

DIGITIZATION AND DIGITAL ARCHIVING

SECOND EDITION



A Practical Guide for Librarians

ELIZABETH R. LEGGETT

Digitization and Digital Archiving

PRACTICAL GUIDES FOR LIBRARIANS

About the Series

This innovative series written and edited for librarians by librarians provides authoritative, practical information and guidance on a wide spectrum of library processes and operations.

Books in the series are focused, describing practical and innovative solutions to a problem facing today's librarian and delivering step-by-step guidance for planning, creating, implementing, managing, and evaluating a wide range of services and programs.

The books are aimed at beginning and intermediate librarians needing basic instruction/guidance in a specific subject and at experienced librarians who need to gain knowledge in a new area or guidance in implementing a new program/service.

About the Series Editor

The **Practical Guides for Librarians** series was conceived and edited by M. Sandra Wood, MLS, MBA, AHIP, FMLA, Librarian Emerita, Penn State University Libraries from 2014–2017. M. Sandra Wood was a librarian at the George T. Harrell Library, the Milton S. Hershey Medical Center, College of Medicine, Pennsylvania State University, Hershey, PA, for more than thirty-five years, specializing in reference, educational, and database services. Ms. Wood received an MLS from Indiana University and an MBA from the University of Maryland. She is a fellow of the Medical Library Association and served as a member of MLA's Board of Directors from 1991 to 1995.

Ellyssa Kroski assumed editorial responsibilities for the series beginning in 2017. She is the director of information technology at the New York Law Institute as well as an award-winning editor and author of thirty-six books, including *Law Librarianship in the Digital Age*, for which she won the AALL's 2014 Joseph L. Andrews Legal Literature Award. Her ten-book technology series, *The Tech Set*, won the ALA's Best Book in Library Literature Award in 2011. Ms. Kroski is a librarian, an adjunct faculty member at Drexel University and San Jose State University, and an international conference speaker. She was the winner of the 2017 Library Hi Tech Award from the ALA/LITA for her long-term contributions in the area of library and information science technology and its application.

Recent Books in the Series Include:

50. *Gaming Programs for All Ages in the Library: A Practical Guide for Librarians*, by Tom Bruno
51. *Intentional Marketing: A Practical Guide for Librarians*, by Carol Ottolenghi

52. *Electronic Resources Librarianship: A Practical Guide for Librarians*, by Holly Talbott and Ashley Zmau
53. *Citation Management Tools: A Practical Guide for Librarians*, by Nancy R. Glassman
54. *Embedded and Empowered: A Practical Guide for Librarians*, by Courtney Mlinar
55. *Creating a Learning Commons: A Practical Guide for Librarians*, by Lynn D. Lampert and Coleen Meyers-Martin
56. *Graphic Design: A Practical Guide for Librarians*, by Valerie Colston
57. *Creating a Tween Collection: A Practical Guide for Librarians*, by Karen M. Smith
58. *Teaching First-Year College Students: A Practical Guide for Librarians*, by Maggie Murphy with Adrienne Button
59. *Reaching Diverse Audiences with Virtual Reference and Instruction: A Practical Guide for Librarians*, by Meredith Powers and Laura Costello
60. *How to Write and Get Published: A Practical Guide for Librarians*, by Tammy Ivins and Anne Pemberton
61. *Library Programming Made Easy: A Practical Guide for Librarians*, by Michelle Demeter and Haley Holmes
62. *Library Volunteers: A Practical Guide for Librarians*, by Allison Renner
63. *Developing a Residency Program: A Practical Guide for Librarians*, by Lorelei Rutledge, Jay L. Colbert, Anastasia Chiu, and Jason Alston
64. *Yoga and Meditation at the Library: A Practical Guide for Librarians*, by Jenn Carson
65. *Fundraising for Academic Libraries: A Practical Guide for Librarians*, by Karlene Noel Jennings and Joyce Garczynski
66. *Developing a Library Accessibility Plan: A Practical Guide for Librarians*, by Rebecca M. Marrall
67. *Terrific Makerspace Projects: A Practical Guide for Librarians*, by Juan Denzer and Sharona Ginsberg
68. *Systems Librarianship: A Practical Guide for Librarians*, by Brighid M. Gonzales
69. *E-Textiles in Libraries: A Practical Guide for Librarians*, by Helen Lane and Carli Spina
70. *Anime Clubs for Public Libraries: A Practical Guide for Librarians*, by Chantale Pard
71. *Digitization and Digital Archiving: A Practical Guide for Librarians*, Second Edition, by Elizabeth R. Leggett

Digitization and Digital Archiving

A Practical Guide for Librarians



Second Edition

Elizabeth R. Leggett

PRACTICAL GUIDES FOR LIBRARIANS, NO. 71

ROWMAN & LITTLEFIELD
Lanham • Boulder • New York • London

Published by Rowman & Littlefield
An imprint of The Rowman & Littlefield Publishing Group, Inc.
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
www.rowman.com

6 Tinworth Street, London, SE11 5AL, United Kingdom

Copyright © 2021 by The Rowman & Littlefield Publishing Group, Inc.

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except by a reviewer who may quote passages in a review.

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Name: Leggett, Elizabeth R., 1985–, author.

Title: Digitization and digital archiving : a practical guide for librarians / Elizabeth R. Leggett.

Description: [Second edition]. | Lanham : The Rowman & Littlefield Publishing Group, [2021]

| Series: Practical guides for librarians; no. 71 | Includes bibliographical references and index.

| Summary: “This second edition has been updated to address new trends and concerns in software. It also addresses questions about preserving materials with no original physical format and preserving media storage devices as important artifacts in themselves”—Provided by publisher.


Identifiers: LCCN 2020035956 (print) | LCCN 2020035957 (ebook) | ISBN 9781538133347 (paperback) | ISBN 9781538133354 (ebook)

Subjects: LCSH: Archival materials—Digitization. | Archival materials—Conservation and restoration. | Digital preservation. | Records—Management. | Electronic records—Management.

Classification: LCC CD973.D53 L44 2021 (print) | LCC CD973.D53 (ebook) | DDC 025.3/4140285—dc23

LC record available at <https://lcn.loc.gov/2020035956>

LC ebook record available at <https://lcn.loc.gov/2020035957>

 The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

Contents



Illustrations	ix
Preface	xi
CHAPTER 1 Why Use Digital Preservation?	1
CHAPTER 2 How Do Computers Store Information?	11
CHAPTER 3 Storing Images	27
CHAPTER 4 Storing Text	41
CHAPTER 5 Storing Audio and Video	57
CHAPTER 6 Storing Born-Digital Materials	71
CHAPTER 7 Floppy Disks and Optical Media	81
CHAPTER 8 Magnetic Tape	97
CHAPTER 9 Hard Disk Drives	111
CHAPTER 10 Flash Memory	125
CHAPTER 11 Cloud Computing	139
CHAPTER 12 Equipment for Digitizing and Editing Archival Materials	155
CHAPTER 13 Metadata and Accessing Information	171
CHAPTER 14 Copyright Law	185

CHAPTER 15 Problems with Digital Archiving	203
CHAPTER 16 Drawing Up Policies	217
Index	237
About the Author	243

Illustrations



Figures

Figure 2.1.	A Typical RAM Chip	21
Figure 2.2.	Common Computer Ports	24
Figure 3.1.	Using a Histogram	39
Figure 4.1.	HTML Plain Text and Viewed in a Browser	49

Tables

Table 2.1.	Decimal and Equivalent Binary Numbers	14
Table 3.1.	Image File Types and Features	36
Table 7.1.	Common Floppy Disk Features	84

Textboxes

Textbox 1.1.	Possible Goals for a Digital Archive	7
Textbox 2.1.	Useful Terms for Storage Capacity	15
Textbox 2.2.	Typical Computer Components	17
Textbox 3.1.	Useful Terms for Visual Data	28
Textbox 3.2.	Some Common Image File Formats	33
Textbox 4.1.	Some Common Text File Extensions	45
Textbox 4.2.	Text File Formats for Archiving	46
Textbox 5.1.	Terms for Audio Data	59
Textbox 5.2.	Some Audio Codecs	61
Textbox 5.3.	Terms for Video Data	65
Textbox 5.4.	Common Video File Formats	66
Textbox 7.1.	Optical Media Acronyms	88
Textbox 7.2.	Optical Media Types	90
Textbox 7.3.	Ideal Storage and Handling of Optical Disks	94

Textbox 8.1. Advantages and Disadvantages of Tape Storage	103
Textbox 8.2. Ideal Storage and Handling of Magnetic Tape	108
Textbox 9.1. Qualities of a Hard Drive	115
Textbox 9.2. Advantages and Disadvantages of Hard Disk Drive Storage	118
Textbox 9.3. Ideal Storage and Handling of Hard Drives	122
Textbox 10.1. Flash Memory Cells	128
Textbox 10.2. Types of Flash Memory Devices	129
Textbox 10.3. Advantages and Disadvantages of Flash Memory Storage	130
Textbox 11.1. Four Models of Cloud Computing	145
Textbox 11.2. Advantages and Disadvantages of Cloud Computing Storage	146
Textbox 11.3. Things to Look For in a Contract	150
Textbox 12.1. Types of Scanners	161
Textbox 16.1. Items That Cannot Be Copyrighted	221
Textbox 16.2. Questions to Consider for Image Data Storage	224
Textbox 16.3. Questions to Consider for Text Data Storage	226
Textbox 16.4. Questions to Consider for Audio and Video Data Storage	228
Textbox 16.5. Features to Look For in Computers	230
Textbox 16.6. Features to Look For When Purchasing Monitors	232
Textbox 16.7. Metadata and Patron Access	235

Preface



The first edition of this book was published in 2014, a mere six years ago. For most subjects, this is a minuscule amount of time, and the differences between a first and second edition might be quite small. You might suppose, then, that a second edition of a book for librarians isn't going to be an awful lot different from the first.

While this book is aimed at librarians and personal archivists, it's really a book about technology, and for technology, six years is an *eternity*. Since the invention of the World Wide Web in the early 1990s, society has become more and more dependent upon the Internet as a means of communication and as a means of finding information. With this dependence comes the need and desire for ever faster, ever more convenient means of accessing this network of computers and the information held on the computers that make up the Internet. Whereas computers were uncommon, often specialized equipment only thirty years ago, today, an average person might own a desktop computer, laptop computer, tablet, and smartphone, all at once—not to mention things like digital assistants. Even our traditionally non-tech devices can now access the Internet. Vacuum cleaners, light bulbs, and thermostats can all be controlled remotely through the Internet using a smartphone—it's possible to go online using a refrigerator!

While no *major* breakthroughs in technology have occurred between the first and second editions of this book, that technology continues to evolve in ways that have a great impact on society, and new ideas that make computing faster, easier, and less expensive are occurring all the time. In the first edition, for instance, it was suggested that CDs were a potential method of backing up data for a small archive or one on a budget. While this is still true, the prices of other methods of data storage have dropped so vastly that it's no longer such a practical suggestion.

The rapid price drop and increase in storage space for flash memory in particular has had a radical impact on modern technology, making it possible to create ever smaller and more versatile computers. As stated earlier, computers capable of accessing the Internet are now ubiquitous.

As devices with computing abilities, devices that connect to the great network known as the Internet, continue to infiltrate everyday life, so our history becomes more and more entwined with these devices. The first edition of this book emphasized the importance of archiving this novel and complex point in history, and so the task only grows more in importance as technology continues to evolve.

As with the first edition, this book seeks to teach you the basics of digital archiving, both preserving materials that are digital in nature, such as software programs and e-mails, as well as creating digital copies of actual, tangible items.

Organization

Digitization and Digital Archiving: A Practical Guide for Librarians is divided into 16 chapters. Each chapter will provide a little history, explain a practical aspect of digital archiving, and help you to make decisions about how to go about forming your own archive.

In chapter 1, you'll start with the basics: learning about what digital archiving is and why it is becoming more prominent and ever more relevant. In chapter 2, you'll move on to learning the basics of how a computer operates and stores data, along with learning some basic computer terminology. Explanations of features to look for in a new computer are also covered.

In chapters 3–5, you'll learn about the optimal archiving formats for images, text, audio data, and video data, and you'll learn how computers recognize and store these types of data. New to this edition, chapter 6 will additionally cover information about some born-digital materials that you may want to preserve, such as software, databases, websites, and e-mail.

Chapters 7–11 will teach you about the common methods of data storage available today, which methods are best for your archive and why, and how the different common methods of data storage work on a physical level. Optical disks, magnetic tape, hard drives, flash memory, and cloud storage are all discussed in these chapters. Chapter 11 also provides an overview of how the Internet works. In this edition, floppy disks are also discussed, as this is a commonly found data storage method that is vulnerable and is a medium that may need to be addressed in some archiving projects.

You'll need some equipment for any digital archiving project, even if it is merely a computer and monitor, and so chapter 12 provides an overview of basic equipment that you may need for your project, explaining what you may need and situations in which certain types of equipment are optimal. You will also learn a little about some of the types of software you may need.

Metadata is an important part of any project and is key to making data usable and accessible, and so chapter 13 will discuss what metadata is and why it is necessary. It additionally addresses concerns that you may have about how patrons access your files and how much access they should have. Chapter 14 helps you avoid legal issues in regard to copyright law, and chapter 15 teaches you about the limitations of digital archiving. At the end, you'll review the important parts of this book, putting it all together in chapter 16.

Using This Book

It's often assumed that everyone has experience with computers. If you do not, don't worry: this book does not make assumptions about your skill level. If you can type a word document, open an image, send an e-mail, and do a web search, you probably know enough to use this book. If you want to create databases or put your collection online, these are a little more complicated tasks, but anyone can create a digital backup of nearly

any kind of information and store it under safe conditions to preserve it for the future. Even if you do feel like you're comfortable with technology, this book seeks to enrich your knowledge, giving you a deeper understanding of what is happening when you interact with a computer and when you use some common technology, such as cloud storage services and social media services.

While this book aims to cover a wide variety of topics, it will not tell you everything that you need to know for your archive. Every archive is different, and so it would be quite difficult for a single book to cover every situation. Instead, it will provide basic information that you can use to get started or to get a basic understanding of digital archiving so that you can learn more.

You do not have to be a librarian to use this book. You could, for instance, be trying to preserve a family history, or you may want to create a local genealogical collection. That's okay, too. This book will guide you, step-by-step, through all the basic information that you need, explaining computers, files and file types, and how to make your collection accessible.

You do not have to read this book in order. If a chapter is irrelevant to you, you can skip it. For instance, if your archive has no audio media, you might want to move ahead to something more relevant to you. If something is of particular importance, you can skip ahead to that chapter. For instance, if you have questions about copyright, you can skip ahead to chapter 14. If something was covered in an earlier chapter that you would need to know to fully understand what you are reading, this will be noted for you. However, the concepts in the book build on one another, so reading in order is recommended. Even if you think a chapter isn't relevant to you, you never know what kinds of ideas and inspiration you might get from learning more about the possibilities of digital archiving.

Understanding computers is easier than it seems, and creating a digital archive doesn't need to be intimidating. Anyone is able to make a difference, helping to preserve both the past and the ever-changing digital nature of our present.



Why Use Digital Preservation?

IN THIS CHAPTER

- ▷ What is digital archiving?
- ▷ What is the difference between digital archiving and traditional archiving?
- ▷ What are some current large digital archiving projects?
- ▷ How can you start your own digital archive? What do you need?
- ▷ How should you use this book?

What is an archive? And, for that matter, what's the difference between an archive and a library? It's a question that's a little trickier to answer than you might think. It's essentially a collection of materials that are stored and maintained and are intended for others to use, but not to own or purchase. But which does that sentence describe: an archive, or a library?

As you might suspect, in the past, the concepts of an “archive” and a “library” were pretty much the same. Probably this is because books were massively expensive and rare, since they were written and illustrated by hand. There might be only one copy of a book in existence, as well, and such a book would be extremely valuable. If a person could afford to have books, then that person would also be able to afford the expenses involved with storing and maintaining those books, blurring the line between a library and an archive.

Over time, however, the cost of books has gone down and their accessibility has gone up. Libraries are now less involved with maintaining information, as an archive is, and are more interested in distributing it. Libraries even have to make decisions about which items are worth keeping in the library and which should be weeded out. There are books around today that are not valuable, or that have so little value that they can't even be *given* away; consider outdated encyclopedia sets, for instance. There are instructions on the Internet for art and craft projects that describe how to turn old books into things

such as lamps and tables, and there are people who fold or carve the pages of books in order to turn a valueless item (at least as far as the ability to sell it goes) into a work of art. This was pretty unthinkable not that long ago, which may be the reason for the need to distinguish between a library and an archive now—at least in part.

An archive is different from a library in that the point of an archive is to preserve materials from the past—in essence, to save a record for the future. Archives don't weed the same way that libraries do, and archives don't consider items to lack value, because everything created by a culture is a record of its past. Information in an archive does not become outdated.

The word “archive” probably evokes the imagery of old books and papers, yellowed and covered in dust. What an archive really stores, however, can be much, much more varied. Again, the point of an archive is to preserve the past for use in the future, so any artifact from the past that has relevance could potentially be stored in an archive.

There is a lot of information available about the past owing to the efforts of archives. For example, we have information about notable people and events from the past because it was common to write letters, and these letters reveal useful information about those people and events. Letters are one of the many things that are stored in archives.

But what about now? Most people don't write letters anymore to inform their friends and relatives about themselves and their lives. This is more likely to be done using e-mail or texts, rather than through letters, which could take days or even weeks to reach the recipient.

You would, of course, want to store the information in e-mails and texts in your archive. People who will be considered of historical relevance in the future are writing to others via e-mails, sharing their thoughts in the form of blogs and vlogs, or even social media. The invention of the Internet and the World Wide Web has been revolutionary for people to share information with one another.

E-mails and texts and other forms of online communication are not the same as letters, though. You can't keep them in a box and take them out to examine them, as they're ephemeral, intangible. You could, of course, print them out and keep them in a box if you wanted to, like a normal letter, but this is clumsy and difficult. Digital communication allows for things that have no tangible equivalent. Suppose that the e-mail has an attachment? A link to a website? An animated GIF image? How will we learn from these types of communications hundreds of years from now?

This is where the concept of *digital archiving* comes in. The idea of digital archiving can refer to a couple of different things, but basically it refers to these:

- Converting tangible materials to digital ones (like scanning a photograph), usually to either preserve the original or to make distribution easier.
- Storing materials that have no tangible form (like the aforementioned e-mails). These are known as *born-digital* materials.
- Storing materials that are somewhere in between these concepts—for example, a CD containing software *sort of* has a physical format (the disk), but it's really more digital in nature. Software is also something that can be archived.

Digital archiving has recently become prominent among librarians and archivists for a few important reasons:

- In the past, computers were too large, expensive, and slow to make digital archiving viable. Only large companies and organizations could use computers for large

amounts of data storage, and even then, the amount of data that could be stored was tiny compared to what is possible now.

- Modern computers and the equipment needed to create a digital archive, such as scanners, are affordable and easy to obtain today, even for an archive with a very small budget.
- Owing to the rise of personal computers as everyday items and the rise of the Internet, more and more information is being created that has no original physical format and is solely digital in nature. This information is quite vulnerable for reasons that will be explored later and, therefore, archiving efforts are needed in order to preserve the history being made right now.
- Also owing to the rise of personal computers and the Internet, your patrons are expecting to have the information that they desire rapidly and possibly without ever entering your archive. For example, a college student can now write a perfectly good research paper without leaving his or her dorm room using resources such as online databases, journals and magazines that post articles online, e-books, textbooks, and blogs written by professionals.

So, digital archiving is a *trend*, but it's not *trendy*, and unless everyone spontaneously decides to stop using computers and the Internet, it's something that will need to be addressed more and more in the future as more and more digital materials are created and people continually turn to the information that can be found online rather than what can be found inside a building.

At this point, though, you may be wondering why it is necessary to have an entire book devoted to digital archiving. Isn't it just like regular archiving? In some ways, yes, it is, and in some ways, it's quite different.

Archiving

People have been attempting to record information for millennia. Images, which can be a form of recorded information, have been around as a form of communication for tens of thousands of years. Images are nice in that they don't require a particular language to understand the meaning, but on the other hand, they can't be very specific about meaning, either, and so something more precise than a picture needed to be created. Sometime in 4,000 BCE, people were using small tokens to indicate numbers or amounts of goods, which were very important for a developing economy and civilization. Writing as it is understood today followed not too long after that, allowing for detailed records to be kept (Valentine, 2012).

Ever since then, writing has been a major form of sharing information in different civilizations across the world. Consider the Internet, for instance, which has an unprecedented amount of written information (and other forms of information, too, such as videos, photographs, etc.) that can be shared nearly instantly all across the globe.

As mentioned before, libraries are limited in that they are physical spaces. There are only so many materials that a library can store within a building. An archive, too, can only keep so many materials simply because there is only so much space to keep the materials.

It may be sobering to archivists, then, to think of preserving the massive amounts of written materials, images, sound recordings, and other materials that exist, since you can't save everything. You'll always be limited by the amount of space and time that you have,

your funding and other resources, and the number of people you have to do the work of preservation.

This limitation changes somewhat with digital archiving, though. There's a reason people prefer things like e-mail and digital photography over regular mail (snail mail) and printed photos. You can save *huge* amounts of information using the technology available today, but physically, it will take up practically no space at all.

As an example of just how much the technology of today can store in comparison to that of the recent past, the floppy disk was the main method of storing data outside a computer not very long ago (it's since fallen largely, but not entirely, into disuse). A typical floppy disk could store about 1.44 megabytes of data (you'll learn more about what exactly this means in the next chapter). By contrast, many modern storage devices can hold *terabytes* of information, a not uncommon amount for current computer hard drives. A terabyte is 1,048,576 megabytes of information.

So, suppose that you were able to store one high-quality image on that floppy disk. A modern external hard drive could hold about 728,178 images of the same quality—more than most people will ever need. But for you, the need to store that many images could well be within the realm of possibility. Remember, this is part of why digital archiving is becoming much more prevalent: Storing 728,178 floppy disks would be exceptionally inconvenient, not to mention that it would be expensive to buy that many floppy disks. But buying and storing one hard drive that stores a terabyte of data is not difficult at all.

Think of things this way. You could have a paper version of a book, which will take up a couple of inches of shelf space. You could also have a CD with the same information on it, but in a digital form. CDs are a fraction of an inch in thickness, and the scenario of a CD only holding one book is an inaccurate situation—a CD could hold quite a few books, depending on how long they are, what kinds of files they are, and whether or not they are illustrated. The digital copies of items take up far less space than the tangible ones.

