CHAPTER

DIGITIZATION OF TEXT AND STILL IMAGES

3

INTRODUCTION TO DIGITIZATION

Digitization has contributed to building a critical mass of scholarly and cultural heritage resources in digital libraries. Nowadays, nearly all new information is created in digital format, and although the number of "born digital" library resources has been rapidly increasing, it still represents a small fraction of recorded knowledge. Historically, scholarly and cultural heritage resources have been recorded on a variety of analog formats, with paper being the dominant material for several centuries. Institutions within the cultural heritage community—libraries, archives, and museums—have also collected information resources recorded on other analog carriers including film and glass negatives, slides, and audio and videotapes. The inclusion of analog materials in digital libraries has required extensive conversion efforts. Since the early 1990s, digitization has been undertaken both by individual cultural heritage institutions and through collaborative initiatives. Mass digitization projects including JSTOR, the Google Book Project, and the Open Content Alliance initiative have contributed a significant body of digitized content. The process of conversion is far from complete, but digital libraries such as Europeana, the Digital Public Library of America, the Internet Archive, and HathiTrust already provide access to millions of digitized scholarly and cultural heritage objects. These large-scale digital libraries represent two decades of intensive digitization efforts undertaken by both smaller institutions and large initiatives like the Google Book Project.

Digitization is the process of creating digital representations of information resources recorded on analog carriers. In essence, digitization is "the conversion process of an analog signal or code into a digital signal or code" (Lee, 2001, p. 3). It involves sampling continuous patterns of physical media and converting them into binary streams of ones and zeros that can be processed and represented by digital devices. The term "digitization" is often used interchangeably with "scanning", but as Lee (2001) points out, there are distinct differences between the two. Scanning refers to the conversion of static textual and visual materials, and motion picture films. Digitization, on the other hand, is a broader concept that encompasses the conversion of all analog media, including sound, video, and 3D objects. Digitization, especially if undertaken for preservation purposes, aims at converting not only the informational content but also capturing unique characteristics of analog materials as much as possible. It is worth noting that digitization involves not only copying, but also transformation of the nature and functionality of information resources. While it might not always be possible to replicate all the attributes of physical objects, digitized versions facilitate new means of access and enable the use of materials that are not possible with the analog form.

Creating digital counterparts of analog scholarly and cultural heritage resources is a complex and challenging undertaking because of the variety of predigital formats and the evolving technological standards and best practices. For the purpose of digitization, analog resources are divided into static and time-based media (ALCTS, 2013; Puglia, 2011). Static media encompass textual resources such

Discover Digital Libraries. http://dx.doi.org/10.1016/B978-0-12-417112-1.00003-X

as manuscripts, printed books, newspapers, and journals as well as still images, including photographs, maps, and two-dimensional art works. Time-based or dynamic media include audio, video, and moving images. Puglia (2011) emphasizes that the complexity of the conversion process increases when we move from static to dynamic media. Static resources can be converted to digital format through an imaging process, while time-based media require playback devices in addition to conversion hardware and software. Digitization of 3D objects poses a different set of challenges. Constructing a digital representation of a three-dimensional artifact involves taking multiple still images with a digital camera from several viewpoints and stitching them together or the use of laser scanning technology that captures millions of points of measurement to represent the shape of 3D objects (Collmann, 2011; Surendran et al., 2009; Valentino and Shults, 2012).

The technological advancements of the 1990s set the stage for the digital conversion of printed materials and unique cultural heritage collections. Improvements in scanning technology, faster computer processing, and the lower cost of digital storage have enabled libraries and other cultural heritage institutions to move forward with the digitization of their analog collections. The growth of high-speed networks, especially of the World Wide Web, has allowed for the sharing of digitized collections with wider audiences. Cultural heritage institutions have primarily focused on the conversion of unique materials from their archival and special collections. The digitization of general collections of printed books and journals has required larger scale efforts and the cooperation of multiple institutions. JSTOR, a project initiated by the Andrew W. Mellon Foundation in 1995, undertook the conversion of back issues of scholarly journals. Mass digitization of books began in the mid-2000s with the Google Book Project and the Open Content Alliance taking the lead (Coyle, 2006; Leetaru, 2008).

Large-scale digitization of archives and special collections is challenging because of the unique and heterogeneous nature of the materials and the complexity of analog formats. Digitization of cultural heritage collections frequently requires attention to preservation issues and careful handling of fragile or rare items. The mass digitization of books undertaken in collaboration with commercial and nonprofit partners, however, has changed the digitization landscape and raised expectations for the digital conversion of unique cultural heritage materials (Erway, 2011; Erway and Schaffner, 2007). In recent years, there has been a growing interest in adopting rapid digitization approaches in order to increase the volume of digitized materials from unique collections in libraries, archives, and museums (Erway, 2011; Miller, 2013; Rinaldo et al., 2011; Sutton, 2012).

This chapter provides an overview of the underlying principles and technologies involved in the conversion of analog materials into digital form. The primary focus is on the digitization of cultural heritage resources. This chapter discusses the main reasons why libraries and archives engage in digital conversion and looks at strategies for undertaking digitization projects. Furthermore, it examines the steps, guidelines, technical specifications, and imaging equipment used in converting textual and photographic materials. The conversion of audio and video is discussed in Chapter 4.

RATIONALE AND STRATEGIES FOR UNDERTAKING DIGITIZATION

Libraries, archives, and museums embraced digitization relatively quickly since imaging technology became widely available in the mid-1990s. Major research and academic libraries as well as national libraries and archives led the first digitization initiatives. The 1998 survey of special collections affiliated with the Association of Research Libraries indicated that 89% of the major research libraries in the

United States were involved in digitization in 1998 (Panitch, 2001). An IFLA/UNESCO study demonstrated that as of 1999, 48% of surveyed national libraries worldwide had digitization programs (Gould and Ebdon, 1999). Two studies conducted by the Institute of Museum and Library Services (IMLS) indicated that smaller libraries, archives, and museums were initially lagging behind, but digitization activities increased in all types of cultural heritage institutions between 2001 and 2004 (IMLS, 2006). Conway (2008) notes that digitization of cultural heritage materials, especially historical photographs, has transitioned from "rarified experiments to nearly ubiquitous activity" (p. 94). Finally, a more recent OCLC survey of special collections in the United States and Canada reported that nearly all participating libraries have completed at least one digitization project and/or have an active digitization program for special collections (Dooley and Luce, 2010).

The IMLS and OCLC studies indicate a widespread adoption of digitization by cultural heritage institutions, although funding, staffing, and expertise are constant challenges. Why have libraries, archives, and museums embarked on digitization projects despite significant costs and the complexity of the conversion process? Kenney and Rieger (2000) provide an overall rationale and list two primary reasons for investing in digitization: (1) to accommodate the changing behavior of users in the digital environment and (2) to maintain the relevance of analog resources for teaching, research, and scholarship. The authors comment that "changing user behavior may jeopardize these resources and their stewardship" (Kenney and Rieger, 2000, p. 1). In examining the role of digitization and preservation a decade later, Conway (2010) observes that "in the age of Google, nondigital content does not exist, and digital content with no impact is unlikely to survive" (p. 64).

WHY DIGITIZE: ACCESS AND PRESERVATION

The question "why digitize?" has been posed since the early days of digitization projects (Smith, 1999). Cultural heritage institutions focus on special and archival collections and specific goals related to curating, access, and the new functionality of digitized materials (Besser, 2003; Conway, 2000; IMLS, 2006; Lee, 2001; Lopatin, 2006; Smith, 2001). The discussion on motivation and rationale has centered on two goals:

- Increasing access to library, archival, and museum collections
- Preserving valuable, fragile, and deteriorating materials

Institutions participating in the IMLS survey in 2006 identified additional goals. Museums gave more weight to making information about their collections accessible to artists, scholars, students, teachers, and the public, while academic and public libraries highlighted providing access to materials via the web, minimizing damage to original materials, and increasing interest in the institution. Access and preservation, however, were consistently ranked as top goals across all institutions (IMLS, 2006).

Increased access to unique cultural heritage materials has indeed been acknowledged as the major benefit of digitization (Cohen and Rosenzweig, 2006; Daigle, 2012; Smith, 1999). "Digitization is access—lots of it," emphasizes Smith (1999, p. 7). Daigle (2012) comments, "open access has been transformative to researchers who are no longer required to travel to the physical location of primary source material" (p. 252). The added value of digitization, however, goes beyond the mere convenience of remote access to surrogate copies of original documents. Researchers point to the advantages of digital image enhancement, the ability to bring together dispersed research materials, and the potential to reach audiences across social and economic boundaries (Besser, 2003; Kenney and Rieger, 2000;

Smith, 1999, 2001). The capabilities of full-text searching and cross-collection indexing afford new ways of exploring and using traditional materials (Conway, 2000; Kenney and Rieger, 2000; Lesk, 2004).

Digitization has removed physical barriers to the discovery and use of rare and fragile resources and those recorded on difficult-to-access formats. Access to rare manuscripts, photographs, maps, archival documents, and museum objects has often been limited because of their value and/or fragile nature. Digitization not only expands the reach of these materials to researchers, students, and the general public, but in many cases it enhances the visual quality of faded and illegible documents (Smith, 1999). In addition, advancements in imaging have enabled the conversion of visual materials recorded on difficult-to-access formats such as film negatives and slides. Digitization offers a new chance to shed light on unique historical collections that were previously inaccessible due to the limitations of analog formats. In fact, digitization has expanded the range of primary sources and presents students and scholars with a new body of historical evidence and even a critical mass of materials for analysis or comparison (Matusiak and Johnston, 2014).

The second main reason that libraries, archives, and museums undertake digital conversion is to facilitate the preservation of valuable analog materials. It is important to make the distinction between:

- Digitization as a means of preventive or "rescue" preservation
- Digitization as a reformatting preservation strategy

Preventive digitization is focused on creating digital copies for access and thus reducing physical use of rare or fragile originals, while digitization as a reformatting strategy has an additional goal of creating high-quality preservation copies of deteriorating analog materials. The benefits of digitization in protecting unique and valuable special collection and archival materials are widely acclaimed. Digitization can assist preservation efforts by limiting handling of original items and providing surrogate copies for immediate use (Gertz, 2007; Lee, 2001; Smith, 2001). Digital versions can also serve as backup copies if original materials are lost or damaged (Rieger, 2008).

The use of digitization as a form of preservation, however, has been more controversial. The concerns focus on the integrity and authenticity of digital data as well as on the stability of digital formats and storage mediums (Gertz, 2007; Smith, 2001). Gertz (2007) acknowledges that a digital copy can serve as a record, if an original object deteriorates or is destroyed, but maintains that digitization is a form of copying, not preservation. Digital technology, though it opens new doors for access and reformatting, has also created a set of new challenges with regard to the preservation of digitized objects. In contrast to established preservation methods such as microfilming, creating paper facsimiles, or photo duplication, digital technology is relatively new and raises questions about the access and retrieval of digitized copies due to the possible obsolescence of hardware and software. Challenges associated with the preservation of digitized objects are at the heart of the debate about using digitization as a reformatting strategy, but they are also part of a broader discussion about digital preservation that encompasses digitized as well as "born digital" materials.

The gradual acceptance of digitization as one option of many reformatting techniques reflects the progress in digital preservation and broader thinking about curating special collection and archival materials in the digital environment. The endorsement of digitization as a preservation reformatting method by the Association of Research Libraries (ARL) in 2004 represents a turning point in the debate, although its focus is primarily on print-based materials (Arthur et al., 2004). The ARL's proposal recognizes digital conversion as a viable option and points out that each preservation reformatting technique has its strengths and weaknesses. The Endangered Archive Programme (EAP) at the British

Library supports digitization as the preferred means of copying archival materials that are in danger of destruction or physical deterioration. This recommendation is particularly relevant in developing countries where other preservation methods such as microfilming may not be available. The Preservation Reformatting Division of the Library of Congress considers digital reformatting to be a preservation method for at-risk archival materials among other options, such as microfilm and paper facsimile copies. In fact, the Library of Congress uses digitization as a preservation method for the reformatting of film, sound, and video recordings (Marcum, 2007).

