPs upon ingest and managed by archives, and the Dissemination Archival Package (DIP) accessed by Consumers. The OAIS model provides a foundation for building and implementing standard and interoperable repository systems.

PREMIS (Preservation Metadata: Implementation Strategies) is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability (Library of Congress, 2015). It specifies the metadata units that a repository needs to maintain core preservation functions. The standard was developed by the OCLC/RLG working group in 2005. Its current development is managed by the Library of Congress in conjunction with the PREMIS Editorial Committee. The standard consists of a Data Model and Data Dictionary. An XML schema is also available to support the implementation of the data dictionary in digital repository systems. Version 2.2 of the PREMIS Data Dictionary is currently available though the Library of Congress (PREMIS Editorial Committee, 2012).

The Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process" (PREMIS Editorial Committee, 2012, p. 3). Preservation metadata spans a number of metadata types, including descriptive, structural, technical, and administrative. The Data Dictionary places a strong emphasis on the documentation of digital provenance (the history of an object) and the documentation of relationships, especially relationships among different objects within the preservation repository.

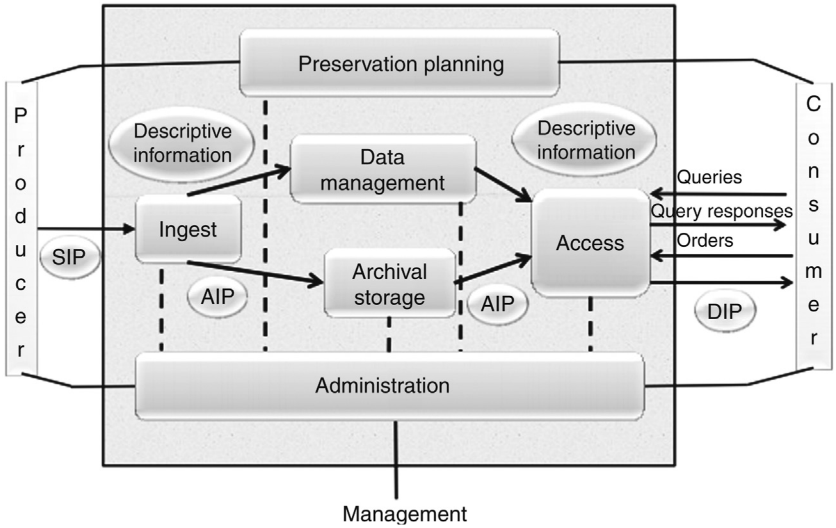

FIGURE 9.2 OAIS Functional Entities (CCSDS, 2012)

PREMIS standard provides a simple data model to organize the semantic units defined in the Data Dictionary and to encourage a shared way to organize preservation metadata (Dappert and Enders, 2010). The following entities are defined in the Data Model:

- Intellectual Entity: a set of content that is considered a single intellectual unit for purposes of management and description, for example, a particular book, map, photograph, or database.
- Object (or Digital Object): a discrete unit of information in digital form.
- Event: an action that involves or impacts at least one Object or Agent associated with or known by the preservation repository.
- Agent: person, organization, or software program/system associated with Events in the life of an Object or with Rights attached to an Object.
- Rights: assertions of one or more rights or permissions pertaining to an Object and/or Agent (PREMIS Editorial Committee, 2012, p. 6).

Fig. 9.3 demonstrates the entities in the PREMIS data model and the relationships between them.

The PREMIS Data Dictionary defines semantic units, not metadata elements. As Caplan (2009) explains, PREMIS does not specify how metadata should be represented or implemented in a repository system; it only defines what the system needs to know and should be able to export to other systems. Semantic units describe properties of digital objects and their contexts or the relationships between them. Each semantic unit defined in the Data Dictionary is mapped to one of the entities in the Data Model. For example, the Object entity is described by a number of semantic units, such as objectIdentifierType or objectIdentifierValue, defined as mandatory (M) and nonrepeatable (NR). Semantic units are presented in a hierarchical structure.