What this means to you is that going digital brings you closer to the ultimate dream of an archivist: saving everything. But before you get too excited, you need to remember a few things. Digital archiving is not the same as traditional archiving. If you're trying to turn a physical collection into a digital one, then this is a very time-consuming process, even if you have the best and most optimal equipment for your particular collection. It's also rather dull and tedious work. So, you'll be restricted both by your available staff and by that ever-limiting factor: time.

If you are preserving materials that are born digital (have no original physical format), things can get complicated. You'll be working with digital files, which need appropriate software programs and equipment to use. This will require making some important decisions about what formats you'll want, whether converting files to another format is appropriate, and more. Some file formats are easy to understand because they have an obvious physical equivalent (a JPEG file is a picture, like a photograph, an EPUB file is an e-book or an electronic book). Some are not (for example, a CSV file is a set of information that is separated by commas and can be imported into a program to upload data).

Understanding rights can be messy with digital materials, as well. Some software formats are proprietary, and some are not. The operating system required to run programs, which may be important depending upon what you archive, is typically proprietary. Digital materials may be under copyright law, and born-digital materials may have their own rules regarding copyright, a topic that will be explored later in this book.

There's another issue, as well. Archivists traditionally work at preserving physical items. Though writing has been around for thousands of years, no one's created a truly stable method of saving information, though some people and organizations have made an excellent effort. There's still a struggle to prevent written materials from decomposing. But no matter what kind of material you use to record information, it is all subject to decay. The various writing materials of the past—such as papyrus, bark, paper, and parchment—are all subject to decay or damage caused by the passage of time. Even materials that one would think are nearly impossible to damage or would withstand the test of time are deceptively delicate. Clay tablets can be eaten by worms, strangely enough. Stone can break, crumble, or simply wear away. Metal is subject to corrosion (Kathalia, 1973).

Electronic information is *no more stable* than any of these. In fact, it's even *less* stable. A book printed on acidic paper (the bane of archivists everywhere) has a longer projected lifespan than stored digital materials (Lazinger, 2001). And, as mentioned before, digital and physical archiving are *not* the same.

Consider this: archivists know a lot about how to keep a book from falling apart. All the typical ailments of a decaying book have treatments: rot can be stopped, for instance, and pages and covers can be glued back together. If you have a book with acidic pages, the pH can be altered with chemicals. There are quite a variety of materials and methods that can be used to preserve all kinds of information with physical formats. After all, archivists have had thousands of years to perfect the art, and there's been a continuous production of new material during that time available for experimentation. You can easily determine that something is wrong with an item like a book from a visual examination, and address it accordingly. This is not true for digital materials.

The bits and bytes of a digital item are not as easy to deal with as tangible items. You can't touch them. You can't stitch a corrupted file back together or add a chemical to prevent a file from decaying. While computers are a part of everyday life for many people and become more user friendly all the time, understanding exactly what they are and how they work is not so simplistic. Even an expert might struggle to understand what is wrong with a file that has gone bad.

It's not as easy to determine what is wrong with a storage device for digital items as it is with tangible ones, nor is it so simple for you to fix it yourself. Part of the challenge of dealing with digital materials is determining how to preserve information that doesn't exactly have a physical format, information that doesn't depend upon the preservation of a specific object, or even information that has *never* had a physical format.

The difficulties with digital archiving will be discussed somewhat throughout this book and will be explored in-depth in chapter 15. But for now, consider the positive aspects of digital archiving. It's an excellent space saver, and your collection will likely see more use if you digitize it, especially if you make it available over the Internet. Even if you don't and you make your digital collection solely available inside your archive, your patrons will very likely appreciate a digital collection, particularly if you work to make it easily searchable, a topic that will be touched upon in chapter 13. Searchability greatly improves the speed with which patrons will be able to locate meaningful information in your collection.

Even if you don't want to make your collection available to the public at all, a digital archive can be useful and of great benefit to future generations. It can serve as a backup to your physical collection, for instance, or, you may simply want to contribute to the overall goal of preserving a culture that is becoming owners of bits and bytes or usage

rights rather than physical objects. There are many benefits to wanting digital collection for your archive, and many ways in which it will benefit others.

Goals of a Digital Archive

While there are many organizations creating digital archives right now, the ultimate purpose for each program may not be quite the same. Therefore, you may need to approach things a bit differently based upon your collection, the goals of your archive, and, to some extent, your philosophies in regard to what an archive should achieve.

It may be that you just want to make your collection more available to the public, which is a simple goal with a clear and useful purpose for your archive. For example, suppose that your archive has a lot of genealogical records on microfilm. Microfilm is really handy and is a great space saver, but it's tedious to use and to both locate and put back into place. You could scan the images on your microfilm reels and create digital images instead, which don't have to be filed away and can be retrieved with a keyword. If your collection is available online, you might find that your patrons are not only increasing in number, but are coming from places far away from your archive. This particular scenario might mean fewer patrons physically coming in through your doors, so you'll need to think about how to show that your archive is getting plenty of use when it comes to getting funding. Though this is a common scenario, there are many other goals that you could have.

Are you interested in preserving the vast amount of information that is available through the Internet? In the past, to share information, you'd need to print a book, flyer, newspaper, newsletter, or similar item that was limited in number and could be kept as part of an archive. Now, many websites make publishing all kinds of materials online easy and accessible—and not just written materials, either. YouTube, for instance, makes it possible for anyone to share videos of nearly anything, while services like Instagram make it possible to share photos. Anyone can make something and share it with the world through the Internet. Is it the goal of your archive to preserve this kind of information?

Information disappears from the Internet without a trace on a daily basis for reasons that will be explored in several chapters throughout this book. If a website is removed by the owner and there is no backup, it has effectively vanished from the world. Errors can cause information to disappear, too. For example, the once very popular social media site MySpace allowed its users to upload their own music, a way for musicians and bands to share their work. In 2019, as part of a server migration issue, files for 50 million tracks were permanently lost (Van Sant, 2019). Due to the efforts of the Internet Archive organization, 450,000 of these songs were archived and still exist (Howard, 2019). Any artist who was not part of this small archiving project and relied on the MySpace service to store their songs, however, is simply out of luck. The efforts of archiving organizations can help prevent these types of cultural losses that can arise from the precarious nature of web-based storage and services and from problems that modern technology can face.

More information is being produced every day than can be meaningfully sorted and organized by librarians, which makes attempts at archiving it a challenge. The difficulty is made even greater by the fact that some people steal materials from others online, effectively creating illegal duplicates of the same information, or the fact that people can easily change, move, or erase information that they have created. Do you want your archive to be part of an effort to organize, track, and record this kind of information?

As an example of a similar possible goal, if your archive is part of a university, you might want to collaborate with the IT department to archive web pages created by the faculty or to organize and create backups of papers written by professors. If something ever goes wrong with your university's servers or faculty members stop maintaining their pages or move on to another university, you have the information that they created and can put it back online if desired. You've probably encountered a "dead link" before, or a link that used to go to a web page but the page is no longer there. Your archive could address issues like this by preserving web pages.

Another problem with digital information is the fact that it's so easily changed. It's impossible to accidentally *erase* a book out of existence, but it's quite possible to destroy an e-book in this manner in just a few seconds, leaving nothing of the original. It's also impossible to go in and erase a paragraph out of a book (even the most meticulously placed Wite-Out will leave a noticeable gap on the page), but this is perfectly possible with an electronic document, leaving no one the wiser. Are you interested in ensuring that born-digital materials have a backup to protect them against accidental loss, or that they have a master copy to which they can be compared for integrity? The second goal is one that's actually being addressed right now, and will be discussed further in chapter 15.

Though being changed or erased will obviously make information unavailable, there's another scenario in which data can be lost: obsolescence. This is one of the most difficult issues in regard to digital archiving and will be addressed throughout the book—particularly in chapter 15. You're probably aware of some computer technologies that are considered obsolete, such as 8-inch floppy disks, 5.25-inch floppy disks, punch cards, and some forms of magnetic tape.

Your archive may be interested in ensuring that data does not become obsolete due to file formats or equipment that is no longer practical to use. This would mean focusing on transferring information from old methods of data storage (like a floppy disk) to ones that are more current or converting old file formats to new ones. This is a more abstract goal, in a way, because your archive doesn't necessarily gain anything or achieve better patron access by moving files around. However, it's a very important goal. Much information has been lost during the rise of computer usage as a method of information storage due to obsolete methods of data storage or by not properly caring for methods of data storage.

As a rather infamous example of this happening, the 1960 census was recorded onto a UNIVAC tape drive, which was probably efficient at the time; computers are excellent when it comes to efficiency. However, when the U.S. Census Bureau attempted to access the stored files a mere sixteen years later, it was discovered that they could not be accessed due to the fact that the UNIVAC tape drive was now obsolete. While the Bureau was able to transfer most records onto new tapes, not everything could be recovered (Lazinger, 2001).

POSSIBLE GOALS FOR A DIGITAL ARCHIVE

- Making it easier to access the archive's information
- Preserving modern digital culture and information
- Assisting other archives with their preservation projects
- Protecting a collection that is unique to your organization
- Making information available while protecting the original object from wear caused by use (like digitizing a book)

If your archive decides to focus on this as a goal, then it's certainly a useful and worthwhile one. Regardless of what you do, though, guarding against obsolescence is something that you'll need to address in order to protect your collection, and so while this might not be the major focus of your project, it is something that will need to be a part of your plans.

Digital Archives

If you create a digital archive, you're certainly not alone in your project. There are a variety of organizations that are attempting to preserve all kinds of materials. When you begin planning your archive, you can look to these organizations for guidance by seeing what they preserve and how, or even asking questions. There are many digital archiving projects, and this book will touch on just a few of the ones that may be of particular interest to you.

If you haven't heard of it already, American Memory, which is part of a larger project known as the National Digital Library, is something that is sure to interest you. The aim of this program is to digitize the Library of Congress's historical collections and make them available to the public. While the program originally distributed their digitized materials via CD-ROM and gave copies to schools and libraries, the cost of this method became prohibitive. Today, the collection is available to the public via the Internet, which is much less costly and much more convenient to patrons. The collection holds a wide variety of materials, such as sound recordings, photographs, maps, sheet music, videos, and writings (Library of Congress, n.d.).

Apart from having an interesting online collection, the Library of Congress is a good place to turn to in general for ideas about how to preserve your information. The Library of Congress has many recommendations online for good file formats to use for preserving digital materials. If you don't quite understand what that means or implies, this will be covered later in this book.

Another interesting example of a large-scale project is the Internet Archive. It is exactly what it sounds like: an archive of the Internet. While not fully complete (such a thing would be nearly impossible), the project records and archives many online web pages, something that may prove to be useful to historians in the future—or possibly even now. However, they don't just store web pages. The project also stores many other materials, including digitized text, audio, and video files (Internet Archive, n.d.).

The Internet Archive promotes itself as an "Internet Library," and in a way it is, but there are a growing number of these today, with many projects attempting to fulfill the same ideal. For instance, Project Gutenberg is yet another digital archiving project. This project's main aim is to digitize classic literature and make it available to the public. Literary works that are no longer under copyright in the United States are digitized and made freely available to U.S. citizens (Project Gutenberg, n.d.). Works like Frank L. Baum's *The Wonderful Wizard of Oz* and Louisa May Alcott's *Little Women* are among the numerous offerings available through this project.

All kinds of things can be archived, so if you have an unusual collection or a very specific goal, this should not deter you from making a digital archive. For example, the National Digital Newspaper Program, which is a program that is part of a partnership between the National Endowment for the Humanities and the Library of Congress, has an extremely specific goal: provide Internet access to historical newspapers from all across the United States as well as information about those newspapers. A variety of

organizations participate in building the collection by digitizing their archival newspaper collections and contributing both the files and information about the newspapers to the program (The National Digital Newspaper Program, 2019). This is something you may want to consider, as well, while you read this book: if you have similar goals to another archive, archives, or an organization, you may want to collaborate, or you may want to consider participating in a larger program.

Starting a Digital Archive

You may be starting completely from scratch with your digital archive. That's okay; this book assumes that you don't have a program or that something about your program could use an overhaul. But you may be wondering if you really have the things that you need to get started; for instance, you may wonder if your budget is capable of handling this, if you have enough people to devote to the project, or if it will take too much time away from your other projects.

There's a lot of confusion when it comes to digital archiving. The standards that you should follow aren't particularly clear. There are multiple groups out there with suggestions for how you should go about things, such as which computers you need to use and at what resolution you need to scan items. Some of what these groups have to say is very confusing. Part of the problem, too, is trying to make standards for creating digital information when the software you need and the capabilities of computers keep changing every year.

Looking at what other people are doing and investigating standards for data creation is a good thing. However, the truth of the matter is this: You can make a digital archive with a fleet of brand-new computers, hoards of cameras and scanners, the latest software, and magnetic tape recorder with a state-of-the-art robotic tape-retrieval system. You can also make a perfectly useful archive with an older computer, a scanner, some software that you downloaded free off the Internet, and an external hard drive that you got from an office supply store. Everything depends upon what you have to work with and what you need to get done. When you read this book, keep the following issues in mind regarding your archive:

Budget. What kind of a budget do you have to work with? Do you have a lot of money, or will you need to make do with what resources the archive already has available? Do you have a donation that will give you a one-time opportunity for equipment that you may not be able to replace soon?

Timeframe. Are you on a time limit, or do you have an indefinite amount of time to add to your collection?

Scope. What kinds of things do you want to archive? Books, music, images, web pages, software programs? Are you preserving the past or the present?

Staff. How many people are available to work on your project? Do you have armies of student workers at your beck and call, or will it only be you working on this project?

Collaborating. Do you want to collaborate with another archive or contribute to an already-existing project?

There is no answer to the above questions that will indicate that you don't have the means or ability to create a useful digital archive unless it is that you do not and cannot have computers and you don't have any time to spare for someone to work on the project. Those are the two things you *must* have: a computer and a worker (and the worker may

be you). Thinking about these questions as you read, however, will help you make some decisions or determine if a proposed method of data storage is not for you.

A single, untrained person can make a usable archive. There are instructions online designed to assist people doing personal genealogy so that they can create a digital archive, and if you are, in fact, a genealogist making a personal archive or if you are working at a single-librarian archive or library, this book can help you, too. While it's obviously optimal to use the best equipment and resources available and to use the standards indicated by groups researching the topic, sometimes that's not possible. Your ultimate goals should be to create a digital archive that suits the needs of your organization and your patrons. If it meets your needs and is accessible, then you've reached your goals.

Key Points

- Digital archiving is a relatively new trend among archivists and librarians that will become more prevalent and more important as time goes on and society relies more and more on information that is solely digital in nature.
- Digital archiving offers many benefits to you, your archive, and your patrons, although there are drawbacks as well.
- Though the goal of preserving the past and present for the future is the same with both digital and traditional archiving, how you should approach achieving your goal is different.
- Digital archiving can have a variety of different goals, and a single project may have multiple goals.
- The scope of a project may be limited by your resources, but it's possible for nearly any archive to create a useful digital collection.

In the next chapter, you'll begin to learn about computers—including an overview of what they are and how they work in general—as well as learn some of the common terminology in regard to computers. You'll also learn about many of the physical components of a computer and what features you can look for in order to purchase the optimal machines for your archive.

References

- Howard, Tanner. 2019. "Please, My Digital Archive. It's Very Sick." *Lapham's Quarterly*. <https://www.laphamsquarterly.org/roundtable/please-my-digital-archive-its-very-sick>.
- Internet Archive. n.d. "About the Internet Archive." Accessed March 31, 2019. <http://archive.org/about/>.
- Kathpalia, Yash Pal. 1973. *Conservation and Restoration of Archive Materials*. Paris: Unesco.
- Lazinger, Susan S. 2001. *Digital Preservation and Metadata: History, Theory, Practice*. Englewood, CO: Libraries Unlimited.
- Library of Congress. n.d. "Mission and History." American Memory. Accessed March 31, 2019. <http://memory.loc.gov/ammem/about/index.html>.
- The National Digital Newspaper Program. 2019. "About the Program." <http://www.loc.gov/ndnp/about.html>.
- Project Gutenberg. n.d. Accessed March 31, 2019. <http://www.gutenberg.org>.
- Van Sant, Shannon. 2019. "MySpace Says It Lost Years Of User-Uploaded Music." NPR. <https://www.npr.org/2019/03/18/704458168/myspace-says-it-lost-years-of-user-uploaded-music>.



How Do Computers Store Information?

IN THIS CHAPTER

- ▷ What is a computer?
- ▷ What is binary, and why is it important in computing?
- ▷ What are the basic units of data storage?
- ▷ What are the basic parts of a computer?
- ▷ How do you choose the best computers for your archive?

In modern times, a lot of people start their day by waking up next to a computer. They go about their day with a computer (often using more than one computer), and then spend their evenings with a computer, as well. In fact, for a lot of people, *not* doing this is such a novelty that there are a plethora of online articles written by people documenting their experiences with “going off the grid,” or trying to live without the convenience of apps and social media.

Of course, this is generally a reference to smartphones, which are essentially very small, very portable computers. The invention of the smartphone has really changed the relationship between users and computers, making the computer not an item merely for work, but an item that is often essential for everyday life.

Computers are indeed pretty ubiquitous these days. They’re even in devices that you might not recognize as being a computer at all; for example, a microwave contains computer-like components. Some modern coffeemakers have computer components. Most modern sewing machines are at least partially computerized.

While some people take little note of these tiny computers, others use computers to control many things in their everyday lives. For instance, some people have decided to have “smart homes,” in which aspects of their environment have computer components and can be controlled with software, such as the heating and air or the lighting.

As computers become tinier and more omnipresent, the difficulty level in operating one of these devices is diminishing, as well. While he or she may not really understand what they are doing, even a toddler can master the “swipe and tap” interaction of a smartphone or tablet, and many parents use this technology to entertain their children because it is so easy to use.

There really is no need to understand how a computer works to use one. However, if you are going to be working with materials that are digital in nature you may find that a deeper understanding of the operation of a computer makes it easier when you need to make decisions or when you run into problems, and so this chapter will discuss some of the basics of computer operation.

Many books on the basics of computers start with a history of computers, oftentimes beginning with a discussion of the abacus. This may seem like an odd choice, given that the abacus really doesn't resemble a computer at all. Certainly you can't text anyone with an abacus, and Google is not abacus-compatible. So, if an abacus is so unlike a modern computer, then why do so many books include a mention of them just before discussing devices that *do* seem much more like a modern computer?

It's better not to think about how an abacus is like a computer, but rather how a computer is like an abacus. An abacus is a set of beads on posts that can be moved up and down to help the user make calculations quickly. It is a calculator. Likewise a computer is, in essence, a glorified calculator. The basic definition of a computer is that it is a device that can compute, or make calculations. The word can even refer to a *person* who is making calculations, “one who computes,” which means that it's the calculations that are the essence of what a computer is.

You may be thinking that you don't really use a computer to make calculations. You use it to check your emails, or listen to music, or play games, or a wide variety of other things. However, from the *computer's* perspective, everything is numbers. The characters in the text of your email are numbers, and the address it goes to is another string of numbers. The notes of music you hear are stored as numerical values, and the colors in the images in a game are also stored as numbers. Every function that a computer does is essentially a calculation—it's just that the user doesn't recognize it that way because the results of the calculation do not look like what we think of as calculations.

Again, while a typical archivist doesn't need a computer to make complex calculations, knowing a little about how a computer works and how it interprets data is helpful to understanding what digital information actually is and how it can be kept safe. This chapter will cover some of the basics about how a computer works, starting with a brief discussion about how computers view data, and will then explore the components of basic computers. When you need to purchase computers, either for yourself or for your patrons, there's a lot of jargon that gets thrown around, most of which involves non-intuitive numbers and acronyms for the hardware that you find inside a computer. Knowing what these parts are and what they do can make decision-making much simpler, and also make it easier to purchase computers that are optimal for your archiving projects.

The Binary System

So, how exactly do computers process information? While you might see an image of a bunny or a sunset on the computer's monitor, there isn't a tiny picture of a bunny or a sunset inside the machine. If you listen to a song, there isn't a miniature copy of a band

inside the computer (although that might be more interesting). As mentioned before, everything to a computer is numbers. A computer can assign a numerical value to a color, for instance, and all the numerical values for colors can combine together to make that aforementioned bunny picture.

Modern computers work in binary, which means that they only use two numbers—ones and zeroes—to make all their calculations. In movies and TV shows and other media, artists and directors might use screens of ones and zeroes to symbolically represent “incomprehensible computer technology stuff.” But that is, of course, not what the inside of a computer looks like.

From a computer’s perspective, there aren’t even ones and zeroes. The calculations inside a computer are actually done through high and low frequencies of electricity. A high frequency represents a one, or “on,” and a low frequency is a zero, or “off.” Think back to the abacus example. The beads symbolically represent different numbers or multiples of numbers. So, in a computer, pulses of electricity also symbolically represent numbers, but there are only two possibilities: one and zero.

A computer doesn’t “know” what a one or a zero is any more than an abacus does, either. A computer is not capable of understanding these concepts; all a computer can do is follow commands programmed by a human, who does know what the concepts of one and zero are. What a computer *can* do is detect high and low frequencies, which humans interpret as ones and zeroes. A modern computer only seems more complex than an abacus because it can do so many calculations in a short period of time, which it does through the use of the *binary* system. Binary means two, referring to the two possible numbers in the system, one and zero.

But a computer doesn’t have to use binary at all. It’s possible to create a computer that uses the much more familiar decimal system, or an octal system, or a hexadecimal system, or essentially any system at all. In fact, the people who designed early models of computers attempted to have values of 0–9, and even to have letters represented in transistors. It was only in 1940 that the current binary model was even proposed (Andrews, 2006).

People tend to prefer the decimal system, which uses the numbers 0–9. So, why would you want to program a computer to use a system that’s not intuitive to human users? The reason is that the binary system is easier for *computers* to handle. Remember, the ones and zeroes for a computer are pulses of electricity. From a computer’s perspective, this makes things very easy to interpret. Anything with a high frequency is a one, and anything with a low frequency is a zero.

If a computer engineer used a number system with eight numbers instead, for example, which is an octal system, this would require the computer to be able to detect a wide variety of signals. It would need to be able to detect whether an electrical signal is low, a little higher, a little higher than that, and so on for eight different frequencies. That leaves a lot of room for inaccuracy and interpretation on the part of the computer. As mentioned before, computers don’t “know” anything, and so they can’t make guesses the way that a human can about information, either. For instance, a human can make a good guess about the true meaning of a misspelled word, whereas all a computer can do is consult a program and provide some likely matches (a process that leads to the often comical or frustrating results of autocorrect programs on tablets and smartphones).