Digitization with new technical capabilities for capturing the content of analog materials brings renewed attention to the preservation of deteriorating historic photographs and audiovisual media. Conway (2010) emphasizes that the preservation of audiovisual collections remains a major challenge of the 21st century and points out that the efforts of the preservation community in preserving paper-based materials have not been extended to audiovisual resources. Archival photographic, audio, and video collections provide a rich and often untapped source of historical evidence, but their preservation is problematic due to complex and deteriorating formats. Ester (1996) notes that "photographic collections are deteriorating, and in many cases, much faster than monographs and periodicals" (p. 2). Still photography and time-based media with motion-picture film, audio, and video have been historically recorded on fragile and unstable carriers, including glass plates, cellulose nitrate- and acetate-based film, and magnetic audio and videotapes. The degrading analog formats lead to unrecoverable information loss. As Koelling (2004) succinctly states, "the point of digital preservation projects is to capture the information held by the original before time turns that information to dust" (p. 12). Fig. 3.1 demonstrates an example of an image scanned from a deteriorated glass plate negative from the Roman B.J. Kwasniewski Photographs at the University of Wisconsin-Milwaukee Libraries.

Digital conversion offers an opportunity to capture the visual and/or audio content of unstable media before they deteriorate even further. In addition, digitization projects restore the usefulness of visual materials as information resources by providing item-level indexing and placing them in the context

FIGURE 3.1 Image Scanned from a Deteriorated Glass Plate Negative

From the Archives Department, University of Wisconsin-Milwaukee Libraries. The image is available at: http://collections.lib.uwm.edu/cdm/ref/collection/mke-polonia/id/31790

of other digital collections (Capell, 2010; Matusiak and Johnston, 2014). Moreover, digitization frees the recording of knowledge from physical carriers that are prone to deterioration and enables further copying without information loss.

Digitization has introduced new dimensions to the dynamic between access and preservation. While most digitization projects are undertaken to extend the reach of cultural heritage institutions and to provide online access to their collections, other initiatives connect access and preservation goals. The two complementary goals—access and preservation—can often be realized through the same digitization project. Digitization assists preservation activities by providing surrogate copies of rare and fragile materials and by offering a reformatting option for deteriorating resources. The use of digitization as a strategy for long-term preservation of analog materials is still debatable, but it is gaining recognition as a selective approach to preserving the content of deteriorating photographic materials and archival collections of time-based media. The debate surrounding digitization for preservation has recently shifted its emphasis from reformatting to the usefulness and quality of preserved items. The goal of digitization for preservation is to capture the content of deteriorating resources and to create high-quality digital assets "worthy of long-term preservation" (Conway, 2010). Detailed discussion of digital preservation can be found in Chapter 9.

DIGITIZATION STRATEGIES AND SUSTAINABILITY

The strategies for undertaking digitization tend to be closely related to the goals and mission of the parent institution. The objectives of access and preservation are commonly shared in the cultural heritage community, but individual institutions may have additional goals, such as supporting curriculum programs, engaging the local community, meeting user requests, etc. No digitization activity is the same because of the unique characteristics of the original materials and differing institutional settings. The goals of the project determine different approaches to selection, technical standards, and the level and quality of digital capture. Overall, the selection of a digitization strategy should be informed by:

- The format and characteristics of the original materials
- The goals of digitization projects/programs
- The current and potential use
- The intended audience

The projects that focus on preservation of deteriorating analog items or those that combine access and preservation goals take a more systematic and resource-intensive approach. In these cases, all items in the collection are digitized at the highest quality affordable, and preservation-quality master copies are created for long-term archiving.

Access-oriented projects may adopt lower conversion standards to increase the amount of digitized materials. This approach can be undertaken when original items are in a stable condition and the holding institutions plan to preserve analog collections. Digitization initiatives then can forgo preservation-quality conversion, opt for minimum standards, and devote the resources to creating access copies for online delivery and meeting users' requests. This strategy allows for faster conversion and a decrease in costs. As mentioned in the introductory section, the Google Book Project has had a significant impact on the digitization of archival and special collections materials, resulting in several calls for a mass approach and a number of pilot projects (Erway, 2011; Miller, 2013; Moore, 2014; Patzke and Thiel, 2009; Sutton, 2012). Rapid digitization technologies and strategies are further discussed at the end of this chapter.

Most digitization initiatives are strategic and proactive with a goal of building digital collections for online access and/or assisting preservation. In practice, some digitization activities are also initiated at user requests. Schaffner et al. (2011) focus on user-driven digitization and provide a flexible, tiered approach to digital reproduction policies and procedures to manage on-demand workflows. The proposed strategies aim at streamlining the process of conversion and adopting minimal standards in order to deliver digital images to users in an efficient and economical way. Daigle (2012) reports on a more systematic approach to fulfilling immediate user requests by scanning larger (than requested) portions of collections to maximize conversion efforts.

Digitization strategies have also evolved over time from being project-oriented to having a more systematic program approach. The early digitization initiatives were experimental in nature, often supported with external funding, and focused on selective collections or discrete uses. The project-based approach, however, poses risks to sustainability because it is limited in duration and scope, and often lacks lasting institutional support. Kenney (2000) comments that "to succeed, digital imaging programs must permeate institutional culture and daily functions" (p. 153). The shift from a project-based approach to programs is based on the premise that digitized collections are recognized as institutional assets. Smith (2001) outlines the key points of a sustainable strategy:

- Integrates digitization into the fabric of library services
- Focuses on achieving mission-related objectives
- Relies on funding from predictable streams of allocation
- Includes a plan for the long-term maintenance of its assets

Sustainable digitization programs need not only an organizational affiliation to support the conversion activities but also an institutional commitment to long-term maintenance and preservation of digital assets. Bradley (2007) emphasizes that digital sustainability is not purely a technical issue. Sustaining digital information requires organizational, socio-technical, and economic infrastructure. Digital sustainability is closely connected to digital preservation as it encompasses "the wide range of issues and concerns that contribute to the longevity of digital information" (Bradley, 2007, p. 151). The concepts of digital preservation are discussed in more detail in Chapter 9.

In the context of digitization, sustainability requires consideration of the entire digital conversion cycle to ensure the creation and management of high-quality, sustainable digital objects. The next section provides an overview of the steps in the digitization process and summarizes the general guidelines for digital conversion.

DIGITIZATION PROCESS

Digitization of static media is the process of converting analog information to a digital format through scanning or digital photography. Time-based analog media, including film, audio, and video recordings are transformed into the digital format with the use of playback devices and analog-to-digital converters. Static materials are represented in digital format by still images, while dynamic media are represented by a time-based sequence of digital audio signals or, in the case of video and moving images, digital sound synchronized with a sequence of images. Regardless of the type of analog material or equipment being used, digitization is a process that involves multiple phases. The basic digitization cycle is similar for all materials, although the complexity increases for time-based media. Digitization

is more than simply scanning or converting audio or video signals. It requires processing and describing digital files so they can be presented and preserved in a meaningful way. The purpose of digitization is not only to convert analog information into a digital signal but also to make it into a functional and accessible digitized object.

DIGITIZATION STEPS

Digitization is a complex process that consists of multiple phases and a number of tasks associated with each phase. Basic digitization steps include:

- Project planning, selection, and preparation of materials for conversion
- Image capture (or conversion of audio and video signals) and creation of master files
- Digital processing of captured data and production of derivative files
- Recording of metadata
- Ingesting digitized objects and their associated metadata into digital library management systems
- Digital preservation of the objects created as a result of the conversion process

Fig. 3.2 demonstrates the multistep process of building digitized collections. Other models of the digitization cycle present similar steps but differ slightly in terminology (Chowdhury and Chowdhury, 2003; Lee, 2001; Zhang and Gourley, 2009). As discussed in the previous section, most digitization projects are undertaken in order to present digitized objects through online collections. Ad hoc digitization activities that fulfill user requests generally do not have the same level of complexity, but they also consist of several steps necessary to process files and store them properly. If items are digitized according to the established guidelines and best practices, they can be reused to meet other requests and be added to digital collections in the future.

The steps in Fig. 3.2 are presented in sequence since this step-by-step approach is often necessary to maintain proper workflows. However, the process is also dynamic, and depending on the nature of a project, some steps may overlap or occur in a different order. For example, in the digitization of published textual records, scanning and creating metadata may take place simultaneously, or metadata may be prepared ahead of digital conversion because descriptive information is readily available. On

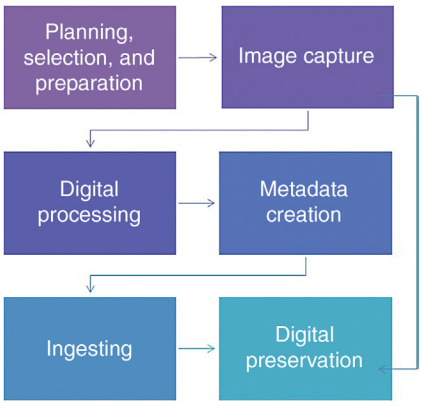

FIGURE 3.2 Digitization Steps

the other hand, in projects converting film negatives or slides, metadata creation usually takes place after scanning because high-resolution digital images provide better access to the visual content. The following section provides a brief summary of each step and associated tasks and activities.

Planning, selection, and preparation of materials is the first and the most critical step that involves defining a project's goals, scope, budget, timelines, and staff roles (Chapman, 2000). As Tanner (2001) notes, "project planning is essential to the successful implementation of any technology based project and particularly one involving digitization" (p. 329). The first phase of the project also includes selecting and assessing analog materials and preparing them for conversion, establishing technical specifications, and selecting equipment. The selection of materials for digitization requires considering multiple factors including copyright status, the format and size of originals, condition and unique characteristics, and requirements for handling fragile items. Selection criteria are discussed in more detail in Chapter 2. Damaged or deteriorating items may require some conservation efforts prior to digitization. The overall assessment of the original source collections influences the selection of digital capture equipment. Many activities in this phase are parallel to collection development, especially in rapid capture projects. Items or entire collections can be selected, processed, digitized, and rehoused as part of one project to improve efficiency of workflows and reduce material handling.

Digital capture represents the heart of the digitization process, in which analog information is sampled by capture devices (scanners, digital cameras, or analog-to-digital converters) and converted into digital signals. Depending on the availability of equipment and in-house expertise, this phase of the project can be outsourced to a digitization vendor. Dale (2007) provides a comprehensive overview of the pros and cons of outsourcing. The files produced as a result of the capture process represent information-rich digital masters, also referred to as archival master files. Digital masters represent "the best copy produced by a digitizing organization" (FADGI, n.d.). The FADGI Guidelines consider a file to be a digital master only if it meets the established technical requirements and has been quality controlled (FADGI, 2010). Master files need to be checked against quality benchmarks to ensure that they accurately represent the content of analog source materials and adhere to the established digitization guidelines. Master files serve as the source of all subsequent files to be derived. In practice, two copies of master files are usually created: one to be saved for long-term preservation, and a second one, often referred to as a service master, to serve as a source for smaller-size derivative files. The high-quality master files are important assets, so they ought to be transferred after capture to a digital repository for long-term preservation. The task of transferring digital masters for preservation is represented in the diagram by an additional arrow on the right.