PREMIS may be implemented in a variety of ways, which offers the potential of broad application across a wide range of preservation contexts. Guenther (2010) explores using PREMIS within a METS

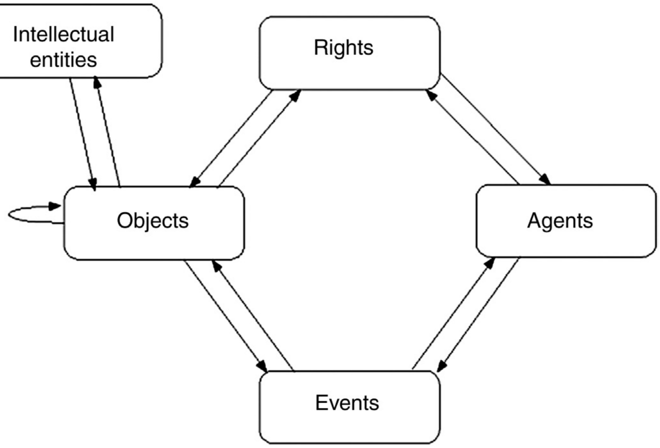

FIGURE 9.3 The PREMIS Data Model (PREMIS Editorial Committee, 2012)

container and points to the benefits of using the two metadata standards together. A number of research studies investigate implementation of PREMIS in practical digital library settings. Alemneh (2009) examined the barriers to adopt PREMIS in cultural heritage institutions. Donaldson and Conway (2010) present a case study in which PREMIS is implemented in the Florida Digital Archive. Findings point to the iterative nature of the implementation process and to the necessity of adopting the standard in the local repository. Donaldson and Yakel (2013) investigated the adoption of PREMIS by several organizations registered with the Library of Congress PREMIS Implementers Group. The researchers confirm the findings of the earlier studies, indicating that many institutions have made the decision to adopt PREMIS, but few have fully implemented it.

DIGITAL REPOSITORIES

Digital repositories are information systems that ingest, store, manage, preserve, and provide access to digital content. The OAIS model provides a conceptual foundation for designing standard-compliant repositories. Digital repositories are a relatively new phenomenon that emerged in the early 2000s. A concept of trusted digital repositories has been advanced to ensure high-level preservation services for all types of repositories. There are several repositories types, including institutional, disciplinary, government, and centralized repositories, which aggregate content from several subsidiary repositories. According to the Directory of Open Access Repositories, most of the content in open access (83.2%) is available through institutional repositories (OpenDOAR, 2015). As Lynch (2003) observes, institutional repositories offer an essential infrastructure for scholarship in the digital age and a potential to revolutionize scholarly communication. He also notes that a key part of the service is to manage technological change and the migration of digital content from one set of technologies to the next.

Institutional digital repositories serve multiple purposes. Their primary goal is to support scholarly communication and provide open access to articles, dissertations, and research data. In addition, they provide platforms for storing and preserving the digital master files created as a result of digitization projects. The boundaries between a repository and digital libraries are sometimes blurred, as repositories also host digital collections for access. The combination of access and preservation functions poses significant challenges. McGovern and McKay (2008) investigated the juncture of institutional repository implementation and digital preservation programs and provided a set of recommendations for leveraging the benefits of institutional repositories to strengthen long-term preservation. A number of research studies examine the current practices of digital preservation in the institutional repository environment (Banach and Li, 2011; Kunda and Anderson-Wilk, 2011; Neatrour et al., 2014; Oehlerts and Liu, 2013).

Trusted digital repositories perform preservation functions. This notion was first introduced in the seminal report, *Preserving Digital Information* (Waters and Garrett, 1996). The authors emphasize the role of trust in managing the identity, integrity, and quality of digital information in archival systems and recommend developing a process of certification. The concept of a trusted digital repository was fully articulated in another foundational report, *Trusted Digital Repositories: Attributes and Responsibilities*, prepared by a RLG/OCLC working group (Beagrie et al., 2002). A trusted digital repository is defined as one "whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future" (Beagrie et al., 2002, p. 5). In order to gain recognition as "trusted," a repository has to have certain attributes that ensure the reliability and authenticity of stored information. The RLG/OCLC group outlines the following characteristics of sustainable digital repositories:

- Accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users
- Have an organizational system that supports not only long-term viability of the repository but also the digital information for which it has responsibility
- Demonstrate fiscal responsibility and sustainability
- Design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it
- Establish methodologies for system evaluation that meet community expectations of trustworthiness
- Be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly
- Have policies, practices, and performance that can be audited and measured (Beagrie et al., 2002, p. 5)