A binary system doesn’t need guessing. It reduces things to essentially yes or no. Is the frequency high, yes or no? Is it on or off? There is no “sort of on” for a computer to interpret. So, even though it’s not intuitive for humans, using the binary system greatly

improves the accuracy of the computer and leaves little gray area for interpretation. It also helps when the electrical signal needs to be strengthened. Even over an area as short as an inch, the electrical signals can begin to degrade, losing the distinction between the high and low frequency. This means that, as the signals travel around the computer, they can lose their strength. If there are only two possibilities, this signal can be “reclocked,” or reset to the original frequencies, thus eliminating data loss. Doing this to a signal that has multiple interpretations is difficult or even impossible without data loss, whereas strengthening this “yes/no” signal is much easier (Dale and Lewis, 2013).

Binary can represent and calculate any number at all with just two numbers. The set of ones and zeroes 110001 is equal to 49 in the familiar decimal system. The two numbers are really the same thing and mean exactly the same amount, it’s just that the system of *representing* it has changed. To show you how this works, table 2.1 has the decimal numbers 1–10 and their binary equivalents.

Table 2.1. Decimal and Equivalent Binary Numbers

BINARY	DECIMAL
1	1
10	2
11	3
100	4
101	5
110	6
111	7
1000	8
1001	9
1010	10

In school, when you first started learning about math and numbers, you almost certainly learned about “places,” or the “slots” used to represent and understand numbers. Thinking back on this can help you with understanding how binary works.

Suppose that you have 2,156 items. You know, intuitively, that this number does not mean that you have a set of 2 items, a set of 1 item, a set of 5 items, and a set of 6 items for a total of 14 items. You can see this number as meaning that there are two thousand, one hundred and fifty-six items present. That’s because you know that each of the places for each number actually represents a *multiple* of a number, not the number itself. For example, the number 2156 is broken down as follows:

- 2 Thousands
- 1 Hundreds
- 5 Tens
- 6 Ones

When in the tens place, five doesn't mean five. It means five multiplied by ten, or fifty. The one in this example is not one item—it's one set of one hundred items. Binary can be thought of in exactly the same way, but it doesn't have the same places. For instance, the number 1011 has these places:

1 Eights
0 Fours
1 Twos
1 Ones

This number isn't 3, nor is it 1,011. It has one set of ones, a set of twos, no sets of fours, and one set of eights for a total, in the decimal system, of 11. While the places for decimal numbers keep going up in multiples of ten, for binary, each new place is double the place before it. So in this sequence, the next place beyond the eights place would be the sixteens place.

You can produce any number with the binary system using any type of mathematical function, though it gets a little more complex when a computer starts working with decimal places and negative numbers, and all computers are limited in the highest number that they are able to calculate (Dale and Lewis, 2013).

Binary can only represent two numbers at a time, and just like in the decimal system, it's necessary to have multiple numbers grouped together to express a larger amount. For instance, a person needs the numbers 1 and 9 in order to express the number 19. It's like that in binary, too, and so there are terms in computer design for these larger, more useful collections of binary numbers.

Bits and Bytes

Computers work in binary. So, the smallest unit of information possible for a computer is either a one or a zero. The term for this unit is a *bit*, which stands for the longer term “binary digit.” Think of a bit being like a coin—the coin itself is a single unit with two possibilities, heads or tails, and a bit also has two possibilities, one or zero. A single number doesn't really do much, though, so there are terms for larger groups of bits.

USEFUL TERMS FOR STORAGE CAPACITY

Bit	One binary unit
Byte	Eight bits
Kilobyte	One thousand bytes
Megabyte	One million bytes (1 thousand kilobytes)
Gigabyte	One billion bytes (1 thousand megabytes)
Terabyte	One trillion bytes (1 thousand gigabytes)
Petabyte	One quadrillion bytes (1 thousand terabytes)
Exabyte	One quintillion bytes (1 thousand petabytes)

One of the most commonly used and useful of these is the *byte*, a set of eight bits that together form a single unit of information. Why eight? It may seem somewhat random, considering that, again, people like to think in tens (probably due to the fact that humans come with an easy method of keeping count, ten fingers). However, eight bits is a useful unit of information for a computer. Eight bits allow for enough numerical combinations to use one binary number to represent every letter of the alphabet in English, including both capital and lowercase letters, some characters, and numbers. Eight bits is also handy for encoding color information, as you'll learn in the following chapter. Because eight provides a significant number of useful number combinations, eight and multiples of eight are used a lot with computer data.

However, a byte is still really a very small unit of information. Talking about how many bytes a modern computer is able to store would be cumbersome, because the number would become very large. That's why there is usually a prefix to the term "byte," like "kilobyte" or "megabyte." The term "kilo" means "one thousand," so there are about 1,000 bytes in a kilobyte; specifically, there are 1,024 bytes in this amount of data. The exact numbers of bytes are rounded for the sake of convenience for both programmers and consumers. When purchasing storage devices, the description of how much it can store on the product packaging and what you actually get will be close, but not exactly the same. A megabyte, by the way, is about a million bytes, a gigabyte is about a billion bytes, and a terabyte is about a trillion bytes.

For a little comparison between these amounts, this chapter has a little over 8,000 words and uses about 113 kilobytes to store in a digital format. It would take about 8,849 copies of this chapter to fill a single gigabyte flash drive, which is considered tiny by today's standards.

When digitizing collections, it can be helpful to know roughly how much storage space you will have available and how much space, on average, the items you are storing will require. This will be useful for determining practical methods of storage as well as what standards will be useful to you; for example, a larger digital photo offers more detail, but also requires more data than a smaller photo.

As mentioned in the beginning of this chapter, bits and bytes don't float around inside a computer. They are composed of pulses of electricity, and this electricity needs actual, physical materials to conduct the signals from one area of a computer to another. All computers need the same basic parts in order to do this.

A Basic Computer

Every personal computer, no matter what kind or which company made it, has three basic components: hardware, software, and firmware. Just about anything that makes up a computer, short of things like the casing and screws that hold it together, falls into one of these three categories. Though these terms sound similar, there is a definite difference between them.

Hardware is composed of the parts of a computer that theoretically can be touched; that is, they're usually tangible in nature. Motherboards, hard drives, and RAM chips are parts of the hardware of a computer. These are the parts that allow the all-important electrical currents to run throughout the device. However, these items all do nothing that is of any use without software.

Software refers to the programs that run on a computer and make the hardware components actually do something. The operating system, word processing programs, and Solitaire games are all examples of software.

Firmware is a little trickier to define, because it's a mixture of both of these things. Firmware items are physical in nature, like hardware, but have software permanently embedded into the item. ROM chips are a type of firmware. However, firmware items are not usually something that concerns an average user and are more important to people who program and design computers.

Of these three components of a basic computer, two categories are visible and tangible, the hardware and the firmware. The hardware, or the “guts” of the vast majority of personal computers, though, is hidden inside a little metal and plastic box, so many people may have no idea of what is actually in there, what it looks like, or how it works. Even if you were to see inside, through a clear casing, looking at the parts wouldn't lend much information about how the computer works just by observation. There aren't many moving parts, and those few moving parts are usually hidden inside their own casings to protect them.

For all the mystery and complexity involved, the insides of a computer are comprised of a rather limited number of items. If you've never seen the inside of a computer, opening one up can be very educational. While all personal computers are made of essentially the same parts, it's best to look at a desktop-style computer, as more compact models, like laptops, require a lot of disassembling to see the parts and reassembling them correctly can be difficult (taking apart a tablet or smartphone is certainly not recommended). A typical desktop is easy to open and to look at the inside in most instances by simply removing a panel, and you don't need to take it apart to see most of the components. If you do want to look at the inside of a real computer, back up your files, take precautions against the hazards of static electricity for the safety of the computer, and be sure that the power is off and disconnected for your own safety. Never attempt to disassemble the power supply of a computer, either; this is a box behind the power socket that connects your computer to an electrical socket. Otherwise, opening a computer is typically safe to do and, in fact, can be a good thing, as it gives you an opportunity to clean out the dust that collects inside. Consult your computer's help manual for more instructions about how to safely clean out a computer to avoid harming the computer or yourself. If you're nervous, though, try working on an outdated or “dead” computer.

TYPICAL COMPUTER COMPONENTS

Central Processing Unit (CPU)

Motherboard

Random Access Memory (RAM)

Read-Only Memory (ROM)

Hard Drive/Solid-State Drive

Ports

The Central Processing Unit

One of the most important items in a computer is the Central Processing Unit, or CPU. If you open a computer, this item may not be visible to you, however. This little device is typically hidden behind a fan, as the CPU gets very hot and needs to be cooled.

The CPU does all the calculations, interprets all instructions for the computer, and coordinates input/output operations. When a user taps a key or clicks a mouse, the action gets processed through the CPU, which can then determine what needs to be done. For instance, if a user hits “A” on the keyboard with a word processing program open, the keyboard sends a signal that gets processed through the CPU, which ultimately determines that what it is supposed to do is to display “A” on the computer monitor.

There is another component similar to the CPU: a GPU, or graphical processing unit. When looking at a computer’s specifications, a dedicated GPU is more powerful than an integrated one. Integrated means that the GPU is part of the CPU and uses some of its processing power; a dedicated GPU is a bit like having another CPU that is solely for processing graphics, improving graphical speed and quality. A video card contains a dedicated GPU along with its own separate memory, circuit board, heat sink, etc.

Everything that a computer does, all the calculations needed to do anything, is processed through the CPU. As a result of handling all those electrical pulses, the CPU heats up rapidly. There are a few ways to control this heat inside a computer, and if you look inside a desktop or a laptop computer, you may see a fan, which dissipates the heat, or a block of metal strips or prongs. The second item is a heat sink, which also helps to dissipate heat. Items like tablets or smartphones do not have the space for things like heat sinks, and so use other means of controlling heat; this can come at the cost of processing speed, as faster processing makes a computer heat up more rapidly.

The fan and the heat sink help the CPU operate safely, but there are a few other components of the computer that affect CPU speed or the operation of the CPU.

Heat is actually a major consideration when designing computers, and the ability to dissipate heat may be a limit on a computer’s abilities. For example, mobile phones could potentially be more powerful than they are using the current technology that exists, but dissipating the heat that would be generated by the phone’s components is a problem. One of the reasons that desktop computers can be more powerful than things like smartphones and tablets is because there is not only more room for more and larger components, but there are more options for dissipating the heat created by these components. For example, some systems use water to transfer heat away from the components, a liquid cooling system (Gayde, 2020).

Clock

How fast a computer is able to process information is highly dependent on the speed of the CPU, and there are several aspects of the CPU that affect the speed. One of these is the clock speed. The *clock* on a computer is not quite like the one that a person checks to see when it’s time to go to lunch or when a movie will start. A computer’s clock is actually a crystal that vibrates, creating a series of pulses at an extremely high frequency. The pulses are carried to all the components of the computer, enabling a computer to synchronize its own activities, sending information and stopping at the same intervals. It’s a little like

how a traffic light directs the flow of traffic. The CPU can't have two "cars" entering at the same time. The clock ensures that all the data is transmitted in an orderly way.

The rate at which this crystal vibrates is represented in Hertz, abbreviated to Hz. One megahertz, or MHz, is a million cycles of this crystal per second. A gigahertz, or GHz, is a billion hertz. The higher this number, the faster the computer can function. This number is also sometimes represented in MT/s, or megatransfers per second, which is one million bytes transferred per second over a bus (Andrews, 2006).

The term "overclocking" refers to the practice of making a CPU run faster than its intended clock speed. This can really boost the performance of a computer, but can also generate a lot of heat and potentially damage the computer if done incorrectly. Likewise, "underclocking" is running a CPU at lower than the intended clock speed, which reduces performance, but also keeps the computer cooler and requires it to use less power, which can be helpful in some situations.

Bus

CPUs, along with working at a certain speed that is regulated by the clock, can transfer certain amounts of information each time the crystal vibrates. For instance, a computer may have a 32-bit or 64-bit processor. This is the number of bits that can be transferred to and from the CPU simultaneously; you'll notice that these are both multiples of 8, the number of bits in a byte. The electronic lines that transfer these bits to different parts of the computer are known as "buses." You can almost think of them like an actual bus, driving information to a certain destination in the machine. If a processor is 32-bit, then it can carry 32 passengers (bits) all at the same time. A 64-bit processor can take 64. Older machines had smaller buses, but 32 and 64 are common sizes now.

Front Side Bus

The front side bus, or FSB, is sometimes mentioned with a computer's specifications. This is a particular type of bus, which enables the CPU to communicate with the outside world. This bus's size is represented in Hertz, just like the clock, indicating how fast it can transmit data (Dale and Lewis, 2013). The front side bus connects the front side of the processor to the rest of the computer. There are other buses used with the CPU, such as the back-side bus, which connects the processor to the internal memory cache, and the internal bus, which enables the CPU to communicate with itself (Andrews, 2006). However, the FSB is usually the one mentioned in a computer's specifications, as it has a big impact on the speed of the processor (Dale and Lewis, 2013).

Cache Memory

A CPU has its own memory that is solely for the CPU. This memory is referred to as the *cache*. This is memory that a CPU can use to temporarily store data that has been recently processed or accessed, which makes further processing faster. For example, suppose that you were reading this book and you had to go and do something else for a few minutes. It would be more efficient for you to just put the book down on your desk rather than put it away. Cache memory works on a similar principle, temporarily storing data that will be needed again soon rather than storing it elsewhere in the computer and retrieving it

again. The more cache memory that is available, the more information that can be stored in this manner.

Multiple CPUs

There are also some terms that go along with the CPU, like “dual-core” or “quad-core.” In older machines, there was a single CPU chip in the machine. However, having more than one makes the calculations required for computer function go much faster.

A speedy CPU makes processing faster and reduces “down time” that occurs when a user is waiting for a computer to finish running a program. However, a fast processor often uses more electricity and is more expensive. Look at the specifications carefully when comparing the CPU on different models; your decision may require some balance and compromise to get the optimal combination.

While the CPU does all the calculating for a computer, there are other essential parts that largely involve memory storage. The CPU is not able to communicate with these parts on its own, however. It needs something else to do this—something that allows the results of its calculations to travel throughout the machine, and a way to get the information to make those calculations in the first place. That something is the motherboard, the largest circuit board in a computer.

The Motherboard

The motherboard is very easy to spot inside a computer. It typically looks like a large green sheet of metal, usually at the bottom or the side of the computer’s casing, with a variety of circuits and small plastic pieces sticking out of it. Most or all of these pieces will have something plugged into them. Everything in the computer connects to the motherboard at some point, and the motherboard is what enables communication between different parts of the computer.

However, it’s likely that not all of the slots will have something plugged into them when you buy a new computer. Often, motherboards have room for more components than are actually installed during the manufacturing process. This can be a good thing, as it enables you to add parts to the computer and allows for some customization. You could therefore purchase an inexpensive machine within your budget, then upgrade it as funds allow. One of the easiest ways to upgrade a machine is to add more RAM, which is a type of memory.

Both laptop and desktop computers often allow for hardware upgrades; if you look at the inside of either, you may see empty spaces or slots left with this in mind. Upgrading is more limited with tablets and smartphones; it is largely assumed that someone using one of these devices is going to be using cloud storage, although some allow for the user to insert an SD card, which allows for more data storage space.

RAM and ROM

RAM chips are long, rectangular objects that fit into little plastic slots on the motherboard. RAM stands for *Random Access Memory*. This is temporary memory in the com-

puter, the contents of which vanish when the user turns off the power. Another term for this is *volatile memory*. These chips serve a very important purpose in the computer, serving as temporary storage for the CPU and containing data until the CPU can process it. Once information has been processed by the CPU, the new information can also be stored temporarily in the RAM chips. If a person types a document in a word processing program, for example, while that person works on the document, both the information needed to operate the program and the information about the document itself are stored in the RAM chips.

RAM chips store binary values using electrical charges, and these charges need constant refreshing or else they are lost, which is why these chips don't keep their memory when the power is turned off. It is possible to create non-volatile RAM chips, but writing data to this type of storage is slower, which makes non-volatile memory impractical for temporary storage. However, this may change in the future: new chips developed by researchers at the University of Lancaster combine the ability to store data with the quick speeds of the current model for RAM chips (Potoroaca, 2020).

Unlike the CPU, in which more is not necessarily better, having more RAM has a noticeably positive effect on computer function, and more RAM is typically better up to a point. If a computer that you are considering doesn't seem to have enough RAM for your needs, find out if the RAM is expandable. Computers don't always come with as much RAM as they can actually use; that is, you may find that there are empty slots for RAM chips inside your computer. You can buy more RAM chips and put them into these slots. While the computer will need to configure itself to accept new RAM and you need to be sure that your RAM chips are compatible with your computer, installing new RAM essentially involves putting an object into a slot, and is very simple. Figure 2.1 shows what a RAM chip generally looks like.

There are a variety of types of RAM, such as DRAM, SRAM, SDRAM, DDR SDRAM, and more. Each type has certain advantages, but it is important to note that you may not have an option when trying to upgrade a computer; newer types of RAM may not be compatible with older computers.

The benefit of RAM memory for the computer is that the memory is changeable. What has been stored in the RAM chips can be erased, altered, or replaced. However, memory that is not changeable is also valuable; this is the kind that is present in ROM chips.

ROM stands for *Read-Only Memory*. These chips don't lose their memory once the user turns off the computer, which is essential for the function of the computer and its

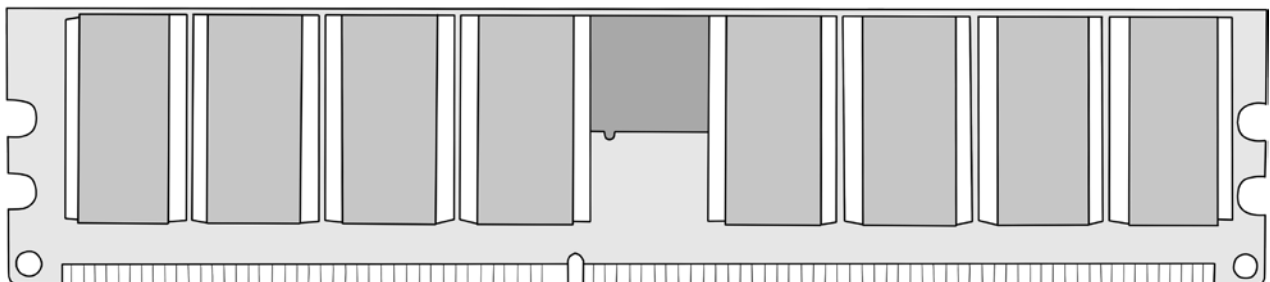


Figure 2.1. A typical RAM chip.

operating system. For instance, once you turn on a computer, the computer requires instructions about what to do next, such as how to look for the operating system software. You may have heard terms like *BIOS*, or basic input/output system, and *POST*, or power-on self-test. These are both programs that run as soon as the user turns on the computer. The BIOS is the first program that runs when the computer starts and controls communications between components of the computer, such as the keyboard. The POST is a test that the computer conducts on itself to ensure that the hardware components are functioning properly. The information for these programs is stored in a ROM chip. Without this information, a computer can't do anything at all.

While an essential part of the computer, the ROM chips are typically not an aspect that you use to compare one computer to another when making a purchase. However, there is another kind of memory that can have a great influence on your decision: the long-term memory for the computer.

Long-Term Storage

In a computer, the RAM chips have temporary, changeable information storage, and ROM chips have permanent, non-changing information storage. However, you need to have a way to save your data and retrieve it later, but also be able to change it if needed. In general, there are two options: a hard drive or solid-state storage. Both types of memory will be covered in more detail later in this book.

A hard drive is the more traditional method of storing data in a personal computer and consists of a series of platters that rotate at an incredibly high speed. Your data is stored in “tracks” on these platters, a little like a minuscule record player, and like a record player, an arm moves about the platter to find your information.

A solid-state drive is a little like the even older method of computer data storage, transistors. It is formed from a series of microscopic “gates,” which store one and zero values as electrons (a one or zero is interpreted as whether or not the gate currently holds an electron).

Solid-state storage is rapidly dropping in price and is very appealing in that it is much more rugged than a hard drive and can withstand a lot more abuse. Hard drives are pretty delicate and can easily be damaged. If you purchase tablets or other portable computers, you also do not have an option: they will use solid-state storage because you can make the storage device smaller than you can with a hard drive and this is necessary for this type of technology. However, hard drives still have their merits, and you will learn more about this later.

In both instances, more is generally better. Computers that can store terabytes of information are not uncommon at all. However, how important this is to you will depend very much on how much information you will need to store locally and whether you will use your software programs locally or “in the cloud,” a concept which will also be discussed in more depth later.

There are many parts inside a computer that enable it to communicate with itself. However, this is all fairly useless if the computer is not able to communicate with the user, as well. A typical computer has a number of ports, which enable the computer to do just that.

A typical computer, no matter what type it is, has at least one port. Whether a computer is a tiny smartphone or a large desktop, it needs to have a port, and having more than one is much more common. Ports might look like a slot, hole, hole with pins, or a raised area with small holes, and they serve an important function—enabling the inside of the computer to communicate with the outside world, and vice versa.

In the past, ports were absolutely essential to computer function, as they were necessary for attaching keyboards, printers, monitors, and more. Some computers now have integrated monitors, integrated keyboards and mice, and can communicate wirelessly for things like printing. However, a port is still helpful, even for wireless devices, as it allows direct access between devices and allows the user to view internal files (for instance, you can connect a tablet to a desktop to offload files to the desktop). Ports on smaller computers are typically also used for charging, and the same cord can be used both for charging the device and connecting it to other computers, so it is normal for such devices to have at least one port (often two, one for charging or connecting and another for headphones).

Peripherals are devices used for communication between the user and computer, with the keyboard and mouse being major methods of putting data and commands into the computer. The computer in turn can communicate with the user through the monitor and other output devices, such as printers. The type, number, and location of ports a computer has may be important to you when purchasing materials for your library or archiving project, as more ports may be more convenient.

Over the years, there have been a wide variety of possible ports for computers. Some have gone largely or entirely out of use, replaced with more convenient alternatives, but this section will describe some of the more common ports, past and present, as you may be using older computers or may even need to use older technology as part of your project. Some common ports on computers (shown in figure 2.2) are:

Universal Serial Bus (USB) ports. These are some of the most frequently found ports on modern computers and can enable a vast variety of peripherals to communicate with the computer, such as mice, printers, keyboards, cameras, and more. Storage media such as flash drives or external hard drives also plug into USB ports. There are actually several types of USB ports, with USB A and micro USB being very common at the moment, and there have been multiple versions of each type. However, this may not be important to you and your computers, as USB connectors are backwards-compatible. It should also be noted that there is another type of USB port that is becoming more common, the USB C. USB Cs, unlike most other ports, do not have a “right way up” and are both smaller and faster than previous USB types.