Digital processing is a phase in which files created as a result of digital capture are edited and transformed to improve their quality and/or to enhance functionality. For example, copies of faded archival documents or maps can be edited to improve their contrast and legibility. In addition, artifacts introduced by the scanning process as well as scratches or dust marks can be corrected to improve the quality of the image. Noise may need to be removed from audio files, and the synchronization of audio and image might need to be improved for video files. It is recommended, however, that digital masters be saved intact, without any changes. The enhanced image should be saved as a second copy of the service master, also referred as to the production master file (FADGI, n.d.). Service master files reflect the changes introduced through digital processing and serve as a source for derivative files to be ingested into digital collections. Digital images of printed text pages can also be processed using Optical Character Recognition (OCR) software to create searchable digital documents. Textual documents

can also be rekeyed to allow for encoding and/or to facilitate full-text searching. Digital processing also includes the task of quality review of service masters and derivatives.

Metadata creation involves selection of metadata schemas, customization of metadata templates, selection of controlled vocabulary tools and input guidelines, and building item-level records. Depending on the amount of information available in original collections, this process may require additional research to provide access points and contextual information. This step includes the recording of descriptive information as well as other metadata types, including administrative, structural, and technical. Quality review of metadata records is associated with this step. Chapter 5 provides more information on schemas and tools used in metadata creation.

Ingesting is a step that brings the transformation of converted items into completion, turning them into usable digitized objects. During this phase, digital content files (images, text, audio, or video) are associated with corresponding metadata records and uploaded into digital library management systems for access. After the collection building process is finalized, digitized objects are available for use. The ingesting tasks can vary depending on the content management system being used for collection building. For example, the CONTENTdm system generates derivative files for access automatically during the ingest process, while Omeka requires the preparation of derivatives in advance. The systems also vary in the level of support for metadata customization and functionality of digitized objects. Chapter 6 provides a comparative overview of six systems currently used in practice.

Digital preservation in the context of digitization refers to the maintenance of digital collections and the long-term preservation of digital master files created as a result of the digitization process. Digital master files represent valuable assets that need to be preserved over time. They serve as a source for future derivatives and as preservation copies for deteriorating analog materials. The activities associated with preserving digital masters include developing a long-term preservation policy and establishing an infrastructure and a strategy for identification, archival storage, integrity and authentication checking, regular backups, refreshing, and migration. A sustainable digitization program requires the development of an institutional approach to digital preservation and involves building a local digital repository or participating in a shared program. Digital preservation is discussed in more detail in Chapter 9.

This brief overview offers some insight into the complexity of the conversion process and necessary tasks to create high-quality usable digital objects. Quality control is an integral part of each step and encompasses procedures and techniques to verify accuracy and consistency of digitized objects (Rieger, 2000). Digitization, like many other technology-intensive projects, requires careful planning, managing multiple workflows, and proper documentation. Good project management is the key to successful digitization initiatives (Chapman, 2000; JISC Digital Media, 2014; Tanner, 2001). Over the years, members of the cultural heritage community have shared their experience and expertise in digitization through openly available tutorials and guides to best practices. The following section provides a summary of general guidelines.

GENERAL DIGITIZATION GUIDELINES

The purpose of guidelines is to ensure the creation of high-quality, sustainable digital objects that support current and intended use and are interoperable and consistent across collections and institutions. The guidelines that have emerged in the cultural heritage community, especially in the United States, are advisory rather than prescriptive in nature. They offer a range of general and technical recommendations but do not constitute a set of formal standards. The *Framework of Guidance for*

Building Good Digital Collections is described as a recommended "best practice" (NISO Framework Working Group, 2007). The most recent guidelines issued by the division of the American Library Association stress that "at this point there is no official standard for digitization, but institutions are discussing how they can collaborate and share digitized content" (ALCTS, 2013, p. 2). This approach offers individual institutions some flexibility but has also resulted in a plethora of published guides and tutorials. Conway (2008) examines 17 guides to best practices in digitizing visual resources and concludes that the lack of standardization has implications for the quality and integrity of digitized objects and may be a hindrance to wider adoption of the guidelines by small and midsize cultural institutions.

The development of best practice guides was spurred by the early adopters of digital technology, such as Cornell University Libraries, the Library of Congress, US National Archives and Records Administration (NARA), and organizations such as Digital Library Federation (DFL), International Library Federation Association (IFLA), and Research Library Group (RLG). Conway (2008) also recognizes the seminal work of imaging specialists and pioneers of digitization, including Michael Ester (1996), Anne Kenney and Steve Chapman (1996), Franziska Frey and James Reilly (1999), Steve Puglia (2000), and Steve Puglia et al. (2004). Their work on imaging concepts and specifications provided the necessary theoretical and technical foundations for developing guides to best practices. The tutorial *Moving Theory into Practice* developed at the Cornell University Libraries has contributed significantly to the training of librarians and archivists in the concepts and procedures of digitization (Kenney et al., 2000). In addition to the guidelines developed by the Library of Congress and NARA, major collaborative digitization initiatives, such as the California Digital Library (2011) and Colorado Digitization Program (BCR, 2008) issued their own sets of recommendations. Those guides to imaging best practices have in turn influenced the development of guidelines at the state and institutional levels (see Appendix A for an annotated bibliography of selected guides).

The majority of published tutorials and guides to best practice focus on static textual and visual resources, but the underlying principles can also be applied to time-based media. The guidelines emphasize digitization at the highest quality to capture informational content and attributes of analog source materials in order to create accurate and authentic digital representations. Recently released guidelines build upon foundational concepts but offer higher technical specifications that reflect the current digital environment. The approach that has emerged is to offer minimum capture recommendations for a variety of static and time-based media with an understanding that unique characteristics of source materials may require variations in the specifications. A set of accepted minimums is, however, recommended to create sustainable digitized content (ALCTS, 2013).

The following list provides a summary of the general digitization principles presented in a number of currently available guides (ALCTS, 2013; BCR, 2008; FADGI, 2010; Yale University, 2010):

- Digitize at the highest resolution appropriate to the nature of the source material
- Use standard targets for measuring and adjusting the capture metric of a scanner or digital camera. Grayscale or color targets provide an internal reference within the image for linear scale and color information.
- Create and preserve master files that can be used to produce derivative files and serve a variety of current and future use needs
- Create digital objects that are accessible and interoperable across collections and institutions
- Ensure a consistent and high-level quality of digitized objects
- Digitize at an appropriate level of quality to avoid recapture and rehandling of the source materials

- Digitize an original or first generation of the source material
- Create meaningful metadata for digitized objects
- Provide archival storage and address digital preservation of digitized objects

The general guidelines assume a use-neutral approach that has been strongly recommended since the early days of digitization projects (Besser, 2003; Ester, 1996; Kenney, 2000). It implies that a source item is digitized once and at the highest level of quality affordable to meet the needs not only of an immediate project but also of a variety of future uses. The goal of this approach is to create high-quality digital representations and to avoid redigitizing in the future. The use-neutral approach is an important component of digitization best practices, as it addresses not only the current needs but also, as Besser (2003) emphasizes, "all potential future purposes" (p. 43). It includes the notion of digital master files (sometimes referred to as archival masters) and derivatives. Ester (1996), who introduced the concepts of digital archival and derivative images, notes "an archival image has a very straightforward purpose: safeguarding the long-term value of images and the investment in acquiring them" (p. 11). In addition to the difference in purpose and use, digital masters and derivatives also differ in regard to file attributes such as size, compression, dimensions, and format.

Digital masters are created as a direct result of the digital capture process and should represent the essential attributes and information of the original material. Digital masters are supposed to be "the highest quality files available" (Besser, 2003, p. 3). They should not be edited or processed for any specific output. Because the process of creating digital masters usually results in large file sizes, digital masters are not used for online display. In fact, many archival formats such as TIFF are not supported by major web browsers. Their primary function is to serve as a long-term archival file and as a source for derivative files. Digital masters are stored in digital repositories for long-term preservation. General recommendations for digital master file creation include:

- Digitize at the highest quality affordable
- Save as an uncompressed file
- Use standard, nonproprietary file formats, such as TIFF for static media (text or still images) or WAV for audio
- Do not save any enhancements in an archival copy
- Use an established file-naming convention

Derivatives are created from digital master files for specific uses including presentation in digital collections, print reproductions, and multimedia presentations. General recommendations for derivative files include:

- Reduce the file size so it can load quickly and be transferred over networks
- Use standard formats with lossy compression such as JPEG
- Use standard formats supported by major web browsers

Table 3.1 provides a summary of formats recommended for digital masters and derivatives based on analog source type. File format is an essential component, as it provides an internal structure and a "container" for digitized content. Unlike physical objects, digital files do not exist in an independent material form. Digital data is stored in file formats and requires hardware and software to be rendered. The Sustainability of Digital Formats site at the Library of Congress provides a working definition of formats as "packages of information that can be stored as data files or sent via network as data streams (also known as bitstreams, byte streams)" (Library of Congress, 2013).

Table 3.1 Recommended File Formats for Digital Masters and Derivative Files

Analog Material	Digital Masters	Derivatives
Text	TIFF	JPEG, PDF
Photographic images (prints, negatives, slides)	TIFF	JPEG, JPEG 2000
Audio recordings	WAV/BWF	MP3
Moving image (video, film)	JPEG 2000/MXF	MPEG-4 (MP4)

File formats vary in their functionality and attributes. The master file format needs to be platform independent and have a number of attributes, such as openness, robustness, and extendibility, to support the rich data captured during the conversion and to ensure its persistence over time as technology changes (Frey, 2000a). The selection of an appropriate format has implications for access across platforms and transfer over networks as well as storage and long-term preservation. The *Framework of Guidance for Building Good Digital Collections* states as one of its principles: "a good object exists in a format that supports its intended current and future use" (NISO Framework Working Group, 2007, p. 26). The section of this chapter on technical factors provides an overview of the recommended formats for static media, including TIFF, JPEG, JPEG 2000, PDF, and PNG. Audio and moving image formats are discussed in more detail in Chapter 4.

General guidelines also include recommendations for establishing a file-naming convention. File names for digital masters and derivatives need to be determined before the digital capture process begins and preferably follow a convention adopted by the parent institution or department. Digital files should be well organized and named consistently to ensure easy identification and access. Systematic file naming helps not only to manage the project but also ensures system compatibility and interoperability. File names can be either nondescriptive or meaningful. Both approaches are valid, but each has its pros and cons (Frey, 2000a; Zhang and Gourley, 2009). Selecting a file-naming convention for digitization requires long-term thinking and a good understanding of the scope of the project and/or the size of the original collection. File-naming recommendations include:

- Assign unique and consistent names
- Use alphanumeric characters—lowercase letters and numbers 0 through 9
- Avoid special characters, spaces, and tabs
- Include institutional IDs (if available)
- Number files sequentially using leading zeros
- Use a valid file extension, such as .tif, .jpg, or .pdf
- Limit file names to 31 characters, including the three-character extension; or if possible, use 8.3 convention (8 characters plus three-character extension)—for example, aa000001.tif

DIGITIZATION OF TEXTUAL AND STATIC VISUAL RESOURCES

Static media encompasses a wide range of textual documents, from handwritten letters to printed books, and an even more complex array of photographic resources and other types of two-dimensional visuals. Photographs, archival records, postcards, rare books, manuscripts, and newspapers represent

FIGURE 3.3 Pixel Matrix of an Image Scanned from a Color Slide

the majority of digitized items in the static media category. Photographs are the most frequently digitized objects. The 2012 survey of digitization activities in Europe reports that 66% of surveyed cultural heritage institutions have digitized photographs (Stroeker and Voegels, 2012).