In addition, the RLG/OCL report discusses methods and strategies for the certification of trusted digital repositories. A regular cycle of certification and audit is recommended for digital repositories to remain trustworthy. The process of certification has gained considerable attention in the last decade, and a number of standards and checklists have emerged, such as TRAC (The Trustworthy Repositories Audit & Certification Checklist), superseded by the ISO 16363:2012—Audit and Certification of Trustworthy Digital Repositories international standard. A range of tools have been developed in Europe, including nestor, DRAMBORA, Platter, and Data Seal of Approval. The recent publications on digital preservation provide an overview of these tools (Brown, 2013; Corrado and Moulaison, 2014).

PRESERVATION REPOSITORY SOFTWARE

Building operational digital repositories requires a technical infrastructure and dedicated software to support the functions of ingesting, storing, managing, preserving, and providing access to digital content. Digital repository software is an area of active development, with several open source and proprietary solutions emerging. The benefits and limitations of open source versus proprietary software are discussed in Chapter 6 in the context of digital library management systems (DLMS).

Many early open source solutions, including DSpace (http://www.dspace.org/), EPrints (http://www.eprints.org/), and Fedora (http://fedorarepository.org/) were developed as part of the Open Access (OA) movement to provide platforms for the open dissemination of scholarly publications. Comparative studies of the early generation of open source repository software that focused on DSpace, EPrints, and Fedora found varying levels of support for preservation functions. In the examined group, Fedora demonstrates the strongest support for features essential to digital preservation (Fay, 2010; Madalli et al., 2012). Fedora is one of the most versatile solutions among open source software, as it provides support for building digital collections for access and performs digital preservation functions. It is often integrated with the new generation of open source systems, such as Hydra and Islandora. DSpace, Fedora, Hydra, and Islandora as multipurpose digital library management systems, are described in more detail in Chapter 6. The new generation of open source preservation repository software, including such software as Archivematica and DAITSS, is built in compliance with the OAIS functional model and implements active preservation strategies.

The following list provides a brief overview of selected open source and proprietary solutions. The review of the software is by no means comprehensive, nor is it meant to serve as a recommendation or evaluation. There are many other options available, especially in the open source category. Archivematica represents a more recent development in the open source category. Rosetta is an example of a proprietary software that is often used in conjunction with digital library management systems.

- *Archivematica* (https://ww.archivematica.org/en/) is an open source, standards-based, integrated suite of software tools designed to process digital objects from ingest to access. Archivematica version 0.10 was released in 2013; as of Nov. 2015, version 1.4 was available. Its functionality is based on the OAIS reference model. It supports a range of preservation standards, including Dublin Core, METS, and PREMIS metadata standards and incorporates the Library of Congress BagIt file packaging format and other task-specific applications. Archivematica can be integrated with digital library management or preservation systems, including AtoM, DSpace, CONTENTdm, Islandora, LOCKSS, and DuraCloud.
- *Rosetta* (http://www.exlibrisgroup.com/category/RosettaOverview) is a proprietary software developed by Ex Libris in collaboration with the National Library of New Zealand. It is intended for managing and preserving digital library resources as well as research data. Rosetta is compliant with the OAIS reference model. It serves as an integrated solution with one central repository, which can be synchronized with many other applications. It is used by a number of libraries in Europe, the United States, and New Zealand. University of Utah's J. Willard Marriott Library uses Rosetta as a preservation system alongside CONTENTdm, which is used for building digital collections (Neatrour et al., 2014).

PARTNERSHIPS AND HOSTED SERVICES

Operational digital repositories can be developed in house with a custom-built approach or by using an open source and proprietary software. Developing and managing an institutional repository requires a significant investment of resources and expertise, and some institutions decide to share the burden and participate in cooperative programs or outsource the preservation functions to a hosted preservation service.

Shared repositories are preservation repositories created through the membership or partnerships of libraries and archives to store and share their digital collections. They represent collaborative efforts to advance digital preservation by reducing cost and sharing expertise. Two models of collaborative initiatives have emerged: (1) centralized infrastructure and services supported by membership fees, and (2) infrastructure and services distributed geographically and shared among participating members. HathiTrust and Portico are two well-known examples of the centralized model, while MetaArchive represents a successful case of collaboration in a distributed environment. Walters and Skinner (2010) argue that interinstitutional repositories provide a sustainable approach to digital preservation. Participating members not only distribute costs but also have an opportunity to leverage expertise across a diverse body of institutions.