Thunderbolt. Like USB ports, a Thunderbolt port can connect a wide variety of devices and is intended to be “universal” in that respect. There are several versions of the Thunderbolt port, with Thunderbolt 3 being the latest. Thunderbolt 1 and 2 ports are a different shape from Thunderbolt 3 ports and are exclusive to Apple products. The Thunderbolt 3 port is compatible with the new USB C, and so cords for one of these ports can be used for the other type, although you may not get the benefits of a Thunderbolt port using a USB C cable, and vice versa.

IEEE 1394 port. This is also known as FireWire. FireWire is a device that was in competition with the USB port, and the two have similar functions and abilities. However, FireWire ports have gone out of use over time.

Video Graphics Array (VGA) port. This is a port used for monitors and connects the graphics card in the computer to the monitor; this is the part of the computer that controls the output that goes to the monitor and may also be called a video card or display adapter. This port is blue, which allows the user to match the blue end of a monitor cord to the correct port.

Ethernet ports. These ports enable computers to communicate with one another in a small location, such as solely within a library or archive, or can connect to a modem or router to enable communication via the Internet.

High-Definition Multimedia Interface (HDMI) port. This type of port enables audio and video communication between the computer and another device, such as a monitor or a flat-screen television. For current computers, this is a typical way to connect a computer to a monitor.

Audio ports. Typical computers have ports for speakers and headphones, and possibly microphones. These may or may not be important depending upon whether or not your archive includes audio or video materials.

PS/2 ports. These are ports specialized for mice and keyboards. They are round and are usually color-coordinated to the end of the mouse or keyboard so that the user doesn't accidentally put the wrong peripheral into the wrong port, since they're the same size. This type of port is more common on older computers, however, and modern mice and keyboards typically plug into a USB port or are wireless.

Serial ports. These ports connect peripherals to the computer and have either 25 or 9 pins. The 25-pin version was phased out for the more convenient 9-pin version. A wide variety of peripherals can be used with this type of port, such as mice or external modems. This is a port that might be found on an older machine (Miastkowski, 2004).

Parallel ports. These have two series of small holes. They are able to transfer 8 MB per second and can connect to a variety of devices, such as printers or scanners, but, like serial ports, are found on older machines (Chen and Mills, 2002).

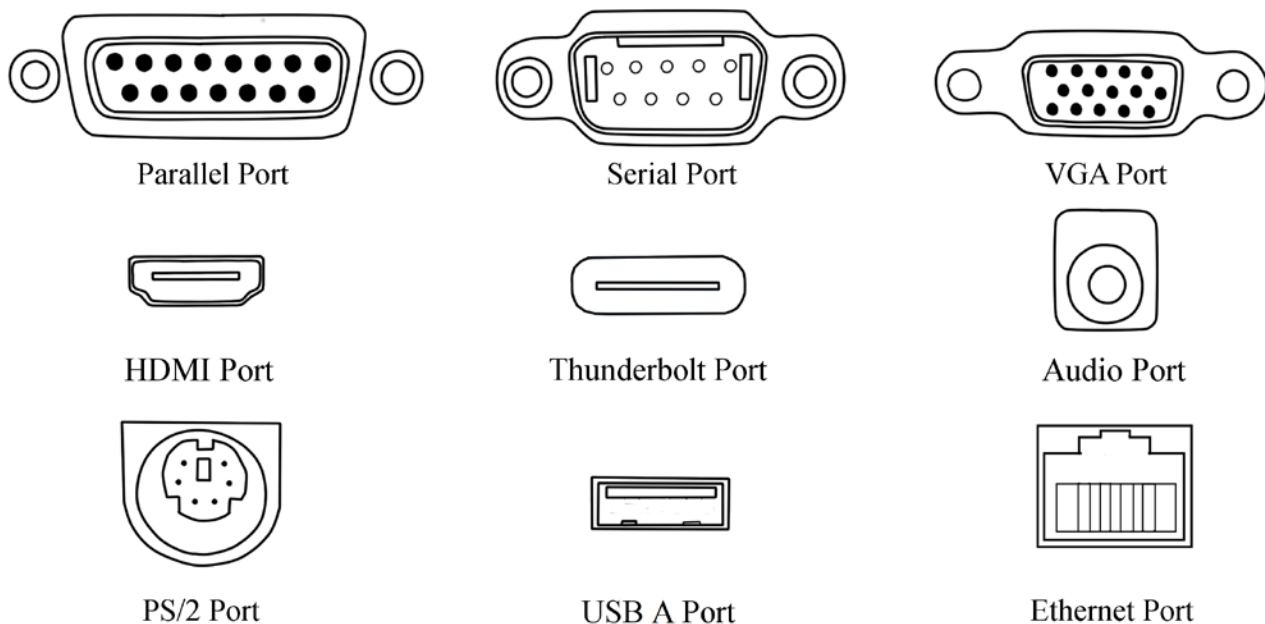


Figure 2.2. Common computer ports.

Ports often require a cable that connects the port to the peripheral. To download photos from a digital camera, for instance, you need a cable that plugs into a port. On many computers, though, there is another option. Secure Digital or SD cards are a source of memory storage for digital cameras, and some computers have a slot specifically for these cards, enabling the user to take the card out of the camera and put it into the computer so that the computer can access the photos. This eliminates the need to connect the camera itself to the computer, a process that can be appealing if the camera is difficult to move, as it might be if you have a permanent photography setup for digitization. Since ease of access is important to working efficiently, you may want to see where the ports are located on a computer, as well.

It should be noted that many peripherals are available in a wireless format. For instance, keyboards and mice are commonly available as wireless peripherals, communicating with a computer using a transceiver, which both receives signals from and sends signals to the device. Not needing wires doesn't mean that wireless peripherals don't need ports, though—the transceiver plugs into a port in the computer. Normal peripherals also draw power from the computer, but wireless peripherals require a power source external to the computer, and so require batteries or charging.

Many ports will be at the back of the computer, but some may be at the front. Take the location of these ports into consideration—USB ports on the front of a desktop computer can be convenient if you need to use them or change what is plugged into them often.

If you are using a laptop computer, the number of ports may be limited. Many new laptops are designed to be quite thin and the number and type of ports is often very limited to allow for this. There are a few options in this case; one of the easy solutions is to use a USB hub, which typically looks like a little box with several USB ports in it. The hub has a cord that plugs into a single USB port, letting you plug in more devices simultaneously.

As another consideration that might be helpful to you, it is also possible to buy converters for ports, which will allow one port to accept peripherals intended for another type of port. For example, suppose that your monitor has a cord intended for a VGA port, but your brand-new laptop only accepts HDMI cables. You could use a VGA-to-HDMI converter, plugging the cord for the monitor into one end of the converter, then plugging the converter itself into the laptop.

Key Points

- Modern computers use the binary system for calculation. Any number in the decimal system, which is what people normally use for everyday tasks, can be represented using a series of ones and zeroes in the binary system. All functions of a computer are simply a set of calculations or numerical values.
- All computers have the same basic essential parts, such as the central processing unit, buses, the motherboard, memory chips, hard drives or solid-state drives, ports, and peripherals for communication.
- Choosing the optimal computer for your archiving project depends upon a number of factors, such as what type of equipment you plan to use, what kind of software you want to run, and how much money you have in your budget.

Deciding what kind of computer is best suited for your archiving project is only the beginning of the decisions that need to be made to ensure that a project goes smoothly. Whether you are archiving born-digital materials or digitizing non-digital materials, you'll most likely need to know about image formats. In the next chapter, you'll learn how computers store image data, what the different common formats for images are, and which ones are best suited for digital archiving. Images are stored in different ways according to their formats, and some lend themselves better to archiving than others.

References

- Andrews, Jean. 2006. *A+ Guide to Managing and Maintaining Your PC*. 6th ed. Boston: Course Technology, Cengage Learning.
- Chen, Li, and Joyce White Mills. 2002. "What's Under Your PC's 'Hood': A Primer for Today's Machines." *Computers in Libraries* 22, no. 7: 14. EBSCOhost.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Gayde, William. 2020. "The Science of Keeping It Cool." Techspot. <https://www.techspot.com/article/1969-cooling-science/>.
- Miastkowski, Stan. 2004. "No More Cable Confusion." *PC World* 22, no. 8: 158. EBSCOhost.
- Potoroaca, Adrian. 2020. "A New Type of DRAM Might Pave the Way to Instant-On PCs." Techspot. <https://www.techspot.com/news/83559-new-type-dram-might-pave-way-instant-pcs.html>.



Storing Images

IN THIS CHAPTER

- ▷ How do computers store and display data for images?
- ▷ Why are computers limited when it comes to storing data for images?
- ▷ What are some of the common file types for images, and how are they useful?
- ▷ How can images be edited for archiving?

Human beings have been trying to record the world in images for thousands upon thousands of years. In fact, in the past, many people who are now considered “artists” were considered to be merely practicing a craft—that craft being the recording of the world they knew.

In 1826, though, something revolutionary happened. This is the year that the first photograph was created by a man named Joseph Nicéphore Niépce, and for more than a hundred years, photography was a process of capturing light intensity and color using photosensitive chemicals that change color or opacity upon exposure to light.

In 1957, however, a new process was discovered. This was the year in which the first digital image was created, a grayscale depiction of the son of computer scientist Russell Kirsch, and people have continued to record images digitally ever since. As mentioned in the previous chapter, it is not uncommon to carry a little computer with you everywhere you go, and it’s typical for a smartphone to have a camera embedded into the device. Humans use this fact to take a lot of photos—in fact, the number of pictures taken by people around the world may now be in the trillions per *year* (Marr, 2018).

Chances are excellent that you will need to deal with digitally rendered images in some aspect of your archive, whether you are digitizing tangible items, like books, maps, or photos, or storing files that are born digital and do not have an original format that is tangible in nature.

You already know from the previous chapter that a computer interprets everything as numbers. When you look at a picture rendered on a computer, though, that might seem very strange indeed. How does a computer store a photograph as a number?

It is also possible for some software programs to render an image as part of the instructions for the program. In this case, there is no original image—the image is created as a part of the instructions. As a simple example, when applied to a web page, the following instructions create a circle upon loading the page that is red and is both half the height and width of its container.

```
.circle { background-color:red; border-radius:50%; height:50%; width:50%; }
```

Because size is part of the instructions, this circle may look different on different monitors.

Pixels

There is a style of artwork that you may be familiar with known as *pointillism*. With this art technique, the artist uses many tiny patches of color to express the image being painted. A very famous painting done in this style is *Un dimanche après-midi à l'Île de la Grande Jatte* by Georges Seurat, in which minute patches of color form a picture of people spending a sunny afternoon at the Island of la Grande Jatte in Paris.

The theory behind pointillism is that the human brain will be able to blend together these dots of color and correctly interpret what is being depicted, even without distinct lines and forms to define the subject of the image. In a way, computers operate similarly, tricking the human eye into interpreting many tiny bits of color as one cohesive image.

The real world is a continuum of information, with infinite variations in color and light. Computers don't like continuums, though. Continuums can't be saved in a file with an intrinsically limited size. Computers instead work with small, distinguishable pieces of data, stored as bits and bytes. So, rather than trying to preserve information about the entire world at once, a computer stores tiny bits of data that, when put together, form meaningful information. For images, this little bit of data is the pixel, which stands for "picture element."

USEFUL TERMS FOR VISUAL DATA

Color depth, bit depth	The number of bits used to store a color
Compression	Removing unnecessary information to make a file smaller
Compression ratio	The ratio between the size of the full and compressed images
High color	Color encoding that uses 16 bits per pixel
Lossless compression	Compression that loses none of the original data
Lossy compression	Compression that discards some of the original data
Pixel	The smallest amount of data for an image
Resolution	The number of pixels in an image
True color	Color encoding that uses 24 or 32 bits per pixel

A *pixel* is basically data about a single color, like red or green. It is the smallest amount of information that a computer can store about an image. Pixels are arranged in a grid pattern, kind of like a mosaic in which each tile is one color. The more pixels that are available, the bigger the image is and the more information there is about the image.

Resolution is a term that refers to the number of pixels an image has. As an example, imagine that you have two identical images of a mountain. The first one is 1000×1000 pixels and the second one is 2000×2000 pixels. They are the exact same image, but the second one has a higher resolution with four times as many pixels. The second image will be larger from the computer's perspective, have a bigger file size, and have more information about that mountain. Details that can't be distinguished on the first photo—like individual trees, for example—might be visible on the second photo because there are more pixels and therefore more information.

Speaking of pixels can be a little confusing because this term can also be used to describe monitors. A monitor or a television screen displays images using a grid of minuscule lights; again, it's a little like a mosaic, but with lights instead of tiles. For a monitor, each tiny bit of light is also known as a *pixel*, the smallest bit of information that can be displayed on the monitor. Again, the image being shown is composed of many tiny dots of light, but the human brain interprets the information all together as one cohesive image. Monitors with more pixels are able to display more detailed images (and are often bigger than monitors with fewer pixels).

When scanning images, you may see terms like ppi or dpi, which stand for pixels per inch or dots per inch, respectively. These are measurements of the image's resolution, as mentioned above. Again, the resolution of a digital image refers to how many pixels there are in an image—that is, the density of the pixels. Though this is usually measured in pixels per inch, it is possible to measure it in other ways, such as pixels per centimeter. The higher the resolution of an image, the higher the quality and the clearer the image and the more information there is available about the image. While dpi and ppi are both used as measurements of resolution, ppi is more accurate in terms of preserving images. The term dpi generally refers to the capabilities of a printer (Bioinformatics and Research Computing, 2008).

It should be noted that the higher the resolution, the more data is needed to store the information for the image, as well. That is, data needs to be stored about each pixel, and so more pixels require more data.

For your archive, once you've determined how much storage you have available and what kind of information needs to be stored, it's best to use the same resolution or to have standards for the sake of consistency. Choosing the optimal resolution is a bit of a balancing act, as a higher resolution is better and captures more information, but a resolution that is too high results in an image that is extremely large and cumbersome to store, retrieve, and view. In addition, for some projects, storing more data may not be more beneficial. For example, for an old photo, there may only be so much data that can be gathered by scanning the photo, so you need a resolution that provides as much information as possible without storing unnecessary information.

While there are a number of factors involved with choosing your ideal scanning resolution, the size of an image file depends in part on how much information is needed to store the colors in that image.

As discussed in the previous chapter, computers encode information using *bits*, which are single numbers (a one or a zero). Color information can be encoded using bits, too. If there is one bit available for colors, then a computer can display two colors because a bit has two choices; for instance, black and white. A zero could mean black, and a one could mean white. In many cases, such as in the case of displaying text, black and white is perfectly acceptable, and a computer can display quite a bit of information using only two colors.

The world isn't in black and white, though. To replicate the real world, more bits are needed to assign more colors. As an example, if there are two bits, then there are four possible combinations for bit values. A programmer could assign a color to each one of these bit values, and the computer would therefore display the colors that the programmer assigned when interpreting an image. This kind of situation might look like this:

Bit value	Displayed color
00	Black
01	Cyan
10	Magenta
11	Yellow

The more bits that are used to encode colors, the more possible colors that a computer can display, assuming that the monitor is able to display the colors, though this isn't much of a problem in modern times. This quality is known as the *color depth* or the *bit depth*, and refers to how many bits are being used and how many possible colors can be displayed. Eight bits allow for 256 total bit combinations, which can therefore represent 256 different colors. Use of eight bits and 256 colors was common in computers in the past. This number of colors is adequate to display images in a fairly faithful manner, but doesn't come close to the variety of colors available in the real world.

Modern computers can use quite a few more bits for each pixel. There is also 16-bit color, also called *high color*, which allows for 65,000 colors. The first five bits are for red; the next six, for green; and the last five, for blue (Orr, 2003). The extra bit for green is because humans are more sensitive to green light than to red or blue, and so the extra bit allows for more shades of green. That makes an image “appear” to be more true to life to the human viewer (Cambridge in Color, 2019). While this produces a sufficient number of colors for many images, there is an even higher level of color for computers, called true color.

True color can have either 24 bits per pixel or 32 bits per pixel. Using 24 bits per pixel allows for around 16 million possible colors. Although no computer can quite replicate the real world, true color comes pretty close, and humans can't really tell the difference (Beekman, 2005). With 24 bits, each color—red, green, and blue—in that order, gets 8 bits or a full byte. It works the same way with the 32-bit version, but the last 8 bits are for controlling transparency, or an area that is clear or has no color. This is important for computer-generated graphics, particularly those intended for use online, as an area with no color will display the color of the background behind the image.

Using 24 bits per pixel has some serious disadvantages, though—notably with file size. With 24 bits, a computer could either encode the information for three alphabet letters (an entire word in some cases, like “cat” or “dog”) or those same 24 bits could store

the information for one tiny dot of color in a picture, a pixel. This quickly becomes cumbersome and takes up a lot of storage space, which can be problematic, since while digital storage is a great space saver, it still costs money. Fortunately, programmers have come up with some ways to make image files a little smaller.

Compression

There's a saying that "a picture is worth a thousand words." With digital images, this is true in a literal sense; image files are big, and a great deal of text can be stored in the same amount of space as a single photograph. With 24 bits of information needed for every pixel in an image that uses true color, images can become very large very quickly. For example, the digital text version of this chapter requires about 68 KB of data to store. In contrast, figure 3.1, found later in this chapter, uses 23,621 KB of data.

So, what is compression? It's possible for a computer to eliminate redundant information about an image and still display the image correctly from the information that is retained. This is known as *compression*. Compressed files still have the necessary information to reproduce a file, but they don't need as many bits as the full file and thus need less space to store. There are different methods of compression depending upon the file type, and there are two general types of compression: lossless and lossy.

With *lossless compression*, the computer looks for patterns or redundancies in an image. For instance, it may detect that there is a large block of plain black pixels in the image. Rather than save the data for each of these pixels, it instead saves a mathematical code, or an algorithm, that lets the computer know what was there in the original. When a person wants to look at the image, a decompression algorithm decodes this information and displays the correct pixels on the monitor. As a very simple example, suppose that a person were grocery shopping and wanted three cans of soup. Rather than write "can of soup" on the grocery list three times, that person could write "can of soup \times 3" or "3 cans of soup." Both phrases have the same information and are interpreted in the same way, but the second method saves two lines of space. This method of compression is known as *lossless* because even though the information is compressed, there is no loss of information and the displayed image is exactly like the original.

Lossy compression methods are able to compress files much further than lossless compression methods. These methods are a little less precise in that they essentially abbreviate the pixels in an image, discarding information—such as minute color variations—that's not necessary for a human to interpret the overall image. If the compression is low or an image hasn't been compressed repeatedly, the viewer might never notice that an image has been compressed. However, because information is lost during compression, the image essentially becomes less accurate and less true to the original every time it is compressed.

For archiving, lossless compression is more desirable. Understanding the difference between lossy and lossless compression will help you choose the best balance between saving file space and having accurate information preserved in your archive. However, there are a few other useful terms to know when it comes to compressing files.

The term *compression ratio* refers to the ratio of the full-size image to the size of the compressed image in regard to how much data is required to store the image. The larger the difference between the first and second numbers in a compression ratio, the smaller the compressed image is in comparison to the full-size image. So if an image has a compression ratio of 1:10, the compressed image is ten times smaller than the original, and if

an image has a compression ratio of 1:20, then the second image is twenty times smaller. For saving space, a higher compression ratio is better.

The term *nonadaptive* means that the software compresses every image in the same way using the same method. *Adaptive*, therefore, refers to software that compresses an image based on the unique characteristics of that particular image. Compression can also be symmetric or asymmetric. This is important for images that will be viewed online. If a compression method is *symmetric*, then compressing and decompressing an image both take the same amount of time and work for a computer. If a method is *asymmetric*, then decompressing an image is much faster than compressing it. This makes for faster loading when patrons want to see an image that your archive might offer online (Orr, 2003).

A lot of options are available when it comes to compression, and different types of file formats offer different possible methods. The type of compression that a particular file format uses can make a difference in which one you find most suitable for your archive.

Image File Formats

When looking at an image file, you might notice that there are some letters at the end of the name that you did not type when naming the file, such as “JPG” or “PNG.” These are called *file extensions*. Their function is to let the computer know how the information is encoded. Without the extension, a computer isn’t able to figure out what it should do with the data. While a computer uses different combinations of bits to encode color data, it also needs information about the order of those bits so that the colors can be displayed correctly to the user. There’s a little more to saving an image than just encoding color.

Because there are only two possible values for each bit, a computer also needs instructions to know what the sequence of numbers it is processing is supposed to be—an image, a document, a video, a program, and so on.

There are quite a few different formats for images, which might seem a little strange. Why isn’t there one universal format for every kind of image data? Wouldn’t that make things easier for everyone: computer designers, users, and the computers themselves? It would no longer be necessary to convert formats, and any image program on any computer would be able to open any kind of image, since they would all be in the same format. Not every type of software can open every type of image file.

It’s true that a universal format would make things easier. However, each type of image format has different virtues that make it appealing for different purposes—and different drawbacks that make it unappealing for other uses. For your archive, some of these formats may have features that seem more useful to you than others. For the purpose of archiving, formats that encode a lot of information are likely to be the most desirable ones. However, if you need images that can be transferred over the Internet, you might want to consider a different format. Most image formats have potential uses for you, and so each should be considered carefully. You may also want to save the same item in multiple formats to gain the virtues of different file types; for instance, you might use a file type with a lot of information for storage, but offer a type that creates a smaller file to patrons viewing your collection online.

SOME COMMON IMAGE FILE FORMATS

BMP	Bitmap file
GIF	Graphics Interchange Format
JPEG, JPG	Joint Photographic Experts Group format
PNG	Portable Network Graphics
SVG	Scalable Vector Graphics
TIFF, TIF	Tagged Image File Format

BMP

A bitmap file, or BMP, is one of the more basic file types and works in a simple manner, with the information for the color of each pixel encoded going from left to right and top to bottom. It has some virtues from the perspective of archiving in that the BMP format is fairly old and well established and BMP files can be opened with a wide variety of programs. However, BMP files are not as suited for compression as some other file types and aren't well suited for transmission over the Internet.

TIFF

TIFF, or Tagged Image File Format, is a format that has the ability to be either lossless or lossy, although it's typically used for lossless storage. TIFF files are usually not compressed and contain a lot of information, and thus TIFF files are often quite large. For the purpose of archiving, though, more is often better, and so a TIFF format can have a lot of appeal to an archivist. The TIFF format is one of the preferred formats for images for the Library of Congress's collections and is one of the best choices for digital image storage (Library of Congress, 2018). However, TIFF is not a good format for displaying images online since they take a long time to download, and most web browsers can't display TIFFs at all, so this is not a useful choice for sharing your collection online.