Textual materials and still images are converted to the digital form through scanning or digital photography. A variety of digitization equipment, from scanners to digital cameras, can be used in the conversion process. During the scanning process, an item is sampled and mapped as a grid of picture elements. A pixel is a single picture element. The captured data are represented by a series of pixels called "raster images" (also referred to as "bitmap images"). Puglia (2000) describes the structure of digitized images: "digitization converts an image into a series of picture elements of pixels, little squares that are either black or white (binary), a specific shade of gray (grayscale), or color. Each pixel is represented by a single or series of binary digits, either 1s or 0s. The pixels are arranged in a two-dimensional matrix called a bitmap" (p. 83). Fig. 3.3 demonstrates an example of a pixel matrix in an image scanned from a 35 mm color slide. The image of an Iranian woman featured in Fig. 3.3 is part of the American Geographical Society Library Digital Archive and is available at http://collections.lib.uwm.edu/cdm/ref/collection/agsphoto/id/102.

Bitmap images are characterized by a number of measures, such as resolution, pixel bit depth, and color mode, and their size and quality are affected by other technical factors including compression. The following section provides a brief overview of basic image measures, digitization equipment, and selected technical recommendations for conversion of textual and photographic resources.

TECHNICAL FACTORS

A range of technical factors play a role in the digitization process and relate to the quality and size of captured images. Paying close attention to image measures, such as resolution, bit depth, and color mode is critical during the conversion process, as these directly impact the quality of digital master files. Other factors, such as compression, need to be determined during the processing stage in the production of derivative files. Technical specifications, including resolution, bit depth, and mode of capture, have to be considered in the selection of the scanners and digital cameras.

Resolution is one of the most important factors, as it refers to the number of times an image is sampled and consequently relates to the amount of detail captured during the scanning process. Resolution specifically refers to the number of dots, or pixels (picture elements), used to represent an image. It is expressed in a number of ways, DPI (dots per inch) or PPI (pixels per inch). PPI refers to the number of

pixels captured in a given inch and is used when discussing scanning resolution and on-screen display. DPI comes from the print environment in reference to the optical resolutions for images and hardware. DPI more accurately refers to output devices, or how many dots of ink per inch a printer puts on the paper or onscreen monitor display. However, the two terms are often used interchangeably. Digitization guidelines recommend scanning at the highest resolution affordable to accurately and fully capture the visual content of the original materials. Scanning resolution depends on the technical specifications of capture devices, so it is important to select a scanner or camera with sufficient optical resolution. Depending on the camera's lenses and support, the achieved resolution can sometimes be different from the optical resolution. FADGI (2010) provides helpful guidelines on sampling frequency. Items scanned at high resolution will result in large digital master files. However, there is no "one size fits all" ideal or standard resolution. The resolution should be adjusted according to the type of source item, its physical dimensions, and the amount of detail that needs to be captured. Digitization guidelines provide a range of recommended resolution measures relative to the types of sources materials and dimensions. For example, a minimum resolution for textual materials without images is 300 ppi, while a photographic 8 × 10 in. print will benefit from scanning at 600 ppi (ALCTS, 2013). Higher resolution provides more pixels and will generally render more detail, but there is also a point when increasing resolution does not yield any additional information.

Pixel bit depth influences the representation of images, rendered in a grayscale tone or a range of colors. It is a measure that "defines the number of shades that can actually be represented by the amount of information saved for each pixel" (Puglia, 2000, p. 85). Depending on the number of bits per pixel, images are represented as black or white, grayscale, or true color. One-bit images are bitonal—either black or white. Eight-bit images are necessary to represent 256 shades of gray tones in photographic images. Most color images require 24 bits per pixel to provide true representation of color. The greater the bit depth, the more information about the source is captured by the scanning device, resulting in a more accurate digital representation of the original. A bit depth of 8 can capture enough information to represent 256 colors or shades of gray. A bit depth of 24 captures over 16 million colors or shades of gray. It is worth remembering that there is a relationship between bit depth and file size. Scanning at a higher bit depth increases the overall file size. The usage of the term has evolved as institutions have moved from legacy scanning to modern raw capture. Currently, an 8-bit file means a 3-channel file with 8 bits per channel, which used to be referred to as a 24-bit image.

Color mode refers to the representation of color in images. Color images consist of three or more channels that represent color information. Several different systems are used to represent color images, with RGB being one of the most common. RGB stands for red, green, and blue, the three channels used to represent digital color images. Computer software combines the three channels for each pixel to determine the final color. An RGB color digital image file consists of three channels, each with 8 bits of data (3 channels × 8 bits = 24 bits). Many cultural heritage institutions process to 16 bits per channel to achieve subtle gradations of color.

Modes of capture refer to the way digitization equipment captures images in relation to the two measures: bit depth and color mode.

- *Bitonal* mode is appropriate for printed text materials without illustrations. Text can be scanned in bitonal mode where one bit per pixel will represent black or white values. Bitonal scanning was used in early digitization projects but now is used infrequently.
- *Grayscale* mode requires multiple bits per pixel (8 bits minimum) to represent shades of gray and is appropriate for scanning photographic black-and-white film negatives, black-and-white

photographic prints, or books and newspapers with grayscale images. Increasingly, color RGB mode is recommended for black-and-white photographic materials since it captures more information from an analog source. Generally, black-and-white prints and negatives will benefit considerably from scanning in RGB. If storage of these files is an issue, they may be converted to grayscale after scanning.

- *RGB color mode* is recommended for items with continuous tone color information. RGB mode is used for all textual and visual resources where color is present in the source item. In addition, archival textual materials or rare books are scanned in color when it is important to capture the aging nature of paper or other artifacts (handwritten notes, stamps, etc.).

Compression is the process of reducing the file size by discarding a certain amount of information. The process is, in most cases, irreversible. Compression is closely related to the quality of images and their size.

- *Lossless compression* discards redundant information and does not impact the quality of images. It allows for storing data in a more compact form. Lossless compression is supported by TIFF and JPEG 2000 formats and can be used for service masters—images created as a result of processing and image enhancement techniques. However, it should not be applied to digital master files. Digital archival masters should be saved as *uncompressed* files. *No compression* is different from lossless compression. Digital master files should retain all information captured during the conversion process.
- *Lossy compression* creates file sizes that are smaller, but it also contributes to the loss of image data and decreases quality (the amount of discarded information depends on the level of compression). It is important to remember that when a compressed image is decompressed, it is no longer identical to the original image. JPEG format applies lossy compression. JPEG file sizes can be reduced by applying compression, which makes them suitable for online access and distribution.

Formats provide a standardized method of encoding and organizing data into files. The digital conversion of textual and photographic materials results in still raster (bitmap) images, a two-dimensional grid of pixels. A variety of formats can be used for storing raster images. The distinction between digital master files and derivatives in the digitization of cultural heritage materials provides a foundation for the selection of formats. TIFF has been recommended as a master format for still images. TIFF has been widely adopted, and, as a recent study of file formats for raster still images indicates, it "has been the format of choice for the cultural heritage community" (FADGI, 2014, p. 3). JPEG and JPEG 2000 have been recommended as derivative formats for photographic images, newspapers, manuscripts, and maps. JPEG 2000 has also been considered as an archival format for master files (Buckley, 2008; Buonora and Liberati, 2008; Van der Knijff, 2011). PDF is recognized as a suitable derivative format for textual documents. PDF/A is a format recommended for archiving digital documents.

- *TIFF* (Tagged-Image File Format) is a stable and widely adopted file format for master files of raster still images. Used since the early days of digitization, TIFF has become the de facto standard for digital masters of digitized static cultural heritage materials. Fleischhauer (2014a) notes, "its endurance in time can be seen as a strength, especially considering the wide array of applications that can read it" (pp. 2–3). Highly flexible and platform-independent, it can be used for storing bitonal, grayscale, and color still images. TIFF combines raster image data with a

flexible tagged field structure for metadata. TIFF supports lossy and lossless compression. It is recommended that digital masters be saved as uncompressed TIFF files, but lossless compression, such as LZW compression, can be used for service masters. Uncompressed TIFF files require a considerable amount of storage space. TIFF is an open and well-documented standard, with the specifications of TIFF Revision 6.0 maintained by the Adobe Systems. The TIFF filenames use .tif or .tiff extensions.

- *JPEG* (Joint Photographic Experts Group) is designed for compressing and thus reducing the size of grayscale and color raster still images. The JPEG standard was published in 1992 and is commonly used on the web and in digital cameras. In digitization, JPEG is used primarily for derivative images to be displayed in digital collections. JPEG applies a lossy compression method, which reduces the file size. The amount of compression can be adjusted. The typical ratio of 10:1 results in very little perceptible loss in image quality. JPEG works particularly well with photographic images of continuous tone, while images with lettering or line drawings may suffer some degradation in quality. The effective compression makes JPEG a particularly suitable format for online display and transfer over the Internet. However, because of the loss of data associated with compression, this file format should not be used for master files. The JPEG format is supported by all browsers. The JPEG file extensions are .jpg or .jpeg.
- *JPEG 2000* is an international standard for the compression of digital still images. It was proposed by the Joint Photographic Experts Group in the year 2000 as an open file format and a compression method with the goal of improving or superseding the original JPEG format. JPEG 2000 provides a new compression algorithm with progressive display, multiresolution imaging, scalable image quality, and the ability to handle large and high-dynamic range images (Buckley, 2008). The JPEG 2000 file format also offers significant improvements over earlier formats by supporting both lossless and lossy image compression. Because of its superior ability to handle large content files and dynamic display with support for zooming and panning, JPEG 2000 has been used as a derivative file format for maps, newspaper pages, and other large images (Fleischhauer, 2014b). At this point, the format cannot be viewed natively in most web browsers and requires a dedicated JPEG 2000 viewer. The potential of JPEG 2000 for storing large master files and as an alternative to uncompressed TIFFs files has also been explored due to its excellent compression performance (Buonora and Liberati, 2008; Van der Knijff, 2011). The acceptance of JPEG 2000 as a preservation format, however, has been slow and a subject of debate in the cultural heritage community (Adams, 2013; Fleischhauer, 2014b). The study conducted by Van der Knijff (2011) also identifies some preservation risks, related to the current format specification in color space and in the handling of grid resolution, which may lead to the loss of some information in future migrations. JPEG 2000 uses .jp2 and .jpx extensions.
- *PDF* (Portable Document Format) is an access format developed by Adobe Systems in 1993 to share and view digital documents. It remained a proprietary format until 2008 when it was released as an open international standard. PDF is used to represent 2D documents in a fixed-layout format. PDF documents maintain the original structure and appearance of source items and can be exchanged across many platforms. PDF is a universal format used to represent both born digital and digitized documents. A popular format in the publishing industry, PDF became a de facto standard for scholarly publications, administrative documents, and many textual documents shared over the web. In digitization, PDF is used as a derivative format to represent multipage objects, such as manuscripts, books, journals, and archival documents. Full-text searching

of digitized documents can be incorporated into PDF derivatives, but it requires additional processing of source images. At a minimum level, digitized historical documents are presented in the PDF format as images—digital facsimiles. Full-text searchability is available for digitized print documents processed with Optical Character Recognition (OCR) software (Turró, 2008; Yongli, 2010). A free and widely available PDF reader can be used as a standalone program or a browser plug-in. A PDF filename has a .pdf extension.

- *PDF/A* builds upon the specifications of PDF but was developed specifically as a standard format to ensure long-term accessibility and the preservation of electronic documents. PDF/A addresses the concerns of the archival community and is recognized as a format for the digital archiving of documents (Dryden, 2008). PDF/A-1 was released in 2005, and the latest version, PDF/A-3, was made available in 2012. It provides "a mechanism for representing electronic documents in a manner that preserves their static visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files" (Lazorchak, 2014). The difference between PDF and PDF/A is in the preservation function, which in PDF/A is achieved by embedding all fonts and metadata within the file so that it can be consistently rendered regardless of the hardware and software used to create or view it.
- *PNG* (Portable Network Graphics) was designed to replace the older GIF format. PNG supports raster grayscale and color image files and offers lossless compression. PNG is supported by all major web browsers and is a popular choice for transferring images over the web. The use of the PNG format in digitization projects is limited thus far. In a recent Library of Congress blog, Fleischhauer (2014b) highlights PNG support for color management and lossless compression and wonders about the potential use of PNG for master files. PNG uses a .png file extension.