- *HathiTrust* (http://www.hathitrust.org/) was launched in 2008 as a collaborative initiative of major research libraries to ensure that the cultural record is preserved and accessible long into the future. Initially, HathiTrust was created to provide a preservation platform for storing a large volume of items digitized through mass digitization projects, such as the Google Book Project and Open Content Alliance (OCA). HathiTrust is also a large-scale digital library, and as such is described in more detail in Chapter 1. As Christenson (2011) emphasizes "at the heart of HathiTrust is a shared secure digital repository owned and operated by a partnership of major research libraries" (p. 95). Currently, there are more than 60 partners in HathiTrust, and membership is open to institutions worldwide. The HathiTrust repository now contains the largest collection of digital volumes outside of Google Books. It represents an example of a "light archive," meaning that the repository also functions as a digital library and provides access to some of their collections. Access is restricted to items under copyright.
- *MetaArchive* (http://www.metaarchive.org/) was established in 2003 as a "community-owned and community-operated distributed digital preservation network" (Walters and Skinner, 2010, p. 264). MetaArchive works as a cooperative, with members paying membership fees but also contributing in-kind with staff, technology, and space. As Walters and Skinner (2010) note, these in-kind contributions keep preservation costs low. The cooperative includes members from over 50 institutions in 13 states and 3 countries. The distributed model relies on the open source software, LOCKSS (Lots of Copies Keep Stuff Safe), developed at Stanford University. Member institutions host servers within their own organizational infrastructures, while LOCKSS software enables connecting the servers in a secure network and replicating the content for preservation purposes. Servers are selected and assigned to content on the basis of their widespread geographical distribution (MetaArchive, 2015). MetaArchive is an example of a dark archive, meaning there is no public access to it.

Hosted preservation services, sometimes referred as "preservation-as-a-service," are digital repositories maintained by nonprofit organizations that provide archiving services for a fee. The pricing structure is usually based on the number of digital objects and/or size in terabytes of the collection(s). There are a number of hosted services available on the market. OCLC DigitalArchive and DuraCloud serve primarily cultural heritage institutions.

- *OCLC DigitalArchive* (http://www.oclc.org/digital-archive.en.html) is an archiving solution for institutions that prefer to outsource the preservation of their digitized assets to a nonprofit organization. OCLC DigitalArchive provides a hosted preservation option to institutions involved in digitization that decide not to build their own repository or participate in collaborative initiatives. OCLC can securely store digital master files for a fee. OCLC DigitalArchive is convenient for institutions using CONTENTdm software, as the ingest of digital master files is integrated with CONTENTdm functionality. There is an additional fee for archiving on top of a CONTENTdm subscription. Like MetaArchive, OCLC DigitalArchive is also a dark archive.

- *DuraCloud* (http://www.duracloud.org/) is a hosted, cloud-based preservation service offered by DuraSpace, a not-for-profit organization founded in 2009 by the stakeholders of DSpace Foundation and Fedora Commons. DuraCloud is one of several open source services supported by DuraSpace. It is focused on providing preservation support and access services for academic libraries and other cultural heritage organizations. In addition to cloud storage, DuraCloud provides services that enable digital preservation, data access, transformation, and data sharing.

Digital technology presents a paradox in the realm of digital preservation. On one hand, it poses a number of challenges and risks for preserving digital content because of technological failures and obsolescence. On the other hand, technological progress offers improved solutions and tools for maintaining digital objects in the long-term in reliable and trusted repository systems. Measurable progress has been made in establishing a conceptual framework for repository systems, developing validation tools, and building a technical infrastructure. Digital libraries are part of this active development, but in practice, insufficient attention and resources are allocated to digital preservation priorities, especially at smaller institutions. Digital preservation practice in the digital library field is still in the early stages of development. While practical guidelines are available for bit-level preservation, more research is needed on capturing and maintaining representation and contextual information. National libraries and large research libraries have taken a lead in establishing preservation programs and developing new solutions and best practices. Hopefully, with an increasing awareness of the importance of digital preservation and the diffusion of best practices and tools, sustainable models will be adopted in the mainstream digital library practice in the near future.