GIF

The Graphics Interchange Format, or GIF, has a rather unusual method of storing color compared to other formats. A GIF image is limited to 256 colors, but it doesn't have to be the same set of 256 colors for each GIF image; each image can use 256 colors from 16 million possible colors. If a person saves an image as a GIF, the program that creates the GIF uses algorithms to determine the optimal colors needed to save the image as faithfully to the original as possible (Matthews, n.d.). This is a technique known as *indexed color* (Dale and Lewis, 2013). GIF images are also capable of having transparent backgrounds, whereas some other formats aren't able to store information for transparent or clear pixels.

Since they only use 256 colors, GIF images only need a few bits per pixel to represent all of these 256 colors. This makes GIF images naturally small. The format can also compress large areas of uniform color by indicating that there are a certain number of pixels with the same color rather than saving the information for each individual pixel (Matthews, n.d.).

GIF images work best in situations in which there are only a few colors, such as line art, logos, or grayscale images, and do poorly for full-color photographs. They can also be used effectively for the web. Because less information is needed for the image, it takes less time for the image to load when viewing a web page than with many other image formats. As another bonus, GIF images can also have the option of being interlaced or noninterlaced. This is another feature designed for web use in particular. If an image is *interlaced*, then it can load in stages on a web page. The lowest resolution image loads first, so the viewer is able to get an idea of what the final image will look like. The picture then reloads in increasing stages of resolution until the highest resolution is achieved. If it is *noninterlaced*, then it loads in stages from the top of the image down. Making an image interlaced increases the file size (Lake and Bean, 2008).

While the GIF format is also considered acceptable by the Library of Congress for storing images, GIFs are likely to be of best use to you if it is the original format of an image, if you are storing web pages, or if you are making your collection available over the web due to the color limitations—that is, it won't faithfully reproduce a scan of a complex, full-color photo or similar images (Library of Congress, 2018).

If creating an archive of born-digital materials rather than digitizing tangible materials, printing hard copies of digital photographs is a valid method of ensuring that those images are preserved. While you will lose the advantage of saving space that digitally storing materials has, you will have both a digital and a tangible copy of the item to preserve, which allows you to have the advantages of both.

PNG

A Portable Network Graphics or PNG file is another image type that is designed for use on the Internet. It was designed to improve upon the GIF format, offering more colors and higher compression than a GIF (Dale and Lewis, 2013). PNG files are smaller than GIF files and allow for true color, whereas GIF files are limited to 256 colors. These files are compressed according to patterns within the image. This compression is lossless, which makes it a desirable format for displaying images on the web, even complex images like full-color photographs. Like GIFs, PNGs can have transparent backgrounds.

This is a somewhat newer file format, which means that PNGs may not be supported in older web browsers (the browsers will not display the image), but this is not an issue in newer browsers. PNGs also cannot be animated, as GIFs can, although there are some extensions to the format that can be animated—the MNG and APNG.

JPEG

The term JPEG, also known as a JPG, stands for “Joint Photographic Experts Group.” This is a very commonly used file format. Rather than storing information pixel by pixel, a JPEG file takes averages of a range of colors to form the image. This is handy from the perspective of storage, since JPEG files can be highly compressed. JPEG images use lossy compression, and the amount of compression can be controlled. This is important to consider when working with JPEGs; if compressed using a lossy compression format, JPEGs will lose information after each edit and save and their quality will become poorer over time. JPEGs are a good choice for photographs, but aren't a good choice for any images

with large patches of uniform color or images with sharp, precise lines. For instance, a JPEG wouldn't be a good choice for a simple logo.

JPEGs are a good choice for complex images that are intended for use online and have a feature similar to GIF's interlaced images. For a JPEG, this is known as *progressive encoding*, and also allows an image to be loaded in stages on a web page. Images without this feature have what is known as *standard encoding* (Lake and Bean, 2008).

Note that there is what is essentially an updated version of the JPEG: the JPEG 2000 (the file extension is .jp2). It offers a number of advantages over the JPEG, but is not as widely supported.

SVG

An SVG, or Scalable Vector Graphics file, is less commonly used than the other file types. The other image formats discussed so far are all what are known as *raster graphics*. This means that the information for the image is stored pixel-by-pixel, encoding the color and number of each pixel. An image doesn't have to be stored that way, though. Instead, an image can be stored based on its shape. This is known as a *vector graphic*.

If you've ever tried to make a standard digital photograph very large or very small, you'll notice that there are some problems. Shrinking an image involves essentially taking averages of all the pixel color values and can cause the image to look distorted or strange. Expanding the image, on the other hand, can lead to a "pixely" look. An image file can't create more information than is available, so it creates an estimate, which leads to the pixely look.

A vector graphic is saved by its *shape* rather than by individual pixels. This has one major benefit—scalability. A vector graphic looks exactly the same regardless of how big or, up to a point, how small it is. There are a few ways to store vector graphics, but the SVG format is one of the common ways and is a preferred format for the Library of Congress (Library of Congress, 2018). Vector graphics can also be converted into a raster image. This process is known as *rasterizing*.

PDF

Better known as a document format, the PDF or Portable Document Format is also a potential way to store images. For example, when digitizing a book, you may be actually storing images of the book rather than its text. Converting those images to a PDF file format would preserve those images in a way that replicates the look and feel of the original object. The PDF format will be discussed more in the following chapter.

Raw Formats

In traditional photography, an image is captured on film treated with photosensitive chemicals. It's not much different for digital photography, but the item that is capturing the information is a device with photosensitive sensors. These sensors detect only light intensity, so they are given filters to additionally capture amounts of red, blue, and green light. A single sensor can only detect one color, and so they are usually arranged in an even array of red, blue, and green sensors known as a Bayer pattern. When you capture the photo, a file is created that captures information from every sensor in the camera (Fraser, 2004).

A *raw* format is the file format that is created upon initially capturing an image with a digital camera, and these formats may be specific to a particular brand of camera. The need to work with a raw format may arise when working with digital photography. These formats often have more data than the format to which they are converted, which is another important consideration to keep in mind (London and Stone, 2012). That is, the raw format contains as much information as possible for the image, while converting the image to a more usable format discards some data (Fraser, 2004). Digital negative, or DNG, files are a type of raw format sometimes used with cameras that is preferred by the Library of Congress (Library of Congress, 2018).

Every file type has different potential uses as well as drawbacks. Table 3.1 summarizes the features of the image file types discussed so far.

Native Formats

The formats listed above are general image formats. Many different programs are able to read these files and display the image back to the user. However, it's important to note that many image files are only able to be read by the software that created the file or by specific types of software. If you don't have the right software, then you can't see the image that is stored with such a format. For instance, the extension "PSD" is the extension for a file created by the photo editing program Photoshop. A general program can't read the information in this file; only programs that are able to open Photoshop files (many painting or photo editing programs can) are able to read the information contained in the file.

A format that is specific to a program like this is known as a *native format*. There are many instances in which you might need to work with native formats, such as when using editing software as mentioned above.

If you have the proper software, working with a native format is not an issue. However, it is possible that you may experience problems arising from native formats. Some-

Table 3.1. Image File Types and Features

FILE TYPE	BENEFITS	DRAWBACKS
BMP	Usable with a variety of programs	Low compressibility, may have limited colors
TIFF	Stores a lot of information, has lossless compression, preferred by the Library of Congress	Creates large files, not suitable for displaying online
GIF	Small files, good for web pages and sending information over the web, can be used for animations and images with transparent backgrounds, can load on a web page in stages, lossless compression	Limited colors
PNG	Appropriate for the web, lossless compression, more colors available than a GIF	Cannot be animated as easily as a GIF, not supported in as many programs as GIFs
JPEG	Usable with a wide variety of programs, compression can be controlled, good for photos on the web, can load on a web page in stages	Lossy compression, loses noticeable amounts of information after repeated saves
SVG	Useful for storing vector graphics, can be rasterized	Usefulness limited to vector graphics
Native formats	Often have a lot of information about an image	Limited by the types of programs that can open and utilize the file

times older files cannot be opened with newer versions of the same software that created the file, and if the software for opening a certain type of file is lost, it may be difficult or impossible to decode the information in a file.

As an example, in 1985, artist Andy Warhol used an Amiga 1000 to create some digitally rendered works of art. In 2014, the data for these images was discovered on floppy disks. While the data could be retrieved, the program that was used to make them was obsolete by decades and the file formats were unfamiliar. In this case, the software was able to be reverse-engineered in order to view the files, but this was a fortunate example. These files could just have easily been forever inaccessible (The Andy Warhol Museum, n.d.).

There are a wide variety of file formats available—both common and uncommon, proprietary and free for use—and it should be noted that the file formats discussed so far are really just a handful of the most commonly used file formats today. This may change in the future as new formats are developed. There are many, many more image formats. Some of these are quite specialized—for example, ICO files are used for icons mainly on computers using the Windows operating system, among many other uses. They are used for the tiny icons that appear at the top of browser windows; these are often used to display a miniature version of a website’s logo.

Similar to files that are in a native format, some file formats are rarely used now, and it’s possible that part of your project will be to convert these files to a more accessible format. This is a very important part of digital archiving, but can be challenging in some ways. Not only do you need a program that can open the file you want to view and save it as the file format that you want to save, but you will need to be able to determine if the file has been successfully reproduced in the second format (Digital Preservation Management, 2014).

Changing a file’s format is inherently going to change the nature of that file, as well. This difference might be unnoticeable by a normal user or it might change quite a bit about the file. For example, the SHG file format, a type of image format, was used for old WinHelp (Windows Help) files, an obsolete file format used for digital help manuals. Because it is, in a way, tied to this obsolete file format, it is difficult to open SHGs at all. Additionally, this type of format is not merely an image: SHG files also feature clickable “hot spots,” and so finding a file format that can replicate this may be a challenge.

Additionally, you’ll need to decide if having a file in the new format is sufficient or if the original needs to be preserved as well as part of the archiving process. This is more likely to be of concern when addressing the archiving of born-digital materials.

Determining the optimal format for your collection is not the only matter that must be considered when creating your archive. Just because an image is saved at the optimal resolution with the best format for your archive doesn’t mean that the quality is inherently good. In some instances, images can benefit from adjustment.

Editing Images

If you have a physical format for an image, like a photo or a painting, it needs to be copied using a scanner or a camera in order to convert it to a digital format. After capturing images, you may want to adjust the quality of the image. For instance, if you scanned a hundred-year-old photo of the inside of an office, and there wasn’t enough light available to get a good photo in the first place, it’s possible using modern technology to adjust the

contrast in order to get a clearer image from a muddy one by digitally altering the contrast. It's important to use methods that don't permanently alter or remove information, however.

Many photo-editing programs have something called a *histogram*. This may not be labeled as such in the program, however, so you may need to browse around your particular program to look for this option. A histogram looks a little like a tiny graph.

As mentioned earlier, 8 bits can produce 256 different colors. When a computer is using true color, each color channel—red, green, and blue—gets 8 bits for color. This means that each color can have 256 different levels of intensity, essentially ranging from very dark to very light. A histogram shows these levels. Histograms can also display the general range of darks and lights in an image, which is particularly handy for black-and-white images.

In a histogram, pure black is represented by a value of 0, and pure white is 255. A program with a histogram gives a person the ability to adjust these values in an image. As an example, suppose that the office picture in the above example is in gray scale. If an image's histogram shows that there are no white values, this indicates that there are a lot of grays in the image, but no strong white points. There are usually some arrows beneath a histogram that allow the user to manually adjust it. If a person moves the arrow to where the histogram levels start, then the brightest grays present in the current photo will become white. This is a good way to get better contrast (and offers a little more control than the more common automatic brightness/contrast adjustment option), which both improves the aesthetics of a photo and makes it easier for the user to interpret a photo. This method doesn't lose information from the image file, but does change the file so that it doesn't look exactly like the original. Figure 3.1 shows a grayscale photo with its histogram, as well as the same photo adjusted for higher contrast.

For color photos, there are also sometimes histograms for each color: red, green, and blue. These show the amounts of each color in the photo, and adjusting them will change that amount, making it more or less intense. This may be of use to you in some situations—for instance, if you find that your scanner captures a particular color poorly, you can adjust things to be more true to the original image. The equipment that you use can have a large impact on the quality or faithfulness of the images that you capture and store. Similarly, it should be noted that monitors need to be carefully calibrated so that they display colors correctly. It is possible, for example, for a monitor to display colors that are too blue in tone or too yellow. If editing color images on an uncalibrated monitor, those images may only display their colors correctly on that monitor.

Key Points

In this chapter, you learned how computers store information for images, how to make image files smaller or more convenient, and why there isn't a single image file format that will suit every archiving situation.

- Image files require a lot of space to store all of the information needed to save and reproduce an image, and the more colors that are available in an image, the more space is needed for storage.



Figure 3.1. Using a histogram.

- For archiving, choose a resolution for your collection that is high enough to be useful, but not so high that it's difficult to store and retrieve images in a practical manner.
- There are a wide variety of file types available for images, all of which have different practical uses for an archive.
- Along with choosing things like color, resolution, and file type, it's also possible to adjust images to make them clearer or more faithful to the original image.

In the following chapter, you will learn about how computers store text and learn about some of the file types available for text storage. Because archiving text sometimes involves capturing images of printed text, there is some overlap between storing images and storing text. In addition, you will learn more about why computers sometimes have difficulty “reading” text, as your patrons will likely be interested in searchable text.

References

- The Andy Warhol Museum. n.d. *Warhol and the Amiga*. Accessed April 30, 2019. <https://www.warhol.org/exhibition/warhol-and-the-amiga/>.
- Bioinformatics and Research Computing. 2008. “Resolution.” <http://jura.wi.mit.edu/bio/graphics/scanning/resolution.php>.
- Cambridge in Color. 2019. “Digital Camera Sensors.” <http://www.cambridgeincolour.com/tutorials/camera-sensors.htm>.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions*. 2014. “Obsolescence: File Formats and Software.” <https://dpworkshop.org/dpm-eng/oldmedia/obsolescence1.html>.
- Fraser, Bruce. 2004. “Understanding Digital Raw Capture.” Adobe. https://www.adobe.com/digitalimag/pdfs/understanding_digitalrawcapture.pdf.
- Lake, Susan, and Karen Bean. 2008. *The Business of Technology: Digital Multimedia*. 2nd ed. Mason, OH: South-Western Cengage Learning.
- Library of Congress. 2018. “Library of Congress: Recommended Formats Statement 2018–2019.” <https://www.loc.gov/preservation/resources/rfs/RFS%202018-2019.pdf>.
- London, Barbara, and Jim Stone. 2012. *A Short Course in Digital Photography*. 2nd ed. Upper Saddle River, NJ: Prentice Hall PTR.
- Marr, Bernard. 2018. “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.” *Forbes*. <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#cfb6ffd60ba9>.
- Matthews, Rick. n.d. “Digital Image File Types Explained.” Accessed April 30, 2019. <http://users.wfu.edu/matthews/misc/graphics/formats/formats.html>.
- Orr, Genevieve. 2003. “Image File Formats.” Willamette University: General Graphics Resources. <http://www.willamette.edu/~gorr/classes/GeneralGraphics/imageFormats/index.htm>.



Storing Text

IN THIS CHAPTER

- ▷ How are letters and other characters encoded in binary?
- ▷ What are the common methods of encoding letters and characters?
- ▷ Which text file formats are best for archival storage, and why?
- ▷ What is searchable text?
- ▷ Which settings are best for scanned documents?

Documents and different forms of written information are very likely to be part of your archive in some respect, and people have been recording written information in a variety of ways for a very long time. The first known samples of writing are in the form of cuneiform tablets from Mesopotamia, and for thousands of years, humans recorded things painstakingly by hand, or sometimes by using things like woodblocks or engravings, which also required a lot of time and labor to produce the original printing block.

This rapidly changed with the invention of movable type, which allowed for characters to be printed over and over again without the need for a human to actually write them. Movable type made it possible to quickly, easily, and cheaply produce written materials en masse, making them available to many people, and people have been trying to create better and better methods of distributing written materials ever since.

In 1897, the first manual typewriter was invented, making it possible for an individual to create their own printed documents, and over the next few decades, several other innovations made it possible to create and re-create text information rapidly. For example, the Flexowriter typewriter used a paper tape to record the characters required for a document, and then that same tape could be used to re-create the document over and over again (Kunde, 1986).

In 1964, the IBM company created an improvement on this method: the MT/ST (Magnetic Tape / Selectric Typewriter), which used magnetic tape to record and re-create

text documents. Essentially, this was a recording of digital material that could be used to reproduce a document. What was particularly innovative about this invention was the fact that magnetic tape, unlike paper tape, is a rewritable method of data storage (Kunde, 1986). This method of storage is actually still in use today and will be covered later in this book.

What this means to you is that there are thousands of years of recorded information that could potentially be stored, and if you are interested in storing born-digital information, what could be considered digital text files have been around for longer than most people probably realize.

Text information essentially falls into two categories: text information that is in the form of a digital image of a printed text item, like a book, and text information that was “born digital” and whose original format is an electronic text file. In the previous chapter, you learned that pixel color and position are key to storing information about an image. For a text document, the character is what is most essential.

Encoding Text

Just like any other kind of data, text data is encoded with bits. Rather than colors, as in image files, the bits represent characters, like the letters in the alphabet. One bit has the possibility to represent two characters, with a 0 standing for one letter and a 1 standing for another. That’s not particularly helpful. More bits are needed to encode the entire alphabet.

Whenever you type on a keyboard, each of those little keys is sending an electrical signal to the computer. You’re communicating that you want the particular key that you press to activate a command in the computer. In the case of a word processing program, the key press is an instruction to display a letter or a character. The computer decodes the electrical signal generated by the key you pressed and converts it to a binary sequence, which is then stored and displayed to you on the screen.

When encoding colors for images, there’s a basic method as to how the bits are encoded that’s pretty logical. In true color, there are three sequences of eight bits. Each of these bit sequences tells the computer “how much” red, green, or blue there is in a pixel, with 0 being none of the color and 255 being the greatest possible amount of that color. For example, 255 red and 0 blue and green would be pure red, while 150 red and blue and 0 green would be purple. This is a fairly logical approach.

Determining how many bits and what sequence of numbers represents each letter in a text file, however, is essentially arbitrary. Any combination of 0s and 1s could represent a letter. This leads to some compatibility problems between different systems of encoding letters in binary.

In the past, there have been numerous ways to encode characters, called *character sets*. However, if different computer and software manufacturers use different methods of encoding text characters, then files for text information are incompatible between computers and between different software programs. One software program will not be able to determine what the characters from another program mean or will display characters incorrectly. For example, if the letter *A* is coded as 001001 for one computer and it’s coded as 100100 for another, then text files created with one computer can’t be interpreted correctly by the other. You can’t share the information, and you’re limited to the kinds of computers your files will work with.

In 1960, there were at least sixty different character sets in use, nine of which were in use by IBM's computers alone. Because the ideal situation is to use formats for digital information that can be opened with a wide variety of software programs, this situation would be a nightmare for archivists. To avoid problems like this, computer manufacturers eventually agreed to use some particular character sets for text in order to increase compatibility (Dale and Lewis, 2013).

One of the first standard character sets was the *ASCII set*, or the American Standard Code for Information Interchange. This is a set that is intended for use by personal computers; there's another standard set, called the *Extended Binary Coded Decimal Interchange Code*, or EBCDIC, which is intended for use by larger computers, like servers and mainframes. The two do not overlap as far as encoding characters goes. For instance, in ASCII, the letter *A* is represented by the sequence 01000001, while *A* in EBCDIC is 11000001 (Fuller and Larson, 2008).

The ASCII set originally used seven bits to represent each character with an eighth bit used as a "check" bit to check for accuracy as the data was transmitted through the computer; this allowed for 128 characters total. This encodes all the letters of the alphabet for English, both lower and upper case, numbers, and basic punctuation. Having separate sequences for upper- and lowercase is necessary because a computer doesn't perceive upper and lower case letters as being the same, like a human can. While a person could read *T* and *t* as being the same letter, "tee," a computer doesn't distinguish this. The bit combinations for these two letters are completely different depending upon whether or not they are capitalized (Dale and Lewis, 2013).

A later version called the *Latin-1 Extended ASCII set* used all eight bits and had 256 characters. This allowed for some accented characters and extra symbols. Like capital and lowercase letters, computers are not able to tell that a normal letter is essentially the same as one with an accent; they are composed of two entirely different sets of bits (Dale and Lewis, 2013).

There's a problem with using only eight bits and 256 characters, though. It works just fine for words in English and quite a few other languages that use the same alphabet. However, it doesn't work well for *every* language. More characters are needed to represent all the languages of the world. The *Unicode character set* is designed to address this problem. Unicode uses 16 bits per character rather than eight, allowing for 65,536 different possibilities. This set includes characters in languages other than English and a wider variety of symbols (Fuller and Larson, 2008). This system can also use more than 16 bits if needed for a character, so it's a flexible method of encoding text data (Dale and Lewis, 2013). For convenience, the Latin-1 ASCII set is encoded in the same way in the Unicode set—that is, the first 256 character codes in Unicode are encoded exactly the same way as the corresponding characters in the ASCII character set. This makes the two sets compatible from a programming perspective (Fuller and Larson, 2008).

This is a very convenient system for archivists as well as programmers. However, a character set only takes into account which character should be displayed. It does not take into account the *look* of the character that is being displayed. This requires another type of encoding.

Though not an essential aspect of storing and retrieving text data, if preserving the original look of a item is important, you may nevertheless find it useful to understand how a computer renders fonts.

Fonts are not an aspect of a word processing program, nor are they inherently embedded into a text file. The font file and the text file are two separate files. When you use a word processing program and choose a font from the menu, what you are actually seeing is a list of font files that are available on the computer you are using. Computers typically come with a large set of fonts available for you to use, and sometimes fonts are specific to a particular brand of computer. When you install a word processing program, it may come with some fonts, as well. You can also add font files to a computer yourself.

There are a wide variety of types of font files, with some of the common ones being TrueType fonts, PostScript fonts, and OpenType fonts. The TrueType format can be used on most modern operating systems—that is, the major operating systems in use today can open this kind of format. PostScript fonts were the first standardized fonts and have been overtaken by the more convenient TrueType format, as PostScript fonts were operating system specific (a font for a Windows machine could not be used on an Apple computer, and vice versa). However, they are still notable in that these files are still being used. OpenType fonts are a more complex format that may contain TrueType data, PostScript data, or both (Felici, 2011).

If you have born-digital materials that you want to store, it may be necessary to also have the font files used in order to get the file to display as the file's creator originally intended. This is not only true of text documents, but web-based materials as well, should you decide to store this type of material.