The file formats used in the digitization of static media demonstrate a high degree of stability, especially in comparison to the still-evolving formats for video recordings. A comparative study of TIFF, JPEG, JPEG 2000, PNG, and PDF, conducted by the Federal Agencies Digitization Guidelines Initiative, indicates that all formats have viable sustainability, although they vary in attributes, capabilities, and cost of implementation (FADGI, 2014).

DIGITIZATION EQUIPMENT

The focus of this overview is on imaging equipment used in the digital conversion of textual and 2D visual resources recorded on a variety of analog carriers, including paper, film negatives, glass plates, or slides. The type of equipment used in the conversion process depends on the condition and format of the analog source, its physical dimensions, characteristics, rarity, and fragility. When considering the format of analog materials, it is important to make a distinction between:

- *Reflective* materials, such as paper used in creating manuscripts, books, maps, drawings or photographic prints
- *Transparent* media, such as film, slides, or glass plates

This distinction has implications for selecting an appropriate capture device and is related to the way scanners work.

A scanner is a device that analyzes the surface of an image, printed text, or transparent film and converts it into a digital image, which is a 2D pixel array. Most scanners use CCD (charge-coupled device) light-sensitive image sensors. In the case of reflective materials, such as paper-based textual

resources or photographic prints, the light is reflected off the surface of the paper and read by a set of light-sensitive diodes that then convert this reading into a digital value. In the case of transparent materials, such as film negatives or slides, the light needs to pass through the material so the sensor can read the image and convert it into a digital file. Transparent materials require dedicated film scanners or flatbed scanners with a transparency adapter.

Cultural heritage institutions have collected historical materials on a variety of analog formats, and their conversion requires versatile digitization equipment. The selection of a scanner or digital camera depends on the physical dimensions and characteristics of analog sources and will greatly impact image quality. Photographic prints can be digitized using flatbed scanners. Larger film negatives such as 4×7 in., 5×7 in., or 8×10 in. can be scanned using flatbed scanners with a transparency adapter, but small negatives of 35 mm require dedicated film scanners. Textual materials on paper (reflective), if they are single-leaf documents, can be scanned using flatbed scanners or even faster sheet-fed scanners if materials are not rare. Bound materials—books and manuscripts—require overhead scanners or digital cameras. Maps and charts that are large and exceed the size of flatbed scanners will need wide-or large-format scanners. And finally, documents on microfilm, such as newspapers, require dedicated microfilm scanners. Digital cameras are increasingly being used in digitization projects, but they require a digital imaging studio with additional pieces of equipment. The price of scanners and cameras has decreased over the years, but the cost still plays a significant role in selecting equipment, especially when it comes to high-end film scanners or large-format digital cameras.

The following section provides a brief overview of types of scanners and cameras used in practice. The examples presented in the figures are meant to illustrate the types of equipment but are not intended to be an endorsement of specific models or companies.

Flatbed scanners are suitable for single-leaf text documents and most photographic prints, provided the material does not exceed the scanner's maximum imaging area. Large-format flatbed scanners and sheet-fed scanners can capture single-leaf oversized materials. Flatbed scanners are used for digitizing reflective materials. Scanning transparent materials, such as glass plates and larger film negatives, requires a transparency adapter. The major limitation of implementing scanners in a digitization program is that they are very slow and require contact of a scanned item with the glass surface of the scanning bed. Flatbed scanners are not suitable for brittle or bound materials.

Fig. 3.4 includes an example of a flatbed scanner, Epson Expression 10000, that is commonly used by archives and libraries for digitization. It provides a scanning area of 11×17 in., high optical resolution up to 2400 ppi, and high bit depth, up to 48 bit for color and 16 bit for grayscale images. A transparency adapter is optional, but it allows for the scanning of large film negatives (4×7 in., 5×7 in., or 8×10 in). The resolution 2400 ppi, however, is too low for small size film (35 mm).

Overhead scanners are used in the image capture of bound materials—books and manuscripts—as well as for fragile reflective materials, such as newspapers, prints, drawings, and small maps. The source of light is on the side, light sensors are at the top, and thus books do not need to be placed face down.

Fig. 3.5 presents an example of an overhead scanner. This scanner has a large scanning area and an integrated glass plate that allows for the flattening of uneven materials. It also includes a motorized book cradle that makes it easier when scanning bound volumes. The resolution is 400 ppi, so this type of scanner works well with books but not with smaller items that require a higher resolution.

Fig. 3.6 presents an example of DT BC100 Book Capture System, a dedicated book scanning system that offers fast, preservation-quality image capture of bound monographs and loose materials, including works on paper, serials including newspapers, loose manuscripts, photos, and drawings. The

FIGURE 3.4 Flatbed Scanner

Image courtesy of Ling Meng, Digital Collections and Initiatives, University of Wisconsin-Milwaukee Libraries.

FIGURE 3.5 Overhead Scanner

Image courtesy of Ling Meng, Digital Collections and Initiatives, University of Wisconsin-Milwaukee Libraries.

system includes a V-cradle and glass plates for flattening materials. For more fragile or rare items, the cradle can be raised so material is about to touch the glass but does not actually make contact.

Film scanners are used in the conversion of transparent media, such as 35 mm slides and film negatives. Some flatbed scanners include transparency adapters, but their resolution and dynamic range are limited in comparison to film scanners. Dedicated film scanners offer higher resolution and are

FIGURE 3.6 DT BC100 Book Capture System

Image courtesy of Digital Transitions Division of Cultural Heritage. www.dtdch.com

FIGURE 3.7 Nikon Film and Slide Scanner

Image courtesy of Ling Meng, Digital Collections and Initiatives, University of Wisconsin-Milwaukee Libraries.

appropriate for the small size of the original transparent material. They also enable scanning without glass. Glass attracts dust which becomes visible at high resolutions.

Fig. 3.7 includes an example of Nikon CoolScan that can create scans of 35 mm film negatives and slides at 4000 ppi. Fig. 3.8 demonstrates a high-end Hasselblad film scanner, which provides high resolution of 6000 ppi and is capable of scanning other film sizes in addition to 35 mm film. It can also scan batches of negatives in a somewhat automated fashion, but of course, high quality and scanning speed come at a price. Fig. 3.9 provides another example of a film scanner, the DTFSK scanner available

FIGURE 3.8 Hasselblad Film Scanner

Image courtesy of Ling Meng, Digital Collections and Initiatives, University of Wisconsin-Milwaukee Libraries.

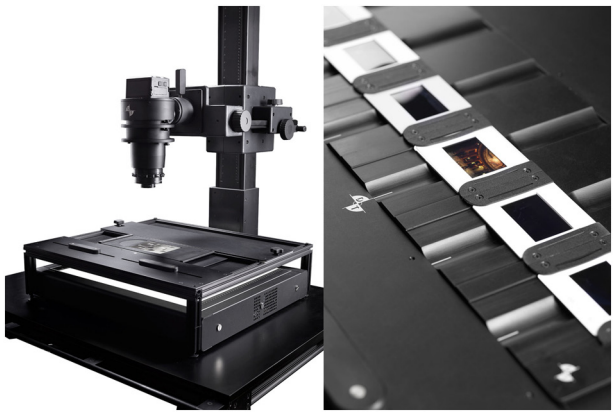

FIGURE 3.9 DTFSK Film Scanner

Image courtesy of Digital Transitions Division of Cultural Heritage. www.dtdch.com

from Digital Transitions. This new generation film scanner offers not only preservation-quality image capture but also speed, being 400 times faster than legacy scanning equipment. It is capable of scanning a wide range of formats from 35 mm up to 11 × 17 in. film.

Digital camera systems provide the most versatile image capture environment, but they require setting up a more robust imaging studio. It is important to make a distinction between digital cameras and camera-based systems used in digitization. In addition to a camera unit, digital camera systems include a number of components, such as a light source, a vacuum easel for flat materials, and a cradle for books or other bound materials. Digital camera systems are designed specifically for cultural heritage applications, of which a digital camera is a very small part. There are even currently emerging technologies that incorporate all of these into a single unit. A camera unit will include a camera body, lens, and specially formatted camera back, such as those offered by PhaseOne, Leaf, and Sinar or Hasselblad. These larger format digital camera systems not only offer high resolution—they are engineered to provide the greatest dynamic tonal and color range and clarity, which can render amazing levels of image quality. Many materials of varying formats can be captured using an overhead camera, and the resulting details in the image file often show such minute details as the support composite fibers, quality of typeset or engravings, and even single bristle lines left from an artist's brushstroke. Oversize items, such as maps, can be placed on a vacuum easel or copy stand or hung on the wall for image capture, and the image quality can even facilitate multiple captures of segments of an item to be merged into a larger image with minimal quality loss rather than a single-frame capture that may not show the greatest detail of a given item.

Fig. 3.10 is an example of a camera system in place at the Denver Public Library's Imaging Services Lab, which is a division of the Western History and Genealogy Department. They have retrofitted their previous view camera system with an RCam from the Digital Transitions Division of Cultural Heritage, which utilizes a PhaseOne back and Schneider lens. It is utilized as the main tool in their digitization work as it can capture multiple format types that may have varying characteristics. These types of camera installations, however, are expensive and require substantial expertise and in-depth training. Many institutions have determined that the initial investment of high-end equipment and well-trained personnel is rewarded by superior image quality, smoother workflows, and rapid production of digital assets.

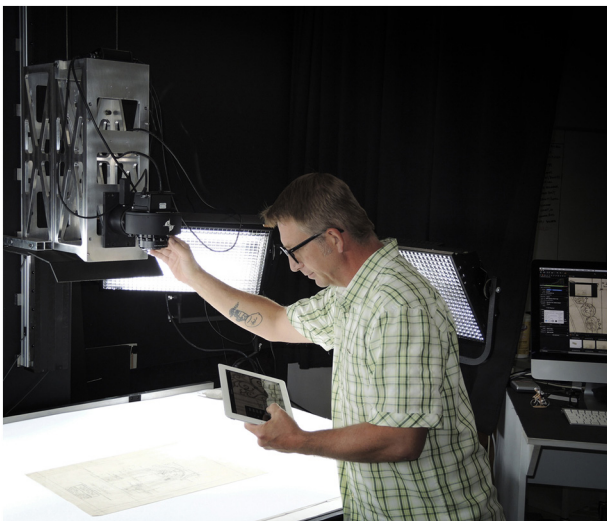

FIGURE 3.10 Large-Format Digital Camera System

Image courtesy of Benjamin Miller. Imaging Services Lab, Denver Public Library.

Less costly options include professional-grade digital SLR (single lens reflex) cameras that work well for books and other prints. Most consumer-level digital cameras are inadequate for the reproduction of special collection materials.

Many digital camera systems involve an intermediary step of working with raw files before they are processed into digital masters. The immediate result of camera-based capture is a raw proprietary file, which at the point of capture does not represent a digital master. Raw files need to be saved in the TIFF format in order to become master files. Photographic capture with its raw workflows is more complex than direct-to-TIFF scanning, but it offers more flexibility and acceleration in imaging process.

RECOMMENDATIONS FOR DIGITAL CAPTURE OF TEXT AND PHOTOGRAPHIC IMAGES

The processes for digitizing textual and photographic materials are similar to a certain extent. For both source types, the output is in the form of digital images, and masters are saved in the TIFF format. However, recommendations for resolution and bit depth vary because of the fundamentally different content, and the amount of detail that needs to be captured is greater for photographic materials. The complexity of photographic processes and formats also demands a wider range of specifications and more versatile digitization equipment. While digitized photographs remain basically as images, digitized text documents need to be further processed and transformed to make the textual content searchable.