REFERENCES

Abrams, S.L., 2004. The role of format in digital preservation. Vine 34 (2), 49–55.

ALA ALCTS: Association for Library Collections & Technical Services. Division of the American Library Association, 2007. Definitions of digital preservation. Available from: http://www.ala.org/alcts/resources/preserv/defdigpres0408.

Alemneh, D.G., 2009. Barriers to adopting PREMIS in cultural heritage institutions: an exploratory study. In: Archiving Conference, Vol. 2009, No. 1. Society for Imaging Science and Technology, pp. 113–118.

Arms, C.R., 2000. Keeping memory alive: practices for pressing digital content at the National Digital Library Program of the Library of Congress. RLG DigiNews, 4 (3). Available from: http://webdoc.gwdg.de/edoc/aw/rlgdn/preserv/diginews/diginews4-3.html.

Ashenfelder, M., 2011. DigitalPreservationEurope. The Signal: Digital Preservation. Library of Congress blog. (October, 14). Available from: http://blogs.loc.gov/digitalpreservation/2011/10/digitalpreservationeurope/.

Banach, M., Li, Y., 2011. Institutional repositories and digital preservation: assessing current practices at research libraries. D-Lib Mag. 17 (5), 1.

Beagrie, N., Bellinger, M., Dale, R., Doerr, M., Hedstrom, M., Jones, M., Woodyard, D., 2002. Trusted digital repositories: attributes and responsibilities. Research Libraries Group (RLG) & Online Computer Library Center (OCLC) Report. Available from: http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

Beagrie, N., Semple, N., Williams, P., Wright, R., 2008. Digital preservation policies study. Joint Information Systems Committee (JISC). Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.214.9056&rep=rep1&type=pdf.

Beaudoin, J.E., 2012a. Context and its role in the digital preservation of cultural objects. D-Lib Mag., 18 (11), 1. Available from: http://www.dlib.org/dlib/november12/beaudoin/11beaudoin1.html.

Beaudoin, J.E., 2012b. A framework for contextual metadata used in the digital preservation of cultural objects. D-Lib Mag., 18 (11/12), 3. Available from: http://dlib.org/dlib/november12/beaudoin/11beaudoin1.print.html.

Berman, F., Lavoie, B., Ayris, P., Choudhury, G.S., Cohen, E., Courant, P., Dirks, L., Friedlander, A., Gurbaxani, V., Jones, A., Kerr, A.U., Lynch, C.A., Rubinfeld, D., Rusbridge, C., Schonfeld, R., Smith-Rumsey, A., Van Camp, A., 2010. Sustainable economics for a digital planet: ensuring long-term access to digital information. Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Available from: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Bradley, K., 2007. Defining digital sustainability. Lib. Trends 56 (1), 148–163.

Brown, A., 2013. Practical Digital Preservation: A How-To Guide for Organizations of Any Size. Neal-Schuman, Chicago.

Candela, L., Castelli, D., Pagano, P., Thano, C., Ioannidis, Y., Koutrika, G., and Schuldt, H., 2007. Setting the foundations of digital libraries: the DELOS manifesto. D-Lib Mag., 13 (3), 4. Available from: http://www.dlib.org/dlib/march07/castelli/03castelli.html.

Caplan, P., 2008. The preservation of digital materials. Lib. Technol. Rep. 44 (2), 1–38.

Caplan, P., 2009. Understanding PREMIS. Library of Congress. Available from: http://www.loc.gov/standards/premis/understanding-premis.pdf.

CCSDS: Consultative Committee for Space Data Systems, 2012. Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS. Available from: http://public.ccsds.org/publications/archive/650x0m2.pdf.

Chowdhury, G., 2010. From digital libraries to preservation research: the importance of users and context. J. Doc. 66 (2), 207–223.

Christenson, H., 2011. HathiTrust: a research library at web scale. Lib. Res. Tech. Serv. 55 (2), 93–102.

Cloonan, M.V., 2015. Preserving Our Heritage: Perspectives From Antiquity to the Digital Age. Neal-Schuman, Chicago.