Specifying which fonts should display on a website is actually a little tricky, and there are essentially three ways to do it. One is to specify a font in the site's coding. This requires that the designated font file must actually be stored on the user's computer. That is, the coding of a web page can specify which font to display, but it will be necessary for the computer that is displaying the web page to have that font file. To ensure the correct look and feel, a developer may specify a primary font to render, then a secondary or even a tertiary font that will be selected in the event that the desired font is not available on the visitor's computer. This can pose a bit of a problem to a web designer who wants a lot of control over how the site will display, because they, of course, cannot control what files are on the computer of the person viewing the website.

There are a couple of ways to get around this. One is to only specify fonts that are installed by default on nearly any computer (Arial and Times New Roman are a couple of examples), known as *web-safe fonts*. Another is to use *web fonts*, which are fonts available online. Rather than using a local file to correctly render a font, the website uses a file available online to do it, so the site should display correctly no matter what computer is being used to display it. Google Fonts is a commonly used service that offers free fonts for this purpose.

Another way to get around this problem is to actually include font files with the files that are used to generate a website. This can have its own set of problems. Fonts intended for the web have their own file types, and different browsers support different font files. Most current web browsers support WOFF or WOFF2 type files, but there are other types of font files intended for older browsers (MDN Web Docs, 2019).

Regardless of what types of text files you are storing, it is important to note that fonts are covered under copyright law. This may cause some restrictions regarding what you can store and what you can do with font files. A later chapter will give you a quick overview of the general rules of copyright law.

Text Formats

Storing text can be a complicated process for an archive. In essence, there are two different types of text that your archive might need to store. The first kind is a text file, text information that has been encoded by a computer and is “born digital.” The second is text that originally had a physical format. Even though it is text from the user’s perspective, it can’t necessarily be stored as a text file. Instead, it needs to be treated as an image.

SOME COMMON TEXT FILE EXTENSIONS

DOC, DOCX	Microsoft Word documents; DOC is the older version
HTML	Hypertext Markup Language file
ODT	Open Document Text
PDF	Portable Document Format
RTF	Rich Text Format
TXT	plain text file
WPD	WordPerfect Document
WPS	Microsoft Works document
XML	Extensible Markup Language file

Text File Formats

There are a number of formats for text files, and choosing an appropriate format can be a tricky process. Many formats and extensions for text files are proprietary and belong to a specific software company. For instance, the extension “.DOCX” indicates a file created with the software program Microsoft Word. A proprietary file like this is typically only compatible with the program that created it—that is, only another copy of the program is able to open a file with this extension, so you can’t use just any word processing program to open a DOCX file. It must be opened with Microsoft Word or another program that is specifically designed to open this file format.

To make things even more complicated, files created by one version of a particular type of software might not be compatible with newer or older versions of the same software. For instance, older versions of Microsoft Word created files with a “.DOC” extension. New versions of the program are still able to open these, but older versions can’t open the new DOCX files without a converter.

These are undesirable qualities for archiving, since it’s best, whenever possible, to avoid proprietary formats. Not only do proprietary formats make it difficult to share information with patrons or with other archives, but they also make the files more

vulnerable to becoming obsolete due to the lack of necessary software. Companies that create software programs can go out of business, potentially making any file that used their particular file format useless and unable to be opened and read. An archiving format must be able to withstand these kinds of setbacks. If you want to archive a digital text file, you may need to convert it into a format that is more general. However, documents can look different in different file formats, so if preserving the original look and feel of a document is important, then you need to choose a file format that is capable of creating the appropriate formatting. There are a couple of file formats for text that are more general and not proprietary and thus are more appealing to be used for archiving.

TEXT FILE FORMATS FOR ARCHIVING

Plain text (TXT)

Rich Text (RTF)

Open Document Format (ODF)

Hypertext Markup Language (HTML)

Extensible Markup Language (XML)

Portable Document Format (PDF)

Plain Text

Plain text is a format that any word processor can open. While there are several types of file formats that use plain text, the basic file extension for a plain text format is “.TXT,” which is easy to remember because it sounds just like the word “text.” This is a format with a lot of appeal because it is able to be opened with a wide variety of programs. This removes all the problems caused by using proprietary software and protects the file from obsolescence due to software problems. When you save your file in a word processing program, the program usually has a “default” format that it automatically saves to. However, you typically also have several options other than the default that you can choose, including TXT files. The program Notepad, as an example, creates files with this extension as its default. The previously mentioned Microsoft Word uses the DOCX format as its default setting, but can have its information saved as TXT files, too.

This file type has some serious limitations, however. There is little formatting available with plain text files—nothing beyond using simple tabs or a return to designate paragraphs. There are no variations in font style or font size available, no tables, and no ability for images or other multimedia elements to be embedded into the file. These kinds of files are useful for pure text, but can’t preserve the original look and format of a document unless that document was already in this format. There are some ways in which plain text can be used with formatting, however, which will be discussed further in a moment.

There are many types of data that can be communicated using just plain text. For example, a CSV or comma-separated value file uses plain text without formatting and is used for storing

and transferring data. Such a file uses commas to distinguish individual pieces of data. For instance, if you wanted to import a list of fruit into a spreadsheet using a CSV file, it might look like this:

orange,banana,apple,peach,pear

The comma character separates each value in the file, making it possible for a program to distinguish where one piece of data ends and another begins. These types of files can be created and edited in normal word processing software and are really a type of text file, although you might not perceive it that way because it doesn't convey data that is relevant to humans, like a book, but rather transfers data that can be processed by a computer.

Rich Text

Rich Text Format files, which have the extension “.RTF,” are simple and able to be opened with a variety of programs, just as TXT files are. They are able to have some formatting, such as changes in font style, size, and color. The program WordPad makes RTF files as its default setting. However, not every program can open these types of files, and they are still somewhat limited in the formatting options available. There is another format available that is capable of much more formatting and a wider variety of information, however.

Open Document Format

The *Open Document Format*, or ODF, became the International Organization for Standardization's, or ISO's, International Standard format in 2006. ODF actually refers to several types of useful documents. The extension “.ODS” is for spreadsheets, “.ODP” is for presentations (a generic term for the program PowerPoint), and “.ODT” is for word processing documents; ODT stands for “Open Document Text.”

An ODT document has formatting capabilities like the somewhat more familiar DOCX format, but has a notable difference in that it is not proprietary. This is an open-source format, which means that it does not belong to a particular software company and is not dependent upon a single company for software appropriate to open an ODT file. The open-source software Apache OpenOffice creates this file format as its default. This software is freely available online (The Apache Software Foundation, 2013). There are many other programs capable of creating this file format, however, including Microsoft Word. The ODT format is one of the file formats preferred by the Library of Congress for text storage. If you have text files that are originally in a proprietary format and have formatting that needs to be encoded, you can convert the file to the ODT format instead, making it more suitable for archiving (Library of Congress, 2018b).

While open document format files are good for storing text that has formatting, it's also possible to incorporate formatting into plain text files, but in a roundabout way. HTML, for instance, is written in plain text but displays formatting when viewed using browser software. Plain text, when used for this purpose, can also contain metadata: information about the file and its contents.

Hypertext Markup Language

Hypertext Markup Language, HTML, is a language that was designed for the purpose of creating web pages and sharing information between computers online. HTML docu-

ments have a similar advantage to TXT and RTF files in that many different software programs known as *web browsers* are able to access the information in an HTML document, and likewise, many different software programs can create HTML documents. HTML is not dependent upon a particular company staying in business. Microsoft Edge, Google Chrome, and Firefox are some commonly used web browsers, but there are many others available to use.

If you've ever browsed around the Internet, you've downloaded a web page. Typical web pages have attractive visual elements, like buttons, pictures, or background colors. All the information for those elements is encoded in an HTML document. Web designers use the HTML language to create, or code, web pages, and their coding is then interpreted by the browser. HTML documents contain *only* text, regardless of what nice pictures might be visible in the browser. The browser is like a translator between the HTML document and the person who wants to see the information contained in the document—that is, the browser turns what would be incomprehensible text into something that makes sense for the user.

Outside a browser, an HTML document is just like a plain text document—no formatting or special fonts whatsoever. An HTML document viewed in a word processing program and not in a browser can look a little odd. To see what this is like, try going to a web page, then right-click in an empty space and select “view source” or “view page source” (on browsers for Macintosh computers with single-button mice, this is a little more complex and involves going through the navigation bar for the option to view the page this way). This will show the plain text file for a web page. You'll see any text that was on the page, but you'll also see words and symbols around the text that don't necessarily make sense. These are the instructions for the browser. Figure 4.1 shows a very simple web page both as it's seen in plain text and as it's seen in the browser Firefox.

The creator of an HTML document uses different sequences of text that act as instructions to the browser and lets the browser know how the document should be displayed. As an example, if a person creating a web document puts the characters `<p>` and `</p>` around some text, then the browser interprets all the text between the characters as being a paragraph. The `<p>` sequence is known as a *tag*, in this instance, a “p” tag; `</p>` is a closing “p” tag. The purpose of a closing tag is to let the browser know when a certain type of formatting ends, though this is not necessary with all types of tags. As another example, the tags `<i>text</i>` tell the browser that the word “text” should be in italics. The tag `<i>` is for italics. So, it's not *necessary* for the original HTML document to have any formatting of any kind—the browser takes the information encoded in the document and displays the information in the way that the designer of the document instructs. The text doesn't need to be in italics in the original document. The browser will convert the text to italics based on the tags. Because only plain text is needed, no formatting is required for proper display.

There are many ways to format a document using the HTML language, with a wide variety of tags, or instructions for the browser regarding formatting and content. However, tags don't apply solely to text in regard to formatting. For instance, the `` tag tells the browser to insert a particular image at a certain location in a web page. A generic example of what this would look like would be ``. The “src” in the character sequence lets the browser know how to locate the image “apicture.jpg.” Other instructions within the `` tag can tell the browser things like how big the picture should be or whether it needs a border.

```

<html>
<head>

<title>The Title of this Web page</title>

</head>

<body>
<p>You can do all kinds of things to a web page, simply by
using text in an ordinary text editor.</p>

<h2>Like Make Headings</h2>

<h2 align="right">Or Change the Alignment of Those Headings</
h2>

<p>Though the text is plain in the editor, you can do a wide
variety of interesting things to it.</p>

<ul>
<li>Like make bulleted lists</li>
<li>Make fonts <i>italic</i>, <b>bold</b>, or <u>underlined</
u></li>
<li><font color="gray">Change the color</font></li>
<li><font size="4">Change</font> <font size="5">the</font>
<font size="2">size</font></li>
<li><font face="verdana">Or even change the font</font></li>

<p>Your patrons may be interested in images, which are easily
incorporated into an html document.</p>


</body>
</html>

```

An HTML Document in Plain Text

You can do all kinds of things to a web page, simply by using text in an ordinary text editor.

Like Make Headings

Or Change the Alignment of Those Headings

Though the text is plain in the editor, you can do a wide variety of interesting things to it.

- Like make bulleted lists
- Make fonts *italic*, **bold**, or underlined
- Change the color
- Change **the** size
- Or even change the font

Your patrons may be interested in images, which are easily incorporated into an html document.




Figure 4.1. HTML plain text and viewed in a browser.

There are a number of benefits to using this method for storing documents. Like TXT and RTF files, many programs can both make and open HTML documents. Both Notepad and WordPad, mentioned above, are capable of creating HTML documents, but there are numerous others, like the program Notepad ++, which is also capable of editing other coding languages. Unlike the TXT and RTF formats, an HTML document can incorporate multimedia elements. A browser program is required to display this properly, however. For example, if the text in an HTML document requires an image, then the HTML file and the image file are still two separate files; opening the HTML file in a text editor won't display the image. The instructions for the browser in the HTML file tell the browser to retrieve the image and insert it in the correct location among the text information when the web page is viewed.

HTML is a format recommended for data storage by the Library of Congress (Library of Congress, 2018b). It's a good format to use if you have multiple text documents that you want to link together or if you are saving information that is originally in this format. It's also handy to know about if you want to make your collection available online. HTML is a very good choice for sharing information, since most computers have a browser program for interpreting HTML documents. Files that use plain text are smaller than other types, since they only have characters and no other data. This makes information in this format ideal for transmitting over the Internet, since transmitting large files takes a long time. However, there are some difficulties with HTML. The HTML language is continually evolving, and so some instructions do become obsolete or are interpreted differently over time. In addition, different browsers can interpret the instructions in an HTML document slightly differently, so the same information may look a little different in different browsers. Web designers sometimes need to create multiple sets of instructions in a web page in order for it to display similarly inside different browsers.

HTML is not the only format or language that can enhance plain text. There is another, similar format that can be of great use to you and your collection, and understanding HTML will also help with understanding this format.

Extensible Markup Language

Extensible Markup Language, or XML, is similar to HTML in quite a few ways. It uses the `<tag>text</tag>` format to communicate information, just as HTML does. However, XML has a very different purpose from HTML. The main purpose of HTML is to instruct a browser program about what information should be displayed and how it should be displayed, and contains information regarding formatting and layout.

XML has no impact whatsoever on the layout or appearance of a document. Instead, it contains information about the contents of a document, or *metadata*. For instance, if you had a book in digital format with XML tags and used the tags `<title>Reading Is Easy</title>`, then the tags can communicate that the title of the book is "Reading Is Easy."

HTML has a set of tags that are universal. Everyone who codes websites must use the same set of tags. Again, it's like using a language—everyone coding with HTML is "speaking" the same language and communicating using the same pool of "words," or tags. With XML, it's possible to make up your own tags in order to communicate whatever information you think is relevant, like a book title or an author. However, this means that you're essentially making up your own language. There are ways around this, though. A *Document Type Definition*, or DTD, has a specific set of tags that you can use. Like using

HTML with its universal terms, this allows for better, more consistent communication (Combs, 2011).

So, what is the point of XML? Well, if you use the appropriate software or a database, it can use a document that is tagged with XML tags to do all sorts of useful things. For instance, if all of your documents have a tag for the book title, a software program can do things such as go and search for the text between the book title tags in every document that you have, then compile a list of every book title that your collection has. A patron could search for a particular book title, too. If all of the chapters in a book are tagged by their titles, a program could compile a table of contents automatically by locating, for example, all chapters with the same title tag, plus chapter titles or numbers based on another tag for this data (Combs, 2011).

As mentioned earlier, there are a lot of different DTDs, or vocabularies for XML, and these are typically designed for a specific type of information (Combs, 2011). That is, there are different standard ways of tagging metadata using XML. One of the preferred text formats for the Library of Congress collections is the EPUB, which includes an XML file that indicates all files included in the EPUB file as well as their intended reading order. Software readers for this format are able to interpret JPEG and PNG files, so images in these formats can also be included in such a document (Library of Congress, 2018a).

Documents using XML have many virtues and can include a lot of useful additional information, but are rather complicated, and learning about HTML and XML takes time. They can be important for your archive, though; XML, for instance, can be very useful for making your digital collection more easily accessible and searchable. There are still more options for storing text information, though, that are simpler and have plenty of appeal to both you and your patrons.

Portable Document Format

The *Portable Document Format*, or PDF, is a file type created by the company Adobe. The major virtue of a PDF is that it “feels” like a book to the user, as a PDF can be separated into “pages.” Like HTML, PDFs are capable of supporting both text and multimedia. PDFs are also capable of other handy features, like searchable text, bookmarks, hypertext links, and both annotations and metadata. PDFs are typically easier to construct than HTML or XML documents, as well.

There are several variations on the PDF format. One of the ones preferred by the Library of Congress for archiving text is the PDF/A-1. Like the ODT, this is an open format approved by the ISO and a number of companies create software that supports this file type. The ISO requires that future versions of software designed to view this file type will be backward-compatible, meaning that old files can be read with new software (PDF/A Competence Center, 2019). This particular format doesn’t allow for audio and video embedding, JavaScript, or other executable files (essentially, a software program), or encryption, so it’s a little more limited than a normal PDF file (Library of Congress, 2019). These limitations are to ensure that a PDF/A-1 file can always be displayed *exactly* like the original; features like video or audio usually require software external to the PDF reader to function, which presents problems for display and archiving (PDF/A Competence Center, 2019).

The PDF format can also store images instead of text, but this isn’t really recommended as a method of image storage (Library of Congress, 2018b). However, if you are

digitizing books, this is a good option to keep in mind, as you can use an image of each page in the book as a page of the PDF. This will also keep multiple images together in the same file and, again, will feel like a book to the reader.

Image Formats

If you are storing scans or photos of text, then you need to use an image file format, even though what you are capturing is text. As discussed in the last chapter, some formats are better for some purposes than others. This is also true of images of text. For instance, JPEGs aren't good with high contrast and sharp edges, which makes JPEGs a poor choice for storing images of text in most instances.

TIFFs are a better choice, since they capture so much data and are lossless, and they are the format recommended by the Digital Library Federation Benchmark Working Group. For pure black-and-white text or text with images that don't require tone, this group recommends that scans or photos are made at 600 dpi with 1-bit tone; this means that only black-and-white tones will be present in the scan, which is good for getting a very clear image with the most readable text possible. If text has or is part of a grayscale image, then they recommend using 300 dpi with 8-bit gray scale. As mentioned in the previous chapter, 8 bits will create 256 different shades of gray, which is sufficient for faithfully reproducing a grayscale image with little noticeable difference between the copy and original. For text that is part of a color image, the group recommends the same dpi, but 24-bit color, or true color (The Digital Library Federation Benchmark Working Group, 2002).

TIFFs are not suitable for display on the web, though. If you want your collection to be available online, you may want to store a JPEG in addition to the TIFF format, or to use a GIF or PNG, which are better with clear color boundaries than JPEGs, depending upon how much color is needed to faithfully reproduce the image. What this means to you is that GIF and PNG files are good with crisp edges, and your text will be less blurry than it would be with a JPEG.

A good image can make it easy for your patrons to read text, just as if the text was in a born-digital format. However, your patrons will most likely want, or even expect, *searchable* text. This term means that a computer is able to search through a document and match a word that the user types to a word found in the document. If the document is a text file, this is a simple matter. If it's a scan or photo of a text document, however, this is more complicated.

Searchable Text

Computers can't "read" in the sense that humans can. A computer, for instance, doesn't know that the word "computer" says "computer." Computers can't tell that the word "computer" on a screen and one in a book are the same thing, either. A computer perceives the typed word "computer" as a series of 8 sets of 8 binary sequences, while an image of the same thing is a set of binary sequences indicating colors. To a computer, these aren't the same thing at all. Although a computer can't "read," what computers *can* do is match things.

With Optical Character Recognition (OCR) software, a computer can turn text in an image into text that is searchable, just as if it was “born-digital” information, which makes it easy for people to find specific words or phrases in a document without actually reading the entire document. This also makes it possible for people who have problems with their vision to be able to access documents, as the digital version of the document can be used with other technologies to assist the visually impaired.

OCR technology is essentially a type of artificial intelligence software, and there are a few different ways that this type of technology can work. Basically, however, the OCR software breaks down an image of a document into blocks of text, then lines of text, then words, then letters. It then compares the shapes of characters to letter shapes in a database to find a match (Holley, 2009).

OCR technology today is pretty accurate, but its accuracy is very dependent upon the clearness and quality of the scanned material, and things such as stains or smears can cause problems for the software. For instance, imagine that there is an ink mark through a word on a printed page. If the line is thin enough to leave a little information about the letter’s shapes, a human can still read a word that’s been disrupted. A computer, however, will find this much more problematic.

There are fonts specifically designed for use with OCR technology, such as the font OCR-A. Invented in 1968, OCR-A was designed to be easily read by both humans and machines. This is a monospace font (all characters are the same width) and is still often seen on checks.

The accuracy of OCR software is typically represented in a percentage of accuracy, and it will guess the right character on an average of the given percentage. It’s possible to improve accuracy by correcting errors in the scan of a document (such as de-skewing pages) and by getting a good image capture in the first place; 300 dpi or higher is a good resolution for images that will be processed in this way (Holley, 2009).

Whether a document has been scanned with one bit, which produces pure black and white, or eight bits, which creates gray scale, can have an impact on your success with OCR technology, as well. If a document is very clean and tidy, with no spots, smears, water marks, or similar flaws, or if the document text has multiple colors, then a monochromatic, 1-bit color is the best choice. This will make the scan faster and more accurate, since the scan doesn’t need to detect the threshold at which a character is a character or the background; if an image is pure black and white, then anything black can automatically be considered a character by the program, and anything white can be ignored (Lais, 2002).

However, if a document has any kind of spotting, smearing, or other types of similar damage, these will show up as black spots on the scan, which may overlap or obscure letters and make it difficult for the software to distinguish letters from “noise,” things that have no relevant information. In this case, the image may be easier for both humans and computers to read if scanned in gray scale. Imagine that you made a photocopy of a document that has a faint coffee stain on it. Though you might still be able to read the original document, the photocopier will copy that coffee stain as being pure black, obscuring letters. This is what would happen in a 1-bit scan. In a grayscale scan, that coffee stain will show up as gray, making it still possible to interpret the letters under the stain.

Even if the best bit depth and resolution are chosen for a particular document, OCR technology is never 100 percent accurate and determining accuracy can be a challenge. Some libraries, including the Library of Congress, use the help of volunteers to review

the accuracy of documents that have been scanned using OCR technology and to make corrections (or to create digital versions of documents that are too difficult for OCR technology to handle at all) (Library of Congress, accessed 2020).

OCR technology is improving all the time, though, and it's even included with some software for creating PDFs. The convenience it offers for both you and your patrons makes this technology worth exploring.

It should be noted that OCR technology is designed for matching printed letters—that is, characters that are made using a machine. Interpreting *handwriting* requires the use of a related technology called ICR, or Intelligent Character Recognition. One type of software may be more suitable than the other, depending upon what types of documents you need to analyze.

Key Points

- Today, text information is encoded in a way to offer a lot of compatibility between documents and software from multiple companies.
- Although compatibility has been improved over time, many types of file formats are proprietary and should be avoided when possible.
- TXT, RTF, and ODF files are all nonproprietary text file formats and are suitable for archiving different types of text files.
- HTML files offer a variety of formatting possibilities, and XML allows for the insertion of useful metadata into a file. PDFs allow for patrons to read text information similarly to how they would read a book, and can store text and image information with equal ease.
- Patrons can benefit from the convenience that searchable text offers. However, to make image files readable by a computer, they need to be processed with OCR software. Using the correct number of bits and resolution for scanning presents the best chance for accuracy with this software.

While text and images are items that are commonly found in archives, you may find that you have the need to store other multimedia information, as well. In the following chapter, you will learn about how a computer processes and stores both audio and video information and learn which formats are best for archiving this kind of information.

References

- The Apache Software Foundation. n.d. "Reading ODF Documents (*.odt, *.ods, *.odp) with Apache OpenOffice." Accessed May 19, 2019. http://www.openoffice.org/why/why_odf.html.
- Combs, Michele. 2011. "XML Indexing." *Key Words* 19, no. 4: 123–34. EBSCOhost.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- The Digital Library Federation Benchmark Working Group. 2002. "Benchmark for Faithful Digital Reproductions of Monographs and Serials." Digital Library Federation. <http://old.diglib.org/standards/bmarkfin.htm>.
- Felici, Jim. 2011. "The Complete Manual of Typography: About Fonts." Adobe Press. <http://www.adobe.com/articles/article.asp?p=1743636&seqNum=2>.

- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine*. <http://www.dlib.org/dlib/march09/holley/03holley.html>.
- Kunde, Brian. 1986. "A Brief History of Word Processing (Through 1986)." Fleabonnet Press. <https://web.stanford.edu/~bkunde/fb-press/articles/wdprhist.html>.
- Lais, Sami. 2002. "QuickStudy: Optical Character Recognition." *Computer World*. http://www.computerworld.com/s/article/73023/Optical_Character_Recognition?taxonomyId=11&pageNumber=1.
- Library of Congress. n.d. "By the People." Accessed February 29, 2020. <https://crowd.loc.gov/>.
- Library of Congress. 2018a. "EPUB, Electronic Publication, Version 3." Sustainability of Digital Formats: Planning for Library of Congress Collections. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000308.shtml>.
- Library of Congress. 2018b. "Library of Congress Recommended Formats Statement 2018–2019." <https://www.loc.gov/preservation/resources/rfs/RFS%202018-2019.pdf>.
- Library of Congress. 2019. "PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4." Sustainability of Digital Formats Planning for Library of Congress Collections. <http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml>.
- MDN Web Docs. 2019. "Web Fonts." https://developer.mozilla.org/en-US/docs/Learn/CSS/Styling_text/Web_fonts.
- PDF/A Competence Center. 2019. "PDF/A FAQ." PDF Association. <http://www.pdfa.org/2011/06/pdfa-faq/>.

Storing Audio and Video



IN THIS CHAPTER

- ▷ How are audio and video data similar to and different from other forms of data?
- ▷ How do computers record and store audio and video data?
- ▷ What are some of the optimal formats for storing audio and video data?

The very first known audio recordings date back to the early 1850s. Created by a Parisian bookseller named Édouard-Léon Scott de Martinville, these recordings were made using a device he invented called a phonautograph, which recorded the sound by tracing a stylus on a paper treated with soot. The earliest recognizable recorded sound is his 1859 recording of a tuning fork. These early recordings were actually a visual representation of a sound—not meant to be played back—but they can be heard now using modern technology using a “virtual stylus” (BBC News, 2008).

The history of audio recording in general is somewhat messy in that there have been a lot of different ways to record audio, some more practical than others, including tin foil, wax cylinders, magnetic wire and magnetic tape, and even the aforementioned paper and soot method.

The history of digital audio recording is also a bit murky, because it’s really a series of inventions rather than a single one at a single defining moment. But the technique used to represent sound digitally, called *pulse-code modulation*, was invented in 1939 by a British telephone engineer named Alec Harley Reeves. In 1957, recording sound digitally on a computer became possible through the work of an engineer named Max Matthews.

Recording sound uses much, much more recent technology compared to the two discussed so far—that is, visual and textual data—but it is certainly a type of data that you may need to store. For example, as mentioned, there have been a variety of ways to record sound, and you may be involved in the process of gathering and storing audio from some of the more obscure and outdated methods of recording audio before they become impossible to play back. As another example, some archives store recordings of people telling

stories, information that may be invaluable to historians in the future (and even now), and so you will need to decide how to best store and preserve these types of materials.

From the perspective of a computer, the data required to store and display an audio file isn't an awful lot different from other types of data, and so you may find much of the explanation for how this works familiar at this point if you've read the previous chapters. However, audio files do have some of their own terminology as well as storage formats that are unique to them, and so there are some new things to learn when it comes to storing this kind of data.

Because audio data is often a component of a video file (unless the video has no sound), this chapter will begin by explaining exactly how a computer records and stores audio information.

Audio Data

In the real world, when you hear a sound, the sound is continuous in nature. Imagine that you are listening to a violin player. Each time the bow is drawn across the strings, it creates a continuous sound. In computer science, this is what is known as *analog* data—continuous information.

But as you know, computers don't like continuous information. They deal with information in small chunks of data, which can be stored as binary ones and zeroes. This is known as digital information, and the process of *digitizing* actually refers to this concept of turning analog data into digital data.

So, if you wanted to record the violinist in the earlier example using a computer, how can it be done? Whenever you record audio information with a computer, that computer doesn't record *all* of the available information. That's impossible (at least for modern machines). Instead, what happens is that the computer takes a sample of all the frequencies that exist at a single instance in time. It keeps doing this at regular intervals so that when you listen to all of these samples being played back, you hear the recording as being continuous.

You may notice that this is a similar concept to how a computer records visual information, as discussed in chapter 3, but by taking samples of color rather than samples of sound. A computer can't record all of the available visual information for an image, so instead it records averages, or enough samples of color to form a full image that a person can understand. A computer can't record all sound either, so it takes samples of sound, as well.

If a computer takes enough samples rapidly enough, then all those samples together create a good representation of the original sound. A rate of 44,000 times per second, or 44 kilohertz, is a good rate for most situations (Fuller and Larson, 2008). The higher this rate gets, the more realistic the sound becomes—up to a point. Once the sampling rate exceeds a certain level, the human ear is no longer able to distinguish an increase in quality.

If a sampling rate is much lower than 44,000 times per second, then the human listener starts to notice the difference between a recording and reality and is able to tell that he or she is hearing a recording—and not a very good one (Dale and Lewis, 2013). However, the optimal sampling rate is also dependent upon what is being recorded: The sound of a single human voice speaking, which you might want to record for an archive of spoken records, accounts, or stories, doesn't need as much information for an accurate

recording as, for example, a recording of a rock band. A human voice can be recorded with good accuracy at an only 11,000 times per second, or 11 kilohertz (Fuller and Larson, 2008).

The sampling rate also has an impact on the highest recordable frequency, or the highest pitch that a sound recording can capture. The highest frequency that can be captured is half the recording rate. For professionally recorded music, the standard recording rate is 44.1 kHz, which makes the top available frequency 22 kHz. This is good, since most people can't hear anything above 20 kHz; so essentially, this is an efficient sample rate for the recording industry (CARLI, 2017a).

Suppose that you wanted a good recording of your violin player, so you decide to use recording industry standards and record him playing at a sampling rate of 44,000 times per second. If he played a three-minute song, then the computer would record 7,920,000 samples of that song in even intervals. When the samples are played back to the listener, the listener is able to perceive each of those little samples as being one long continuous sound, even though it actually is not.

Like image files, audio files can become quite large quite rapidly. Even if only one bit was needed for each sample, this one song would be a little less than 8 megabytes worth of information. That's around 78 times the amount of information needed to store the digital version of this chapter, and in reality, much more data is needed to store the information for each sample, since you need more than one bit per sample.

Audio files don't have to be huge, though. Like image files, audio files can be compressed to save space. The software used for this is known as a *codec*, which stands for Compressor-Decompressor (Lake and Bean, 2008). A lot of terms used for compressing visual data also apply to audio data—for example, the terms *lossy* or *lossless*. Remember, lossy means that some data is lost during compression, and lossless means that no data is lost and the file can be restored to exactly how it was originally. Typically, lossy methods are more efficient and can compress files to much smaller sizes than lossless methods, but lossless methods are usually superior from an archiving standpoint.

CARLI, or the Consortium of Academic and Research Libraries in Illinois, recommends a sampling rate of 44 kHz at the minimum, and 96 kHz for archival-quality recording (CARLI, 2017a). This is a very high rate that will record a lot of data. You will need to determine if a higher sampling rate is advantageous to your archive or not. This is also only applicable for digitizing music from analog. If you are storing a recording that is already digital, you'll essentially be storing whatever rate the creator of the file used.

Though these terms are relevant to storing nearly any kind of data, several terms are relevant to audio data in particular.

TERMS FOR AUDIO DATA

Bit depth	The number of bits used for a recording
Channel	Whether the recording was in mono sound (one microphone recording) or stereo (two or more microphones recording)
Ripping	Transferring audio data from a CD (a digital data storage format) to a computer (can also apply to video data)

Bit Depth

Along with the sampling rate, recordings have a certain *bit depth*. This refers to how many bits are used to store the data for each sample taken during a recording. A typical sample will use between 8 and 24 bits. A bigger sample size contains more information about the recorded sound and creates a more accurate recording that is more true to reality. However, a bigger sample size also requires more space to store the data. With your recording of the violinist, the 44,000 sampling rate with a small 8-bit sampling size would create a 63,360,000-bit file (before any compression), and a more robust 16-bit sampling size (16 bits is standard for the recording industry) would create a 126,720,000-bit file. CARLI recommends a recording bit depth of 24 bits, which will provide a lot of data (CARLI, 2017a). While you always want to be as accurate as possible, the amount of storage space you'll need to store your collection is something you should take into consideration for practical reasons.

Channels

Audio data can also be recorded in channels—either mono or stereo. *Mono* recordings capture a single sound using one source (there's typically only one microphone involved). *Stereo* recordings are more complicated and use at least two microphones during recording to capture sound from two different directions.

Why would you want to do this? Well, people have two ears and hear audio information coming from different directions in the real world. Recording audio data in stereo gives a sense of realism to the recording and makes it more true to life. When you listen to a recording in stereo using headphones, for example, the exact sounds you hear may be different on each side of the headphones, making it seem to your brain as though you're actually there, with the original source of the sound, such as a concert, opera, or a comedy monologue. To hear this effect on a computer, you need at least two speakers.

It's important to note that stereo recordings require twice as many bit captures as mono recordings, and thus are much larger. Professionally produced CDs and DVDs typically have stereo sound. However, mono sound still has a number of applications, and older recordings, such as cassette tapes and vinyl records, may be in mono sound.

Ripping

Another term that might be of importance to you is *ripping*. Ripping is the term used for copying audio data from a CD onto a computer hard drive. There are some legal ramifications involved in this if the CD in question is under copyright. Ripping a CD that contains copyrighted information onto a computer is not necessarily illegal, but doing so with the intent to distribute the information is. However, you may have an agreement with the copyright holder in regard to archiving the data, and creating a copy of audio data for the purposes of archiving only is usually perfectly legal (Lake and Bean, 2008). You'll learn more about CD technology in chapter 7 and more about copyright law in chapter 14.

Audio Data Formats for Archiving

Audio data hasn't been in common use for computers as long as image or text data has. While sound capabilities are standard on modern computers, in the past, sound was an optional feature. Sound only began becoming a standard feature in the mid- to late 1980s.

Sound cards are not required for computer function. Very early computers had no sound cards whatsoever and created simple beeps using internal speakers. In essence, the function of a sound card is to translate audio data into a sound that can be heard using speakers or headphones, or to do the reverse and translate sound from a microphone to audio data (Lacoma, 2018).

File formats for audio data are not as well established or well known by general users as the other file formats covered so far in this book. Like other types of data, there are a wide variety of audio recording formats that you can use. However, many of these formats are proprietary or otherwise require specific software to run, so choosing one that suits your needs can be more complicated than choosing one for visual data. The software that you use may only play a limited number of types of audio data or may not be able to convert formats easily because of these proprietary formats. This section will address a few of the many possibilities for storing audio data.

SOME AUDIO CODECS

AAC (Advanced Audio Coding)

AIFF (Audio Interchange File Format)

ALAC (Apple Lossless Codec)

AMR (Adaptive Multi-Rate)

FLAC (Free Lossless Audio Codec)

MP3 (MPEG-2 Audio Layer 3) Files, MPEG-4

MIDI (Musical Instrument Digital Interface) Files

OGG (Ogg Vorbis)

Opus

WAVE (WAV) Files or BWF (Broadcast Wave Format) Files

WMA (Windows Media Audio)

WAVE

The WAVE format was originally created by Microsoft and IBM, who worked jointly on the development of the WAVE format (Costello et al., 2012). WAVE files are a

commonly used file type for sound that can be suitable for archiving purposes, and WAVE files can be used for storing both mono and stereo sound.

The WAVE (also known as a WAV) file is the basis for some of the Library of Congress's preferred preservation formats (Library of Congress, 2018). The preferred type of file is a PCM WAVE, which means that the file type is a WAVE file, and the audio is encoded using pulse-code modulation or linear pulse-code modulation (PCM and LPCM, respectively), which are standard ways of representing audio in a digital format.

WAVE files are created in units of data called *chunks*. Each chunk is a set of data that can either represent metadata (information about the file and its contents) or actual audio information. A variation of this format, the Broadcast Wave Format (BFW), contains metadata specifically for facilitating the exchange of the file between broadcasters, and also contains timestamp information (Library of Congress, 2017a). Metadata will be discussed further in chapter 13.

AIFF

The Apple company's AIFF file is very similar to the WAVE, and the two formats are virtually interchangeable as far as sound capability goes. Most sound-editing programs will work with both file types (Costello et al., 2012). Though not ideal in that they are proprietary formats, both formats are so commonly used that finding compatible software is not an issue.

MP3 Files

A format that may be more familiar to you is the MP3 format. MP3 is short for *MPEG-1 audio layer 3*. MPEG is an acronym for the Moving Pictures Experts Group, which is a group that develops standards for video compression. There are also MPEG-2 and MPEG-4 formats, which are used primarily for video data; MPEG-4 can also be used for archiving audio data. MPEG-3 was never fully developed and is not in use.

In 2017, key patents for the MP3 were not renewed, which may have an effect on the development of future software; this could mean that more software will incorporate this format, as it can be more freely used now by developers, or that the recommended replacement for the MP3, the AAC, overtakes it in popularity due to its advantages in sound quality and file size (Fox, 2017).

The MP3 format is a highly desirable format primarily owing to how well it compresses audio data. What happens in an MP3 file is that the file removes or reduces sounds that humans can't normally hear, cutting out irrelevant information, in a way (Andrews, 2008). This greatly reduces the amount of data needed to reproduce the file, thus requiring less storage space.

The MP3 format has a lot of appeal to the general public for several reasons. These are the files that were initially used with portable media players, often just referred to as "MP3 players." The high level of compression makes it possible to store a lot of data on one of these players; the format can also be used by a variety of other devices and software, a very good feature to have in digital archiving. It's also popular because it makes downloading audio data from the Internet easier; the larger a file is, the longer it takes

to download, and so smaller files are better from this perspective because they take less time to acquire.

MP3s may also appeal to patrons if you want your collection to be available online, since chances are good that they will already have the necessary software to play your audio files and find this format to be comfortable and familiar. However, the MP3 format also uses a lossy method of data compression, which is always less preferable for archiving than lossless methods. While, for this reason, you may not want an MP3 to be your main archiving format, MP3s are very good for sharing, and so you may want to store multiple copies of a file—one in the WAV format, for instance, and one as an MP3.

AAC

Advanced Audio Coding, or AAC, was designed to replace the MP3 and uses some of the same methods of reducing file size by eliminating audio data that humans can't normally hear. However, it has a better sound quality. Like the MP3, this is a lossy format, and it is also proprietary, and so is not a preferable choice for archiving. However, it has the same merits as an MP3 file for sharing data online. It is also popular as the audio component of MPEG-4 videos and is the audio format of choice for the online video sharing service YouTube (Fox, 2017).

FLAC

The Free Lossless Audio Codec, or FLAC, is, as its name would imply, a lossless format. It is otherwise similar to the MP3 and is notable in several ways. Because it is a lossless format, it has a higher quality to the sound, but can transmit quickly over the Internet like an MP3. What is particularly interesting about this format, however, is that it is nonproprietary. However, adoption of this format has been slower than some of the other formats for a variety of reasons, and generally additional software is required to convert a file to this format (Pendlebury, 2018). Still, it is a notable format that may be of use to an archive.

MIDI Files

Musical Instrument Digital Interface, or MIDI, files digitally simulate the sounds of real instruments, like drums or violins. These files are used to create synthesized music. Rather than recording a performance of these sounds, MIDI files play the appropriate notes with the appropriate instrument effect. You can almost think of it as being digital sheet music that can play itself. Rather than being a recording of an actual performance, a MIDI file is instead a set of instructions for how a synthesizer can play a piece of music or a sound. In the past, this was highly advantageous, because MIDI files are just instructions and these files are much smaller than actual recordings of performances. Because they use so little data, MIDI files were often used for things like background music and sound effects in old computer games. MIDI files are still in use today, and can be used by musicians for composing music and other creative applications.

When you play an MP3, AIFF, or WAV file, you're playing an exact reproduction of a recording. MIDI files are instructions. While this saves a lot of space, it can be problematic in that MIDI files are somewhat dependent upon the quality of the software that is using these instructions to reproduce the sound (Lendino, 2013). For instance, imagine

the difference between giving a copy of the sheet music for Beethoven's Fifth Symphony to a middle school band and giving the same music to a professional symphony orchestra. The instructions are exactly the same, but the quality of the results is quite different. MIDI files are not as commonly used today as in the past, but still have some use and may be something you want or need to archive. Again, MIDI files were common in things like older computer games, and so you may need to save MIDI files if you want to preserve old software as part of your project.

In many ways, storing audio data is similar to storing data for images. Video data is also similar in many ways to both of these types of data, since it requires both kinds of data—image and audio—to create the video.

Video Data

The history of film is a little less clear-cut than some of the other media discussed so far, as a lot of different people were involved in its initial development. However, a notable event in the history is the development of the kinetoscope by the Edison Company in 1891. This device allowed a person to view a brief movie, one viewer at a time. Movies that could be shown to an audience were developed soon afterward, in 1895, by the Lumière brothers (National Science and Media Museum, 2011).

The history of *digital* cinematography, however, has been much, much shorter. Digital video cameras became available in the 1980s, and an early film recorded using this technology was *Julia and Julia*, an Italian film made in 1987. Although it was shot digitally, it was transferred to 35 mm film afterward. One of the early films to be shot in a “high-definition” digital was *Once Upon a Time in Mexico*, made in 2001 (Daniele, n.d.).

The move from film to digital has actually been remarkably slow and filled with quite a bit of controversy. Digital and film formats each have a different look and feel, and one might appeal more to a director than the other. In addition, digital films are also much more prone to obsolescence due to equipment problems and are more difficult to upgrade to a more advanced format than films shot using traditional methods (Daniele, n.d.).

Explaining exactly how video data is stored is a little complicated. The general principle behind it, however, is simple, and an easy way to explain it is to consider a digital video as having the same principles as an animation. As you probably know, traditional, hand-drawn animation is created when an artist draws the same picture over and over with incremental degrees of change between the drawings. When the images are shown together, one following after another in rapid succession, it creates the illusion of movement, as if all those still images were moving. You may notice that this is also similar to how audio data recorded by a computer works—by playing tiny samples of sound in rapid succession.

Digital videos essentially work the same way as animation and sound recordings. They are composed of a number of still images that, when shown rapidly one after another, create the appearance of movement. If you've ever seen a reel of video film or the inside of a VHS tape cassette, you'll notice that the film for the video is actually composed of a number of images, one after another on the tape. These are shown rapidly at a steady pace to create the look of movement. The principle used for film-based videos is exactly the same in digital format, it's just that the digital video shows digital images made from pixels rather than ones made from film.

TERMS FOR VIDEO DATA

Rendering	Converting a video file from a native format to another format
Timecode	A method of keeping track of video data, often used for synchronizing audio data
Video codec	Software used to compress video data

Though there is some variation, a video typically has 30 frames per second. This means that for every second that goes by in a video, you're actually seeing 30 still images shown one after another, which are going by too fast for you to even notice. For software that creates video data, it's important to keep track of all these frames. One of the ways in which this is done is by using a *timecode*, which is a method of uniquely identifying frames and determining when they will display. A timecode in video editing also helps the creator of a video keep track of events, such as when specific audio data needs to start playing.

Just like audio data, video data can be compressed using software known as a *video codec*. You'll need to choose a codec that you find acceptable for your video data. Lossless codecs are preferable, but most video codecs are lossy. Though there are some codecs that have the same name as a file format, they are not necessarily the same thing. For instance, you can use an MPEG-2 *codec* to compress a file that is not in the MPEG-2 format. Some of the lossless options are Huffiyuv, Lagarith (similar to Huffiyuv), and JPEG 2000. You can also choose to simply not compress the file, but this will result in an enormous file for storage (CARLI, 2017b). You can use lossy methods, but you should be aware that these are not the best choice for archiving.

Video files are typically created in a native format, similar to a native format for an image, as discussed in chapter 3. Remember, a native format is a file format that works specifically with a particular device or a particular type of software. Video native formats have the same problems as far as archiving goes as native formats for visual data do in that they normally can't be opened by anything other than the software program that created the file. The video file therefore needs to be converted into a different file format, a process known as *rendering* for video data. There are quite a few data formats for videos, and like audio file formats, you'll face a challenge in that most of them are proprietary and specific to a company.

Archiving Video

When you archive a born-digital video, you'll essentially be using the formats and specifications used by whoever created the video, just as you would for audio data (though you may need to convert the file format to something more acceptable). If you're digitizing information, however, you'll need to make some decisions.

One of these is about the *resolution* of the video image. Resolution is measured by the number of horizontal pixels by the number of vertical pixels (pixels were covered in the beginning of chapter 3). More is better, up to a point. A standard television today

has a 640×480 resolution, while a high-definition television can have a resolution up to 1920×1080 pixels (CARLI, 2017b).

Similarly, the *aspect ratio* is the width of a video image divided by the height. For example, the aspect ratio of standard cathode-ray tube televisions (before the rise of flat-screen televisions) was 4:3, meaning that video for these televisions has images that are slightly wider than they are tall. At some time, you may have seen a notice when watching a movie that says something along these lines: “This film has been formatted to fit your screen.” This is because films intended for movie theaters have a different aspect ratio than television sets, and some of the image needs to be cropped in order to be more easily viewed; this is a loss of information. This is also what “widescreen” versus “fullscreen” means when you purchase a copy of a movie; widescreen means that the aspect ratio is true to the original and has not been cropped. Your video data will appear best at the aspect ratio at which it was originally captured.

As you decided with audio data, you’ll need to decide how large your samples will be when you capture video data. A larger sample size, or bit depth, provides more image data and is therefore superior from an archiving perspective. This does, of course, create a larger file size. CARLI (2017b) recommends a sample size of 30 bits.

Video data can also use either interlaced or progressive scanning. With *interlaced* scanning, each image essentially captures half of the relevant data, with one image displaying information on odd-numbered lines and another on even-numbered lines. Analog video, such as a VHS tape, is usually made in this way. When digitizing film that was created in this method, sometimes the frames become blurred from small amounts of movement between one frame and the next. *Progressive scanning* captures all data in a single frame. This is considered the superior method as far as archiving goes, and most born-digital video uses this method (CARLI, 2017b).