Text digitization focuses primarily on legibility issues. Scanned pages of books, newspapers, journals, and other textual documents can be presented as images and/or as searchable text. Digitized text not only has to be legible to the human eye but also needs to be recognized and processed by software if searchable text is to be created. The size of text in original text material is an important factor when determining technical specifications. Higher resolution and bit depth are recommended for documents

Table 3.2 Minimum Digitization Recommendations for Textual Materials

Original Material	Scanning Resolution	Bit Depth	Capture Mode	Notes
Books and other text without images (nonrare)	300 ppi	8 bit	Grayscale	The resolution may be adjusted according to the detail to be represented.
Books and other text with images (nonrare)	400 ppi	8 bit or 24 bit	Grayscale or color	The resolution may be adjusted according to the detail to be represented. Capturing in color (24-bit RGB) is recommended.
Rare books	400 ppi	24 bit	Color	Increasing resolution may be necessary for less standardized fonts.
Manuscripts	400 ppi	24 bit	Color	Increasing resolution may be necessary for difficult to read handwritten documents.

with smaller typeface. Features of the original documents, such as handwritten versus print, their rare nature, or the presence of illustrations, also play a role in adjusting the specifications.

Table 3.2 is based on the most recent recommendations prepared by the Association for Library Collections and Technical Services, a division of the American Library Association (ALCTS, 2013). It is important to remember that the listed specifications represent minimum recommendations. Increasing resolution is highly recommended for rare books with ornate and irregular fonts and for manuscripts where legibility might be an issue.

Page images produced as direct output of digital conversion are not searchable. Digitized documents can be converted into fully searchable text either through manual keying or automatic processing using OCR software. A hybrid approach may combine OCR processing with a review for accuracy and manual correction. Current OCR technology can process typed or printed text with various degrees of accuracy, but is not capable of converting handwritten text. Manuscripts and other handwritten documents must be typed to become searchable. Because of a large body of digitized historical documents and manuscripts, there is a strong research interest in developing solutions for machine recognition of handwritten text (Romero et al., 2011; Sánchez et al., 2014). However, these efforts remain in an experimental stage.

In practice, most searchable text of digitized printed or typed documents is created through OCR software (Lesk, 2004; Yongli, 2010). OCR is the process that converts the text of a digitized printed page into a searchable text file. This is accomplished as the OCR software analyzes scanned page images, recognizes groups of characters (words), compares them against its dictionary, and finally translates the characters into machine-readable digital text format. Cultural heritage institutions have used a wide range of proprietary OCR software, including ABBYY FineReader, Adobe Acrobat, OmniPage, or Readiris Pro. There is also a growing interest in using open source OCR tools, such as Tesseract, that provide more customization and specifically support OCR processing of digitized historical documents (Blanke et al., 2012). OCR technology is primarily used for creating searchable text of digitized books, journals, and newspapers. Increasingly, it is also viewed as an important component in large-scale digitization of modern archival collections (Miller, 2013).

The accuracy of text produced with OCR software varies. The high rate of 99% can be reached in recognition of English-language typed or printed legal or business documents (Rice et al., 1996). However, accuracy decreases for historical documents, such as newspapers, where the rate of uncorrected OCR can be as low as 68% (Klijn, 2008). The performance of OCR software plays a role, but the accuracy of the processed text also depends on a number of other factors, including:

- Quality of the scanned images
- Legibility of the original text
  - fonts
  - contrast between printed text and page background
  - layout
- Language and script

The quality of OCR text is related to the condition of the original source materials (Klijn, 2008). The accuracy rates decrease for historical documents printed with rare and difficult-to-recognize fonts and for documents with complex layouts, such as newspapers (Holley, 2009; Tanner et al., 2009). Non-Latin language scripts, such as Arabic, pose another challenge. Digitization projects involving historic Arabic-language periodicals demonstrate relatively low accuracy of OCR text (Matusiak and Abu Harb, 2011). In addition, the artifacts present in historical documents, such as tears, speckles, poor printing, or bleeding, reduce the quality of OCR output.

Photographic images represent the most difficult materials to convert in the 2D static media category, and yet they are often the first candidates for digitization projects undertaken by cultural heritage institutions. The authors of *The Minimum Digitization Capture Recommendations* emphasize, "accurately reformatting historic photographs is among the most challenging of the static media types" (ALCTS, 2013, p. 23). This difficulty stems from the diversity of analog materials that have been used to record photographic images, from glass plates to different types of negatives. The original negatives and prints come in different sizes ranging from 35 mm to 8 × 10 in. Further, nitrate- and acetate-based film negatives used for several decades of the 20th century are chemically unstable and present preservation risks. However, as previously discussed, digitization provides a tremendous opportunity to capture the visual information of deteriorating film-based materials. The negatives may require some conservation efforts such as cleaning or straightening prior to scanning. Moreover, representing the visual content of photographs in the digital format presents a unique set of challenges related to tone and color reproduction.

The presence of original negatives along with prints in many archival photographic collections causes a dilemma in choosing which source to digitize. In general, it is recommended to digitize from the most original source (i.e., the negative). The general digitization principle, "digitize an original or first generation of the source material," is especially applicable to photographic collections. Frey (2000b) points out that "because every generation of photographic copying involves some quality loss, using intermediates immediately implies some decrease in quality" (p. 114). There are, however, some exceptions to this rule, especially in cases where there are substantial differences between the negative and the print. In some cases, photographic prints have been custom made, and scans from these derivatives will surpass a straightforward scan from an original negative. The negatives also can be in poor condition due to deterioration. In such situations, scanning from an intermediate is a better solution. In the case of artistic photography, it makes sense to scan both the negative and the print(s) if both are available.

The physical characteristics of the photographs need to be assessed in order to select the most appropriate equipment and to determine the technical specifications. The medium, format, and size are the primary factors affecting the equipment selection. Original source collections need to be evaluated with respect to:

- Number of images to capture
- Size of the original photographs
- Format and medium—reflective (photographic prints) versus transparent (film negatives, glass plates, or slides)
- Condition and unique characteristics of the original items
- Requirements for handling fragile originals

The size of original photographs is particularly important because digitization recommendations for photographic images are commonly based on the image spatial dimensions. Since film negatives and prints come in different sizes, a common practice is to use the number of pixels on the long dimension as a measure and adjust the resolution accordingly. In the early digitization projects, 4000 pixels on the long side were recommended as a minimum dimension. With improvements in the capabilities of imaging equipment and the lower cost of digital storage, 6000–8000 pixels on the long edge are currently recommended, especially for larger transparencies.

Table 3.3 provides a list of recommended technical specifications for a range of reflective and transparent photographic materials. The table does not cover all photographic types and sizes. The goal of this sample is to demonstrate that there is a significant variation in resolution that needs to be adjusted according to the size and type of the source item. This summary is based on the minimum recommendations prepared by the Association for Library Collections and Technical Services (ALCTS, 2013). The ALCTS guidelines include minimum specifications for other types of static media, such as maps, drawings, aerial photography, etc. As emphasized in the ALCTS document, these recommendations serve as a starting point, and many still images may require higher resolution and greater bit depth (16 bit for grayscale and 48 bit for color).

Table 3.3 Minimum Digitization Recommendations for Photographic Images

Original Item Dimensions	Scanning Resolution	Bit Depth	Capture Mode	Spatial Dimensions
Reflective Materials: Photographic prints				
8 × 10 in. print	400 ppi	8 bit or 24 bit	Grayscale or color	3200 × 4000 pixels
5 × 7 in. print	625 ppi	8 bit or 24 bit	Grayscale or color	3125 × 4375 pixels
4 × 5 in. print	800 ppi	8 bit or 24 bit	Grayscale or color	3200 × 4000 pixels
4 × 2.5 in. print	1200 ppi	8 bit or 24 bit	Color	4800 × 3000 pixels
Transparent Materials: Film negatives and slides				
8 × 10 in. film negative	800 ppi	8 bit or 24 bit	Grayscale or color	6400 × 8000 pixels
4 × 5 in. film negative	1200 ppi	8 bit or 24 bit	Grayscale or color	4800 × 6000 pixels
35 mm film negative or slide	4000 ppi	8 bit or 24 bit	Grayscale or color	5480 pixels on the long edge

The recommendations for digital capture of text and photographs have evolved slightly in response to changing technology, but the basic specifications have been relatively stable in the past two decades. Although there are no formal standards, the existing digitization guidelines and practices for static 2D materials have been well established. In contrast to preservation formats for time-based media that are still evolving, TIFF as an archival master format for text and images is stable and widely accepted. Digitization of textual documents and photographs is a matter of efficient practice rather than experimentation, although the conversion of versatile photographic materials is not free of unique challenges. In 2002, Lynch commented on the progress in digitization of cultural heritage materials: "we're getting pretty good at digitizing material at scale. We have a wealth of experience and a large number of successful projects" (Lynch, 2002, p. 3). In the past decade, cultural heritage institutions have built upon the early experiences and have increased digitization efforts of unique archival and special collections, although not on a massive scale.

RAPID DIGITIZATION

Rapid approaches have been proposed as an alternative to the resource-intensive, preservation-quality digitization to scale up the conversion of archives and special collections (Erway, 2011; Erway and Schaffner, 2007). Erway and Schaffner (2007) note, "as a community, we have spent more than two decades painstakingly pursuing the highest quality in our digitization of primary resources" (p. 2). The authors of this influential OCLC report recommend "shifting gears" and adopting a more flexible approach to digitizing archives and special collections. The impacts of the Google Book Project, changing user expectations, shrinking budgets, and a desire to maximize digitization efforts, are mentioned as primary reasons for shifting into rapid capture techniques and strategies.

In the archival community, the interest in mass digitization has also been spurred by the discussion about revamping the traditional processing practices and adopting the principle of More Product, Less Process (MPLP) (Greene and Meissner, 2005). Meissner and Greene (2010) address the debate surrounding MPLP in the follow-up article and articulate the general principles of MPLP:

- Make user access paramount: get the most material available in a usable form in the briefest time possible
- Expend the greatest effort on the most deserving or needful materials
- Establish an acceptable minimum level of work, and make it the processing benchmark
- Embrace flexibility: don't assume all collections, or all collection components, will be processed to the same level
- Don't allow preservation anxieties to trump user access and higher managerial values (pp. 175–176)

Although the MPLP principles were originally proposed to alter the practices in processing of physical archives, they have been also adopted in digitization to advance the large-scale conversion of archival and museum collections (Miller, 2013; Moore, 2014; Sutton, 2012).

The overall goal of rapid capture is to lower the cost of the conversion process and accelerate the rate of digitization in order to deliver more content to users. This approach has been proposed for user-initiated digitization (Schaffner et al., 2011) as well as for the mass conversion of archival and other unique cultural heritage collections (Miller, 2013; Moore, 2014; Patzke and Thiel, 2009; Rinaldo et al., 2011; Sutton, 2012). Schaffner et al. (2011) emphasize that user needs and improved access must drive

all digitization efforts and add that "user requests must not be bogged down by fine-tuning images and metadata" (p. 6). In the context of archival collections, rapid imaging is seen as a solution not only to speed up the process of conversion and increase access but also to address the issue of backlogs in archival processing (Meissner and Greene, 2010; Miller, 2013; Moore, 2014).