Cloonan, M.V., 2007. The paradox of preservation. Lib. Trends 56 (1), 133–147.

Conway, P., 2007. The relevance of preservation in a digital world. Northeast Document Conservation Center (NDCC). Available from: https://www.nedcc.org/free-resources/preservation-leaflets/6.-reformatting/6.4-the-relevance-of-preservation-in-a-digital-world.

Conway, P., 2010. Preservation in the age of Google: digitization, digital preservation, and dilemmas. Lib. Q. 80 (1), 61–79.

Corrado, E.M., Moulaison, H.L., 2014. Digital Preservation for Libraries, Archives, and Museums. Rowman & Littlefield, Lanham, MA.

Dappert, A., Enders, M., 2010. Digital preservation metadata standards. Inf. Stand. Q. 22 (2), 156–166.

DCC: Digital Curation Centre, 2004–15. What is digital curation? Available from: http://www.dcc.ac.uk/digital-curation/what-digital-curation.

Dolan-Mescal, A., Farwell, M., Howard, S., Rozler, J., Smith, M., 2014. A digital file inventory of the Queens College Special Collections and Archives. OCLC Syst. Serv. 30 (2), 78–90.

Donaldson, D., Conway, P., 2010. Implementing PREMIS: a case study of the Florida Digital Archive. Lib. Hi Tech. 28 (2), 273–289.

Donaldson, D., Yakel, E., 2013. Secondary adoption of technology standards: the case of PREMIS. Arch. Sci. 13 (1), 55–83.

DPM Tutorial, 2003–15. Digital Preservation Management. Cornell University Library. Available from: http://www.dpworkshop.org/.

Fay E., 2010. Repository software comparison: building digital library infrastructure at LSE. Ariadne, 64. Available from: http://www.ariadne.ac.uk/issue64/fay.

Guenther, R., 2010. Metadata to support long-term preservation of digital assets: PREMIS and its use with METS. In: Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop. Association for Computing Machinery, ACM, New York, p. 14.

Harvey, R., 2010. Digital Curation: A How-To-Do-It Manual. Neal-Schuman, New York.

Higgins, S., 2008. The DCC curation lifecycle model. IJDC 3 (1), 134–140.

Higgins, S., 2011. Digital curation: the emergence of a new discipline. IJDC 6 (2), 78–88.

Johnston, L., 2010. Developing bit preservation services at the Library of Congress. In: Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop. Association for Computing Machinery, ACM, New York, p. 12.

Jones, M., Beagrie, N., 2008. Preservation Management of Digital Materials: A Handbook. Digital Preservation Coalition (DCP). Available from: http://www.dpconline.org/advice/preservationhandbook.

Kenney, A.R., McGovern, N.Y., 2003. The five organizational stages of digital preservation. In: Hodges, P., Bonn, M., Sandler, M., Wilkin, J.P. (Eds.), Digital Libraries: A Vision for the Twenty-First Century, A Festschrift to Honor Wendy Lougee. The University of Michigan Scholarly Monograph Series. Available from: http://quod.lib.umich.edu/s/spobooks/bbv9812.0001.001/--digital-libraries-a-vision-for-the-21st-century.

Kunda, S., Anderson-Wilk, M., 2011. Community stories and institutional stewardship: digital curation's dual roles of story creation and resource preservation. Portal Lib. Acad. 11 (4), 895–914.

Lavoie, B., Dempsey, L., 2004. Thirteen ways of looking at digital preservation. D-Lib Mag. 10 (7/8), 1082–9873.

Lazorchak, B., 2011. Digital preservation, digital curation, digital stewardship: What's in (some) names? The Signal: Digital Preservation. Library of Congress blog. (August 2). Available from: http://blogs.loc.gov/digitalpreservation/2011/08/digital-preservation-digital-curation-digital-stewardship-what%E2%80%99s-in-some-names/.

Lee, C.A., 2011. A framework for contextual information in digital collections. J. Doc. 67 (1), 95–143.

Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., Woods, K., 2012. BitCurator: tools and techniques for digital forensics in collecting institutions. D-Lib Mag. 18 (5), 3.

Lee, C.A., 2010. Open archival information system (OAIS) reference model. Encyclopedia of Library and Information Science 4020–4030. 3rd ed. Taylor & Francis, Abingdon, UK.