The terms rapid capture and rapid imaging are often used interchangeably. The approaches to speeding up the conversion process and digitizing at scale, however, often go beyond the imaging phase. A number of minimalist strategies can also be applied to selection, metadata creation, and quality control to reduce the amount of time and resources devoted to each step. Those strategies focus on establishing an acceptable minimum level of work and applying it consistently across the entire project. Rapid digitization is used here as an umbrella terms that encompasses a range of techniques and strategies, including:

- *Selecting en masse* entire collections without "cherry-picking" individual items. Miller (2013) recommends digitizing archival collections in their entirety, without archival processing, at the point of accession. The focus on scanning an entire collection removes an element of selection that can slow down the digitization process.
- *Adopting minimum technical standards* during image capture and using the preset standards for resolution, color mode, and format for all items in the project. For example, access copies of archival textual documents can be created by scanning in the bitonal or grayscale mode (rather than color), at a lower resolution, and saved in the PDF or JPEG format, without creating archival TIFF files. This approach implies that preset technical specifications are applied to the entire project, which works well for collections with materials in uniform formats. However, it is problematic for collections with versatile formats where technical specifications need to be adjusted based on the size and physical characteristics of original items. Grouping items with similar characteristics and scanning them in bulk is an effective strategy for improving efficiency, while maintaining the quality of digitized items.
- *Automating image capture process* or parts of it. The use of rapid imaging equipment and techniques varies depending on the type and format of original materials. For example, loose-leaf or unbound, nonrare documents can be scanned efficiently using sheet-fed scanners (Moore, 2014). Scanning robots with a high-speed page flipping mechanism can be used to digitize books and other bound materials at extremely high volume, although they are not suitable for books with foldouts or loose pages (Rinaldo et al., 2011). Other semi-automated technologies include the use of conveyor belts to move items quickly through imaging systems. Fig. 3.11 demonstrates a rapid capture system with a conveyor belt that has been developed by Picturae, a company based in the Netherlands. The system with a conveyor has been used for digitizing at high speed the Herbarium sheets for the Naturalis Biodiversity Center in Leiden, Netherlands. The system is capable of digitizing more than 40,000 Herbarium sheets per day. The same system has been recently used at the Smithsonian Institution for digitizing a large numismatic collection from the National Museum of American History (Kutner, 2015). The collection of 250,000 historic bank notes became the Smithsonian's first full production rapid capture digitization project (Kutner, 2015).
- *Applying minimum metadata* for item and collection-level description. The reduction in the amount of resources devoted to metadata creation is one of the hallmarks of mass approaches. Generally, no new descriptive metadata is generated in large-scale projects to avoid the

FIGURE 3.11 A Rapid Capture Digitization System with a Conveyor Belt Developed by Picturae

© Picturae, Picturae's Herbarium Digistreet at Naturalis Biodiversity Center, used with permission.
https://picturae.com/uk/digitising/herbarium-sheets

resource-intensive research process that accompanies metadata creation on an item level. Only preexisting descriptive information is used for item-level metadata without additional subject headings or other access points (Moore, 2014; Sutton, 2012). For archival collections, metadata can be created on a folder or series level and supplemented by full-text searchability generated by OCR (Miller, 2013; Moore, 2014). In addition to the use of OCR technology, the proponents of the mass approach also see user-generated tags as a possible source of descriptive information (Miller, 2013; Sutton, 2012).

- *Streamlining workflows* and integrating digitization activities that can involve digitizing the same items in bulk or scanning bound volumes or a group of archival materials in a folder as single digital objects. Selecting items of the same size or in close range improves the efficiency of image capture process since the camera does not need to be refocused or recalibrated as frequently. Large-scale conversion of archival collections moves away from scanning individual items and creating item-level metadata to folder-level digitization and collective description (Miller, 2013). Streamlining also includes replicating and reusing the established standards and procedures.

The outlined strategies and techniques can be used selectively or in combination in order to facilitate faster production and large-scale digitization. Streamlining workflows is the strategy that can be used effectively in almost all projects. Erway (2011) acknowledges the difficulties of digitizing special collections at scale but also notes that "what makes a capture operation efficient is the ability to streamline workflows by setting up equipment and workflows for one set of characteristics and then capturing a mass of similar items, thereby limiting the adjustments done in between captures" (p. 17).

Rapid digitization strategies and technologies have been tested in pilot projects and in some cases integrated into regular digitization programs to increase productivity. The Smithsonian Institution has used rapid captures in a number of prototypes and in large-scale production (Crawford, n.d.; Kutner, 2015). Erway (2011) reviews a number of large-scale digitization projects in libraries, archives, and museums to see how ideas of rapid capture are being put into practice. The review focuses on equipment, throughput, and bottlenecks in the digital capture process and presents a variety of strategies and challenges due to the heterogeneous nature of special collections.

Moore (2014) reports on a case study of implementing a range of rapid capture strategies and workflows in digitizing a collection of papers at the University of Minnesota Archives. The collection of university papers and publication produced and distributed en masse, in standard document formats, and primarily with textual content made a very good candidate for routine digitization. The papers were digitized according to the preset minimum standards using a scanner with an automatic document feeder. Searchable text was generated through OCR, and only minimal descriptive metadata were added. The case of digitizing university papers at the University of Minnesota Archives demonstrates that rapid capture can be adopted successfully when materials are relatively homogenous in nature and are being digitized for access, primarily for their informational content rather than intrinsic value.

Rapid and minimalist strategies, however, can rarely be implemented in such a uniform and straightforward manner in the digitization of unique archival and special collection materials. Collections consisting of a variety of historic textual and visual resources typically require varying levels of detail in image capture and resource description. While searchable text can be generated through OCR for printed materials, manuscripts, images, and sound and video resources are difficult to discover without accompanying metadata. The minimalist approach to metadata is particularly debatable since putting large quantities of digitized materials online, without accurate description, does not guarantee that resources will be discovered, especially if users rely on keyword searching. More user studies of user information seeking and use behaviors in the context of large-scale digital projects are needed to examine if the mass approach does indeed serve user needs better. Some researchers recognize the challenges associated with the large-scale digitization initiatives, especially in digitizing unique materials, and emphasize the need to balance speed with quality and completeness (Rieger, 2010).

In practice, some projects assume a hybrid approach and attempt to balance minimalist strategies with preservation-quality, so called "boutique" digitization. Sutton (2012) presents a case study of digitizing correspondence, photographs, journals, and drawings of the John Muir Papers at the University of the Pacific Library. The project adopted rapid capture in digitizing correspondence but used high-resolution color scanning for photographs, drawings, and journals to provide greater detail and clarity. Minimum metadata were applied consistently to all digitized items, although transcripts have been created for correspondence. The author acknowledges that strictly minimalist metadata practices may be challenging for resource discovery and notes that the impact "needs to be fully assessed to ensure that this approach does not overly compromise the ability to meet user needs and expectations for discoverability in the online environment" (Sutton, 2012, p. 58).

Rapid capture techniques and strategies pose a number of challenges in regard to resource discovery and quality of digitized materials. Furthermore, the issues of digital preservation have not been discussed in the context of mass projects. A number of questions remain unanswered about the level of digital preservation for materials generated primarily for access and which quality may not be acceptable for long-term preservation. Rapid approaches, however, are an indication of a maturing digitization landscape and recognition that archival and special collections in a stable condition may require

more diversified conversion standards and processes. The current practice allows for a differentiation between preservation-quality digitization, rapid capture focused on increasing access, and hybrid models with varying levels of imaging and metadata standards.

In addition to undertaking the conversion of archival and special collections on a mass scale, the digitization of audiovisual collections represents a current and challenging area of research and practice. The issues associated with the conversion of historical time-based collections are discussed in Chapter 4.

REFERENCES

Adams, C., 2013. Is JPEG-2000 a preservation risk? The Signal: Digital Preservation. Library of Congress blog, (January 28). Available from: http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/.

ALCTS, 2013. Minimum digitization capture recommendations. Association for Library Collections & Technical Services. Division of the American Library Association. Available from: http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations.

Arthur, K., Byrne, S., Long, E., Montori, C.Q., Nadler, J., 2004. Recognizing digitization as a preservation reformatting method. Microform Imaging Rev. 33 (4), 171–180.

BCR, 2008. BCR's CDP digital imaging best practices [updated version of Western States digital imaging best practices]. Bibliographical Center for Research. Available from: http://mwdl.org/docs/digital-imaging-bp_2.0.pdf.

Besser, H., 2003. Introduction to Imaging. Getty Publications, Los Angeles, CA.

Blanke, T., Bryant, M., Hedges, M., 2012. Open source optical character recognition for historical research. J. Doc. 68 (5), 659–683.

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

Buckley, R., 2008. JPEG 2000: A practical digital preservation standard? Technology Watch Report 08-01. Digital Preservation Coalition.

Buonora, P., Liberati, F., 2008. A format for digital preservation of images. D-Lib. Mag. 14 (7/8), 1.

California Digital Library (CDL), 2011. CDL Guidelines for Digital Images. Version 2.0. Available from: http://www.cdlib.org/services/access_publishing/dsc/contribute/docs/cdl_gdi_v2.pdf.

Capell, L., 2010. Digitization as a preservation method for damaged acetate negatives: a case study. Am. Arch. 73 (1), 235–249.

Chapman, S., 2000. Considerations for project management. In: Sitts, M.K. (Ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center, Andover, MA, pp. 31–44.

Chowdhury, G., Chowdhury, S., 2003. Introduction to Digital Libraries. Facet Publishing, London.

Cohen, D.J., Rosenzweig, R., 2006. Digital history: A Guide to Gathering, Preserving, and Presenting the Past on the web. University of Pennsylvania Press, Philadelphia. Available from: http://chnm.gmu.edu/digitalhistory/.

Collmann, R., 2011. Developments in virtual 3D imaging of cultural artefacts. Ariadne (66), 4. Available from: http://www.ariadne.ac.uk/issue66/collmann.

Conway, P., 2000. Overview: rationale for digitization and preservation. In: Sitts, M.K. (Ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center, Andover, MA, pp. 5–20.

Conway, P., 2008. Best practices for digitizing photographs: a network analysis of influences. In: Proceedings of IS&T Archiving 2008, Imaging Science & Technology, Berne, Switzerland, June 24–27, pp. 94–102. Available from: http://deepblue.lib.umich.edu/handle/2027.42/85229.

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

Coyle, K., 2006. Mass digitization of books. J. Acad. Librariansh. 32 (6), 641–645.

Crawford, K., n.d. Rapid capture open house at the Archives of American Gardens. Smithsonian Institution. Digitization Program Office blog. Available from: http://dpo.si.edu/blog/rapid-capture-open-house-archives-american-gardens.

Daigle, B.J., 2012. The digital transformation of special collections. J. Lib. Adm. 52 (3/4), 244–264.

Dale, R.L., 2007. Outsourcing and Vendor Relations. Preservation Leaflets, 6.7. Northeast Document Conservation Center, Andover, MA. Available from: http://www.nedcc.org/free-resources/preservation-leaflets/6.-reformatting/6.7-outsourcing-and-vendor-relations.

Dooley, J.M., Luce, K., 2010. Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives. OCLC Research, Dublin, OH. Available from: http://www.oclc.org/content/dam/research/publications/library/2010/2010-11.pdf?urlm=162945.

Dryden, J., 2008. PDF/A-1: a ray of light in the digital dark age? J. Arch. Org. 6 (1/2), 121–124.

Erway, R., 2011. Rapid Capture: Faster Throughput in Digitization of Special Collections. OCLC Research, Dublin, OH. Available from: http://www.oclc.org/research/publications/library/2011/2011-04.pdf.

Erway, R., Schaffner, J., 2007. Shifting Gears: Gearing Up to Get Into the Flow. OCLC Programs and Research, Dublin, OH.

Ester, M., 1996. Digital Image Collections: Issues and Practice. Commission on Preservation & Access, Washington, DC.

FADGI, 2010. The Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files. Federal Agencies Digitization Guidelines Initiative. Available from: http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf.