Lesk, M., 2014. Foreword. In: Corrado, E.M., Moulaison, H.L. (Eds.), Digital Preservation for Libraries, Archives, and Museums. Rowman & Littlefield, Lanham, MA, pp. xv–xvi.

Library of Congress, 2013. Sustainability of digital formats: planning for Library of Congress Collections. Available from: http://www.digitalpreservation.gov/formats/index.shtml.

Library of Congress, 2015. PREMIS: preservation metadata maintenance activity. Available from: http://www.loc.gov/standards/premis/.

Library of Congress, n.d. On the leading edge. Available from: http://www.digitalpreservation.gov/series/edge/index.html.

Littman, J., 2007. Actualized preservation threats: practical lessons from Chronicling America. D-Lib Mag. 13 (7), 5.

Lynch, C., 2000. Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. Washington, DC: Council on Library and Information Resources. Available from: http://www.clir.org/pubs/reports/pub92/lynch.html.

Lynch, C.A., 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. Portal Lib. Acad. 3 (2), 327–336.

Madalli, D.P., Barve, S., Amin, S., 2012. Digital preservation in open-source digital library software. J. Acad. Librar. 38 (3), 161–164.

McGovern, N., 2007. A digital decade: where have we been and where are we going in digital preservation? RLG DigiNews 11 (1), 1.

McGovern, N.Y., McKay, A.C., 2008. Leveraging short-term opportunities to address long-term obligations: a perspective on institutional repositories and digital preservation programs. Lib. Trends 57 (2), 262–279.

MetaArchive, 2015. Methodology. Available from: http://www.metaarchive.org/methodology.

NDSA: National Digital Stewardship Alliance, 2014. 2014 National Agenda for Digital Stewardship. Available from: http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf.

Neatrour, A., Brunsvik, M., Buckner, S., McBride, B., Myntti, J., 2014. The SIMP Tool: facilitating digital library, metadata, and preservation workflow at the University of Utah's J. Willard Marriott Library. D-Lib Mag., 20 (7), 2.

NISO Framework Working Group, 2007. A framework of guidance for building good digital collections. 3rd ed. Available from: http://www.niso.org/publications/rp/framework3.pdf.

Noonan, D., 2014. Digital preservation policy framework: a case study. EDUCAUSEreview, (July 28). Available from: http://www.educause.edu/ero/article/digital-preservation-policy-framework-case-study.

Oehlerts, B., Liu, S., 2013. Digital preservation strategies at Colorado State University libraries. Lib. Manag. 34 (1/2), 83–95.

OpenDOAR: Directory of Open Access Repositories, 2015. Open Access repository types—worldwide. University of Nottingham, UK. Available from: http://www.opendoar.org/index.html.

PREMIS Editorial Committee, 2012. PREMIS Data Dictionary for preservation metadata, version 2.2. Available from: http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

Rosenthal, D.S., 2010. Format obsolescence: assessing the threat and the defenses. Lib. Hi Tech. 28 (2), 195–210.

Ross, S., 2002. Position paper on integrity and authenticity of digital cultural objects. DigiCULT: Integrity and Authenticity of Digital Cultural Heritage Objects, 1 (August).

Ross, S., 2012. Digital preservation, archival science and methodological foundations for digital libraries. New Rev. Inf. Network. 17 (1), 43–68.

Rothenberg, J., 1995. Ensuring the longevity of digital documents. Sci. Am. 272 (1), 42–47.

Seadle, M., 2008. The digital library in 100 years: damage control. Lib. Hi Tech. 26 (1), 5–10.

Seadle, M., 2011. Archiving in the networked world: by the numbers. Lib. Hi Tech. 29 (1), 189–197.

Thibodeau, K., 2012. Wrestling with shaper-shifters: perspectives on preserving memory in the digital age. In: Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation, pp. 15–23. Available from: http://www.ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf.

Walters, T.O., Skinner, K., 2010. Economics, sustainability, and the cooperative model in digital preservation. Lib. Hi Tech. 28 (2), 259–272.

Waters, D., Garrett, J., 1996. Preserving digital information. Report of the Task Force on Archiving of Digital Information. Washington, DC: The Commission on Preservation and Access. Available from: http://files.eric.ed.gov/fulltext/ED395602.pdf.