FADGI, 2014. Raster Still Images for Digitization: A Comparison of File Formats. Federal Agencies Digitization Guidelines Initiative. Available from: http://www.digitizationguidelines.gov/guidelines/FADGI_RasterFormatCompare_p3_20140417.pdf.

FADGI, n.d. Glossary. Federal Agencies Digitization Guidelines Initiative. Available from: http://www.digitizationguidelines.gov/glossary.php.

Fleischhauer, C., 2014a. Comparing formats for still image digitizing: Part one. The Signal: Digital Preservation. Library of Congress blog (May 14). Available from: http://blogs.loc.gov/digitalpreservation/2014/05/comparing-formats-for-still-image-digitizing-part-one/.

Fleischhauer, C., 2014b. Comparing formats for still image digitizing: Part two. The Signal: Digital Preservation. Library of Congress blog (May 15). Available from: http://blogs.loc.gov/digitalpreservation/2014/05/comparing-formats-for-still-image-digitizing-part-two/.

Frey, F., 2000a. File formats for digital masters. Guides to Quality in Visual Resource Imaging Council on Library and Information; Digital Library Federation; Research Libraries Group, Washington, DC. Available from: http://oclc.org/research/publications/library/visguides/visguide5.html.

Frey, F., 2000b. Why are photographs different? In: Sitts, M.K. (Ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center, Andover, MA, pp. 111–119.

Frey, F., Reilly, J.M., 1999. Digital Imaging for Photographic Collections: Foundations for Technical Standards. Image Permanence Institute, Rochester.

Gertz, J., 2007. Preservation and Selection for Digitization. Northeast Document Conservation Center, Andover, MA. Available from: http://www.nedcc.org/resources/leaflets/6Reformatting/06PreservationAndSelection.php.

Gould, S., Ebdon, R., 1999. IFLA/UNESCO Survey on Digitisation and Preservation. IFLA Offices for UAP and International Lending in cooperation with IFLA Programme for Preservation and Conservation, Boston, MA.

Greene, M., Meissner, D., 2005. More product, less process: revamping traditional archival processing. Am. Arch. 68 (2), 208–263.

Holley, R., 2009. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib. Mag. 15 (3/4). Available from: http://www.dlib.org/dlib/march09/holley/03holley.html.

IMLS, 2006. Status of Technology and Digitization in the Nation's Museums and Libraries. Institute of Museum and Library Services. Available from: http://www.imls.gov/assets/1/AssetManager/Technology_Digitization.pdf.

JISC Digital Media, 2014. Project management for a digitisation project. Available from: http://www.jiscdigitalmedia.ac.uk/guide/project-management-for-a-digitisation-project/.

Kenney, A.R., 2000. Projects to programs: mainstreaming digital imaging initiatives. In: Kenney, A.R., Rieger, O. (Eds.), Moving Theory into Practice: Digital Imaging for Libraries and Archives. Research Libraries Group, Mountain View, CA, pp. 153–175.

Kenney, A.R., Chapman, S., 1996. Digital Imaging for Libraries and Archives. Cornell University Libraries, Ithaca, NY.

Kenney, A.R., Rieger, O.Y., 2000. Introduction. In: Kenney, A.R., Rieger, O. (Eds.), Moving Theory into Practice: Digital Imaging for Libraries and Archives. Research Libraries Group, Mountain View, CA, pp. 1–10.

Kenney, A.R., Rieger, O.Y., Entlich, R., 2000. Moving Theory into Practice: Digital Imaging Tutorial. Cornell University Library, Ithaca, NY. Available from: https://www.library.cornell.edu/preservation/tutorial/.

Klijn, E., 2008. The current state-of-art in newspaper digitization: a market perspective. D-Lib. Mag. 14 (1). Available from: http://www.dlib.org/dlib/january08/klijn/01klijn.html.

Koelling, J.M., 2004. Digital Imaging: A Practical Approach. Altamira Press, Walnut Creek, CA.

Kutner, M., 2015. Museums are now able to digitize thousands of artifacts in just hours. Available from: http://www.smithsonianmag.com/smithsonian-institution/museums-are-now-able-digitize-thousands-artifacts-just-hours-180953867/?no-ist.

Lazorchak, B., 2014. New NDSA Report: The benefits and risks of the PDF/A-3 file format for archival institutions. The Signal: Digital Preservation. Library of Congress blog (February 20). Available from: http://blogs.loc.gov/digitalpreservation/2014/02/new-ndsa-report-the-benefits-and-risks-of-the-pdfa-3-file-format-for-archival-institutions/.

Lee, S.D., 2001. Digital Imaging: A Practical Handbook. Neal-Schuman Publishers, New York, NY.

Leetaru, K., 2008. Mass book digitization: the deeper story of Google Books and the Open Content Alliance. First Monday 13 (10).

Lesk, M., 2004. Understanding Digital Libraries. Morgan Kaufmann, Boston.

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

Lopatin, L., 2006. Library digitization projects, issues and guidelines: a survey of the literature. Libr. Hi Tech 24 (2), 273–289.

Lynch, C., 2002. Digital collections, digital libraries and the digitization of cultural heritage information. First Monday 7 (5–6). Available from: http://firstmonday.org/issues/issue7_5/lynch/index.html.

Marcum, D.B., 2007. Digitizing for access and preservation: strategies of the Library of Congress. First Monday 12 (7), 1. Available from: http://firstmonday.org/ojs/index.php/fm/article/view/1924/1806.

Matusiak, K.K., Abu Harb, Q., 2011. Digitizing the historical periodical collection at the Al-Aqsa Mosque Library in East Jerusalem. In: Walravens, H. (Ed.), Newspapers: Legal Deposit and Research in the Digital Era. De Gruyter, Berlin, pp. 271–290.

Matusiak, K.K., Johnston, T., 2014. Digitization for preservation and access: restoring the usefulness of the nitrate negative collections at the American Geographical Society Library. Am. Arch. 77 (1), 241–269.

Meissner, D., Greene, M.A., 2010. More application while less appreciation: the adopters and antagonists of MPLP. J. Arch. Org. 8 (3–4), 174–226.

Miller, L.K., 2013. All text considered: a perspective on mass digitizing and archival processing. Am. Arch. 76 (2), 521–541.

Moore, E.A., 2014. Strategies for implementing a mass digitization program. Prac. Technol. Arch., 3, (November 2014). Available from: http://practicaltechnologyforarchives.org/issue3_moore/.

NISO Framework Working Group, 2007. A Framework of Guidance for Building Good Digital Collections, 3rd ed. NISO Framework Working Group. Available from: http://www.niso.org/publications/rp/framework3.pdf.

Panitch, J.M., 2001. Special collections in ARL libraries: results of the 1998 survey sponsored by the ARL Research Collections Committee. Association of Research Libraries.

Patzke, K., Thiel, S.G., 2009. Digital Imaging: Theory Joins Practice. In Ross, D.I. (Ed.), New issues in librarianship: juried papers at the ALA 2009 ALA Annual Conference, American Library Association, Chicago. Available from: http://kuscholarworks.ku.edu/dspace/bitstream/1808/11178/1/theory_practice.pdf.

Puglia, S., 2000. Technical primer. In: Sitts, M.K. (Ed.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center, Andover, MA, pp. 83–102.

Puglia, S., 2011. Choosing and using digitization technologies. Keynote Presentation at the Digital Commonwealth, Fifth Annual Digital Library Conference, Danvers, MA, April 26, 2011. Available from: http://www.masshist.org/pub/digicomm/digicomm_2011conf_puglia.pdf.

Puglia, S., Reed, J., Rhodes, E., 2004. Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files-Raster Images. National Archives and Research Administration, Washington, DC. Available from: http://www.archives.gov/preservation/technical/guidelines.pdf.

Rice, S.V., Jenkins, F.R., Nartker, T.A., 1996. The Fifth Annual Test of OCR Accuracy. University of Nevada Las Vegas, Information Science Research Institute, Las Vegas. Available from: http://www.stephenvrice.com/images/AT-1996.pdf.

Rieger, O.Y., 2000. Establishing a quality control program. In: Kenney, A.R., Rieger, O. (Eds.), Moving Theory into Practice: Digital Imaging for Libraries and Archives. Research Libraries Group, Mountain View, CA, pp. 61–83.

Rieger, O.Y., 2008. Preservation in the Age of Large-Scale Digitization: A White Paper, CLIR Publication 141. Council on Library and Information, Washington, DC.

Rieger, O.Y., 2010. Enduring access to special collections: challenges and opportunities for large-scale digitization initiatives. RBM 11 (1), 11–22.

Rinaldo, C., Warnement, J., Baione, T., Kalfatovic, M.R., Fraser, S., 2011. Retooling special collections digitisation in the age of mass scanning. Ariadne 67. Available from: http://www.ariadne.ac.uk/issue67/rinaldo-et-al.

Romero, V., Serrano, N., Toselli, H.A., Sanchez, A.J., Vidal, E., 2011. Handwritten text recognition for historical documents. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage. Available from: http://aclweb.org/anthology//W/W11/W11-4114.pdf.

Sánchez, J.A., Bosch, V., Romero, V., Depuydt, K., de Does, J., 2014. Handwritten text recognition for historical documents in the transcriptorium project. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, ACM, pp. 111–117.

Schaffner, J., Snyder, F., Supple, S., 2011. Scan and Deliver! Managing User-Initiated Digitization in Special Collections and Archives. OCLC Research, Dublin, OH. Available from: http://netweb.oclc.org/content/dam/research/publications/library/2011/2011-05.pdf.

Smith, A., 1999. Why Digitize? Council on Library and Information Resources, Washington, DC. Available from: www.clir.org/pubs/reports/pub80-smith/pub80.html.

Smith, A., 2001. Strategies for Building Digitized Collections. Digital Library Federation, Council on Library and Information Resources, Washington, DC.

Stroeker, N., Voegels, R., 2012. Survey Report on Digitisation in European Cultural Heritage Institutions 2012. ENUMERATE Thematic Framework. Available from: http://www.enumerate.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf

Surendran, N., Xu, X., Stead, O., Silyn-Roberts, H., 2009. Contemporary technologies for 3D digitization of Maori and Pacific Island artifacts. Int. J. Imaging Syst. Technol. 19, 244–259.

Sutton, S.C., 2012. Balancing boutique-level quality and large-scale production: the Impact of "more product, less process" on digitization in archives and special collections. RBM 13 (1), 50–63.

Tanner, S., 2001. Librarians in the digital age: planning digitisation projects. Program 35 (4), 327–337.

Tanner, S., Muñoz, T., Ros, P.H., 2009. Measuring mass text digitization quality and usefulness: lessons learned from assessing the OCR accuracy of the British Library's 19th century online newspaper archive. D-Lib. Mag. 15 (7/8). Available from: http://www.dlib.org/dlib/july09/munoz/07munoz.html.

Turró, M., 2008. Are PDF documents accessible? Inform. Technol. Libr. 27 (3), 25–43.

Valentino, M., Shults, B., 2012. Creating a digital library of three-dimensional objects in CONTENTdm. OCLC Syst. Serv. 28 (4), 208–220.

Van der Knijff, J., 2011. JPEG 2000 for long-term preservation: JP2 as a preservation format. D-Lib. Mag. 17 (5/6), 1–9. Available from: http://dlib.org/dlib/may11/vanderknijff/05vanderknijff.html.

Yale University, 2010. Digitization shared practices – still images. Version 1.0. Available from: http://www.yale.edu/digitalcoffee/downloads/DigitalCoffee_SharedPractices_%5Bv1.0%5D.pdf.

Yongli, Z., 2010. Are your digital documents web friendly? Making scanned documents web accessible. Inform. Technol. and Libr. 29 (3), 151–160.

Zhang, A.B., Gourley, D., 2009. Creating Digital Collections: A Practical Guide. Chandos, Oxford.