© Video Data Formats for Archiving

Like formats for audio files, there are many formats available for video files. Most of these are not particularly suited for archiving, since they are proprietary. However, you may want to know what kinds of video formats are available anyway, so that you know what types of formats you have and what type might be suitable to display to your patrons (for example, which formats are good for use online). This section will cover some possible file types; however, there are many, many ways to store video data.

COMMON VIDEO FILE FORMATS

AVI	Audio Video Interleaved
DPX	Digital Moving-Picture Exchange
FLA	Flash
MJ2	Motion JPEG 2000
MOV	QuickTime Movie
MPEG	File format created by the Moving Pictures Expert Group
SWF	Shockwave Flash or Small Web Format
WMV	Windows Media Video

MPEG

As you learned earlier in this chapter, MPEG stands for the Moving Pictures Expert Group, a group that develops standards. There are several variations on the MPEG file format, with MPEG-1 typically being used for audio data (the MP3 format). MPEG-3 is not in use; MPEG-4 is usually used for transmitting video data online; and the MPEG-2 format is for general use with video data.

As MPEG-1 is capable of creating highly compressed audio files, so too is the MPEG-2 format capable of creating highly compressed video files. With this file format, frames that have no appreciable amount of change from one frame to the next are discarded (Andrews, 2008). Remember, there are 30 frames per second of video in a normal capture, and so there are instances in which some of those 30 frames have no new meaningful information. Getting rid of some of those frames helps compress the file, and using frames with no new data for this means that the viewer won't notice the difference. However, this is a lossy method of data storage, and while you might not readily notice the difference, it's still preferable to use a lossless method of data storage. Like MP3s, though, you might want to use one of these types of formats for your patrons, and use a different format for actual storage.

Motion JPEG 2000

Motion JPEG 2000 is an open international standard with a lossless codec, so no data is lost during compression. Rather than compressing the file by removing frames, as the MPEG formats do, each frame is individually compressed (Pearson and Gill, 2005). This method is different from other methods of video compression in general, and means that very little information is lost during the compression process, making it more appealing from an archiving standpoint than the MPEG formats, for instance. This format is commonly used for digital cinema, or movies shown digitally in a movie theater (*PCMag*, 2013).

Flash

Flash files are particularly desirable for transferring video information online (Fuller and Larson, 2008). Flash files have either the extension FLA or FLV. An FLA file is the file used to create a video, and an FLV file is the final product that can be shared with others. Flash animations are often used online or as components of web pages. This technology was originally developed by Macromedia, but the company was purchased by Adobe Systems in 2005. One of the unusual and appealing features of Flash animations is the fact that they typically use vector images rather than bitmapped or raster images for the frames (Library of Congress, 2017b). Vector images are scalable, as discussed in chapter 3, which means that they look the same regardless of the size of the image, whereas bitmapped or raster images can become pixelated when resized. Shockwave Flash, or SWF, is a related format with similar features; however, it can be used for many things other than videos, such as small games or programs. Both of these formats are particularly useful if you want to display information online or to archive information from the Internet.

MOV, WMV, and AVI

The MOV file extension indicates a QuickTime movie; QuickTime is a format developed by Apple. Similarly, the WMV is a Windows Media Video file, which is a format developed by Microsoft. AVI is an Audio Video Interleaved file, also created by Microsoft. These are all proprietary formats, which makes them less desirable for archiving than some of your other options. However, they are very common and there is a lot of software available to work with these formats, which may make them appealing for sharing your archive's data online or for letting your patrons interact with your archived data. CARLI recommends AVI and QuickTime in particular as alternative options to the more desirable MXF format. Though AVI and QuickTime are less desirable for archiving, they are commonly used by the general public, so you may find it easier to find useful software and resources for these formats (CARLI, 2017b).

Chances are good that you won't be archiving video data, especially not commercially produced video data. However, with cameras so readily available today this may very easily change in the near future. After all, creating and sharing videos online is a common hobby among many people; it can even become a career, in some cases. Preserving this type of data may very well become quite important for preserving information about the present for the future.

Key Points

- Audio information is stored in much the same way as other types of data for a computer, and can be compressed using both lossless or lossy compression.
- The realism and accuracy of audio information is dependent upon the sampling rate and size and whether the recording is in mono or stereo sound.
- Audio information is stored in a variety of formats, but unlike other types of data, audio files tend to be proprietary in nature, adding challenges to archiving.
- Video data is a combination of both audio and visual data, and is stored as a series of still image frames played back to the viewer at a rapid rate.
- Videos can also be compressed using either lossy or lossless methods.
- Videos can be stored using a wide variety of formats. The original format and necessary quality level can be factors that determine what format you choose for your archive.

You've learned quite a bit about how computers store data for several types of data as well as how computers store data in general. This book will cover one more type of data—born-digital data whose original format is less concrete, such as software, websites, and e-mails.

References

- Andrews, Jean. 2006. *A+ Guide to Managing & Maintaining Your PC*. 6th ed. Boston: Course Technology.
- BBC News. 2008. "Oldest Recorded Voices Sing Again." <http://news.bbc.co.uk/2/hi/technology/7318180.stm>.

- CARLI (Consortium of Academic and Research Libraries in Illinois). 2017a. "Guidelines for the Creation of Digital Collections: Digitization Best Practices for Audio." http://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_audio.pdf.
- CARLI. 2017b. "Guidelines for the Creation of Digital Collections: Digitization Best Practices for Moving Images." http://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_video.pdf.
- Coalson, Josh. 2009. "FLAC." Xiph.Org Foundation. <https://xiph.org/flac/>.
- Costello, Vic, Susan A. Youngblood, and Norman E. Youngblood. 2012. *Multimedia Foundations: Core Concepts for Digital Design*. Waltham, MA: Focal Press.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Daniele, Carina. n.d. "Film to Digital: The Growth of Cinema." Computers in Entertainment. Accessed June 6, 2019. <https://cie.acm.org/blog/film-digital-growth-cinema/>.
- Fox, Alexander. 2017. "The Difference between MP3, AAC, FLAC and Other Audio Formats." Make Tech Easier. <https://www.maketecheasier.com/difference-between-mp3-aac-flac-other-audio-formats/>.
- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology Comprehensive*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Lacoma, Tyler. 2018. "What Is a Sound Card?" Digital Trends. <https://www.digitaltrends.com/computing/what-is-a-sound-card/>.
- Lake, Susan, and Karen Bean. 2008. *The Business of Technology: Digital Multimedia*. 2nd ed. Mason, OH: South-Western Cengage Learning.
- Lendino, Jamie. 2013. "At 30, MIDI Is Still Misunderstood." *PCMag*. <http://www.pcmag.com/article2/0,2817,2418880,00.asp>.
- Library of Congress. 2017a. "Broadcast WAVE Audio File Format, Version 1." <https://www.loc.gov/preservation/digital/formats/fdd/fdd000356.shtml>.
- Library of Congress. 2017b. "Macromedia Flash FLA Project File Format." Sustainability of Digital Formats Planning for Library of Congress Collections. <http://www.digitalpreservation.gov/formats/fdd/fdd000132.shtml>.
- Library of Congress. 2018. "Library of Congress Recommended Formats Statement 2018–2019." <https://www.loc.gov/preservation/resources/rfs/RFS%202018-2019.pdf>.
- National Science and Media Museum. 2011. "A Very Short History of Cinema." <https://blog.scienceandmediamuseum.org.uk/very-short-history-of-cinema/>.
- PCMag*. n.d. "Definition of: JPEG 2000." Accessed December 14, 2019. <http://www.pcmag.com/encyclopedia/term/58992/jpeg-2000>.
- Pearson, Glenn, and Michael Gill. 2005. "An Evaluation of Motion JPEG 2000 for Video Archiving." Archiving 2005 Final Program and Proceedings, 237–43. http://archive.nlm.nih.gov/pubs/pearson/MJ2_video_archiving.pdf.
- Pendlebury, Ty. 2018. "What Is FLAC? The High-Def MP3 Explained." CNET. <https://www.cnet.com/news/what-is-flac-the-high-def-mp3-explained/>.



Storing Born-Digital Materials

IN THIS CHAPTER

- ▷ What is software?
- ▷ What is a programming language, and why is it important?
- ▷ What do you need to archive software?
- ▷ What is a database? Why might it be necessary to archive one?
- ▷ How can websites be archived? What are the limitations on archiving websites?
- ▷ How can e-mails be archived?

When you think of a computer, you are most likely thinking of a machine that is capable of a few specific things: it can take an input, calculate a value, communicate that value to the operator, and store that value. While, as discussed earlier in this book, a computer can do many things, these are the essential functions of a computer.

It might surprise you to learn that the first computer program preceded the first operational computer. The development of the first modern computer is often attributed to Charles Babbage, who designed (but did not complete) a device called the Analytical Engine in the 1830s. Owing to a number of issues, this theoretical computer was never built in full. Mathematician Ada Lovelace nevertheless wrote an explanation of how this device could be used to calculate Bernoulli numbers (what Bernoulli numbers are is a little complicated to explain, but they're an important concept in high-level mathematics). This description, written in 1843, is often considered the first computer program. Why this is so will be explained shortly.

Archiving in the past involved caring for and preserving tangible objects. In chapter 3, regarding images, it was even suggested that printing out digital photos was something

to consider to keep information safe. We have historical letters, for instance, from writers, artists, presidents—all sorts of people. These letters tell us important things about these people and how they lived and who they knew.

In current times, however, people are creating information that does not have a true tangible form. These letters tell a lot, too, but they are often in the form of things like instant messages, e-mails, or message boards. Preserving this type of information has an entirely different set of challenges from preserving a tangible object.

Some types of electronic information cannot have a tangible equivalent, as well. Software programs are typically like this; these are instructions that must be read by a computer and that perform a function using that computer. There isn't a good, tangible equivalent to an operation of this type for an archivist to preserve.

This chapter will give some background information to help you learn more about the problems and possibilities of preserving born-digital information, information that is originally digital in nature and may be impossible to store and use in any other way. It should be noted that this might involve a wide variety of materials, but this chapter will address a few common ones: software, databases, websites, and e-mail.

One of the many challenges before you is the fact that there have been so many different types of computers and many different ways of creating software. Why does this matter?

Operating Systems and Applications

What exactly is software? As you learned at the beginning of the chapter, Lovelace's set of instructions for calculating Bernoulli numbers is often considered the world's first software program. You may wonder why this is, as it's just a set of instructions, not a program.

A computer program is really just a set of instructions. In fact, sometimes programs are written out in "pseudocode," which is a set of general instructions that would make sense to a human, before the actual coding is done. This can be helpful for working out exactly what the program will need to do and the order in which it will need to do it before spending the time to work out how the "computer readable" version needs to be written.

In this sense, creating an application is very simple, as it is merely a set of instructions telling a computer what to do—such as adding numbers or displaying characters on a screen. Part of the reason software development becomes complicated is because computers are extremely literal and cannot determine what the purpose of an application is, unlike a human who might be able to figure out the unwritten parts of a set of instructions (e.g., when asked to drive to the store to get groceries, getting into the car first is understood without its needing to be said). A computer cannot execute unwritten instructions, nor can it "know" to *not* execute instructions that are in error. For example, it is possible to create a program that has no way to complete the execution of the instructions, as the instructions have an infinite loop that tell it to repeat a certain part of the instructions endlessly. In general, infinite loops are undesirable and can cause memory issues and other problems with a computer (and are sometimes written on purpose as part of computer malware, or malicious software programs).

Things also get complicated in that there are many, many different programming languages. That is, there are a variety of ways to construct these instructions, and like human spoken languages, each language has its own rules and syntax, its own "vocabulary." Coding languages also have their own strengths and weaknesses and some are designed

for creating specific types of software. As an example of how a program's code looks, the following is a simple program in the coding language "Perl" that prints the text "Hello World!" to a computer screen (a program like this is often a new programmer's first application, as a tradition):

```
#!/usr/bin/perl  
  
use strict;  
  
use warnings;  
  
print"Hello World!\n";
```

There is something curious about this sample code, though. As you know, computers only deal with numbers. They are not capable of understanding words, and although you might not quite understand how this program works, you certainly understand most of the words that make it up.

Programming languages actually exist for the convenience of the programmer. It's exceedingly difficult to write a program solely in number values, and few people are capable of such a complex and tedious task. Once a program is written, it needs to be *compiled*. What this means is that the program is translated from programming language, which a human can understand, to *machine language*, which a computer can interpret.

The machine language specifies what a computer should do using numbers. As a very simple example, when you sit at a keyboard and press the letter *A*, a program is required to interpret that you have pressed a key and that key is the letter *A*. Depending upon the software program, this may have a variety of functions. For a simple word-processing program, this is an instruction to display the letter *A* on the monitor and to put the letter *A* among the information required for the document file you are in the process of creating.

Another important part of this process is the operating system. "Operating system" is a term that you are almost certainly familiar with, and chances are also excellent that you can name a few operating systems, such as Windows, Linux, macOS, iOS, or Android, to name a few standard and mobile-oriented operating systems.

You have to have an operating system to operate a modern computer. But what does it do, exactly? Well, it's actually much like what the name implies: an operating system is software that coordinates the operations of the computer. That is, it coordinates the instructions from any applications that need to be run with the actual hardware of the computer. This can involve a wide variety of different functions, such as coordinating temporary data storage to RAM chips or writing data to more long-term storage, like a hard drive.

Why is all of this important to archiving? Well, if you want to archive software, this can be a rather complicated project. Remember, a computer is needed to execute the commands in a software program. To run a program, you may need a specific type of computer and a compatible operating system. Because an operating system is required for modern software applications (it was not in very early computers), you need a copy of the operating system intended to work with the instructions of the software program for you to actually run it.

Another complication can arise if you are storing compiled code as opposed to source code. Compiled code, again, is code that has been "translated" for a machine to interpret. *Source code* is the original code used by a programmer to design a software program, still

in a human-interpretable language. Commercial software is almost always sold in the compiled format, and terms of service typically prohibit trying to discover, based on the compiled code, what the source code was.

In contrast, you may also come across something known as *open-source code*. The term *open-source* can refer to software that is distributed with the source code accessible to the user. Software like this is typically noncommercial and the developers often invite users to contribute their own code in order to improve the software.

To thoroughly archive software, you may need a variety of files. The following are some of the things the Library of Congress recommends having for archiving software.

Source Code

Having the software's source code is ideal whenever possible. Again, this is the original instructions for a software program, in a human-understandable programming language. Why is this important? If you have the source code and someone who understands the language (some programming languages are no longer in use), then it's possible to determine what the software did and how it worked. It's potentially possible to "translate" the software into another language if the original source code is known. The Library of Congress also recommends that the compiler used for the software is noted in the metadata (metadata will be covered later in this book), or even that the compiler itself, if it is unique, be stored with the software. Remember, computers need programming language "translated" so that it is understandable by a machine. So, the compiler is necessary for the computer to be able to run source code (Library of Congress, 2018b).

Consider your goals when deciding what to store. It may be desirable to preserve exactly how a software program was written, or just being able to preserve the function of the program might be good enough, depending upon what you want to do. As an interesting example, the Internet Archive has a malware collection, which allows users to experience malware programs created in the 1980s and 1990s, but does not actually run the malware aspect of these programs.

Operating System

The operating system is also required (Library of Congress, 2018b). Again, communicating between the program instructions and the physical hardware of a computer, the operating system is needed to actually execute the instructions of a software program. This might be simple or complex. There are only a few operating systems in use currently (relatively speaking), but this has not been the case at all in the past. There have been many operating systems in the past, some of which were common and popular and some of which were not. Operating systems can also be designed for different types of computers—that is, an operating system intended for a personal computer will be different from that designed for a server computer. It should also be noted that some people design their own operating systems as a hobby. (Although it would be unlikely that your archive would store an operating system this unique, it might be possible if the developer eventually becomes someone of historical importance, develops something unique from the code, etc.).

The *version* of an operating system may also be important. Operating systems, especially since the rise of the World Wide Web, need continual updating to match both cur-

rent advancements in computer technology as well as advancements made by malicious users. That is, people who want to use technology for harmful activities are continually finding new ways to bypass security or to exploit the way that code is written in order to conduct these activities. The software for operating a computer needs to change in order to thwart this type of activity. What this means to you is that a program designed for one version of an operating system may not work well on another due to changes in that system's coding.

It's traditional for software programs and operating systems to designate their latest version with a number—for example, Version 2.0, Version 3.1, and so forth. The higher the version number, the more recent it is. Some companies additionally have a name that goes with the number. For example, the Android operating system, used for mobile devices, traditionally uses the names of sweets to designate a version, such as Jelly Bean, Lollipop, and Marshmallow. MacOS often uses the names of animals. This system can make it difficult to determine what the latest version is, although it is probably much more entertaining to the end users, as well as the developers.

Platform

The platform is also required to operate software. A software platform is a complicated concept, as it can mean quite a few different things. In this case, it basically means whatever it is that you require to run the software. For example, if you have a web-based application, then you're also going to need the browser program, like Chrome or Firefox, that can run it. A platform could be the operating system, or it could be other programs running on that system. Things that are not a software program but are required to run it could potentially be considered the platform.

An emulator is a software program that is designed to emulate another. That is, if the original platform is unavailable, you can run an emulator program to have a computer behave like a different computer. As an example, the Windows operating system MS-DOS is no longer in use, but it was popular for a long time, and so there are many programs designed for this operating system. The program DOSBox is an emulator that can be installed on a modern computer and is designed to behave like the operating system MS-DOS. This emulator can run programs designed for the MS-DOS operating system.

Along with all of the programs and possibly hardware that are required for archiving software, copyright is another consideration to make. Commercial software is typically under copyright, and so the information cannot be made publicly available. However, archiving commercial software may be quite important. As an example, in 2004, the FDA needed to track down the recipients of an improperly processed batch of botulin (used to make Botox). This information was contained in a file created using a type of business software. However, the file was created using the 2003 version of the software—it was no longer readable with the 2004 version. The software manufacturer was not able to provide a copy of the 2003 version. Eventually, this file was opened using a copy of the older software archived by the National Software Reference Library (Owens, 2012).

Software is often used to process data of some type, and this is something else that you may need to archive: pure information.

“Database” is one of those words that get used a lot without a lot of explanation. You might hear about accessing databases in a movie, or hear a story on the news about hackers getting into a database. Obviously, it’s a way to store and retrieve data, but what exactly is a database?

Suppose that you wanted to store temperatures readings recorded in your town or city. You could think of each temperature recorded as a singular point of data. You then decide to collect your temperature readings, and you organize your readings by, say, date. The collection of readings is now a *data set*. Now, suppose that you and another library in another town are doing the same thing, and you combine your data. Now you have two sets of temperature information with dates and locations. If you have software known as a *database management system*, or DBMS, you can do all kinds of useful things with your data sets—such as search for temperatures by date or town or sort the temperatures in meaningful ways (like low to high, or viewing the highest temperature recorded). What you have created with your multiple related data sets is a *database* (USGS, n.d.).

A database can be very simplistic or quite complex. For example, it’s possible to create a simple (but useful) database using Microsoft Excel’s spreadsheet software (and other types of spreadsheet software). The table format of spreadsheet software makes it easy to view and understand a dataset, and the software can be used to relate multiple datasets together so that data can be organized and retrieved in meaningful ways.

When interpreting a database, a computer needs a way to separate each unit of information from all the others. In the example of a temperature database, for example, the computer needs to know that the temperature collected on July 2 is separate information from the temperature collected on July 3. There are a variety of ways to do this, and so there are many file formats that can be used. Chapter 4 covered one of the possibilities: a CSV file. A CSV, or comma separated value, file separates each point of data with a comma. So, your temperature set might look like this: “July, 2, 92, July, 3, 93,” and so on. The software used to interpret the data can determine the month, day, and temperature from this information, because each time a comma appears in the data, it indicates to the software that this is new information. There are other ways to do this, too, such as using XML tags. XML files were discussed to some extent in chapter 4 and will be explored further in chapter 13. As a brief explanation, however, XML uses tags, or words, that both separate information and indicate what the information is. In the example, XML tags could be used to flag the months, dates, and temperatures, which can be helpful. For instance, the following is an example of how XML tags might flag the different types of information:

```
<month>July</month>  
<date>2</date>  
<temp>93</temp>
```

In a way, databases operate invisibly. When you interact with information online, for instance, you may be unaware or not think about the fact that the data that you provide for a website is stored in a database. For example, if you sign up for an online account in order to shop at a website, your name, address, and information like credit cards and

phone numbers are all stored in at least one database. This makes it possible to retrieve the data at a later point when you sign in to your account again.

Storing the information contained within a database might be something quite important for an archive. However, because they operate in the background but are integral to the function of many software programs, databases also offer another complicating factor in an endeavor to store information. If you want to preserve the data of a website, this can be problematic, as you won't get the information a website stores without its corresponding database or databases.

Websites

In modern times, there is a peculiar conundrum. The invention of the World Wide Web has led to the decline of printed media, such as newspapers and magazines. It has also led to an unprecedented explosion of information. Anyone can contribute information for others to view online using a wide variety of means.

This means that a lot of information that may be relevant in the future is only available through the Internet using websites to share that information. In chapter 4, you learned a little about how a website works. It's written as plain text, but with code to designate paragraphs, headings, divisions, tables, and so on. The code used for websites is HTML. There is also a code for controlling the style or the "look" of a website, CSS. If a site is more complex, it may also require image files for pictures, video files, files for custom fonts, and more.

These are the requirements for a simple website. However, many websites allow users to interact with the site—for example, users can make purchases, add comments, and so on. If a website is more complex and/or allows for user interaction, software program files are also required. The software code for websites is typically written in a scripting language such as JavaScript or PHP; scripting languages are a special type of programming language that does not need a compiler because it is going to be interpreted by another program. In this case, the browser program is what is used to interpret the instructions of the code. Again, a browser program is a program used to retrieve and view websites, and it is what interprets instructions on a website and shows the site to the user.

What this ultimately means is that a website can actually require a huge number and variety of files, depending upon the purpose of the site and how large it is. To properly work, a website might also have databases, which are organized sets of data, such as names, addresses, and so on.

It's possible to create an archived version of a site using a software program that works a lot like a web crawler. A web crawler is a program designed to search through websites, primarily to determine what keywords pertain to the site so that it can be retrieved using a search engine and to search for more sites using any links on the site. Likewise, a specialized program can be used to "crawl" through websites and store the code about the site.

One of the potential problems with archiving websites is the fact that HTML and CSS are evolving languages. The latest versions of these languages are HTML5 and CSS3. These updated versions are capable of much more than their previous versions, and some aspects of these languages have become obsolete over time—that is, some instructions are no longer used because there is a better way of doing things now, or because of potential issues or lack of support from more recent browsers. As an example, in the

past, a program was required for animations on a site. Now, it is possible to code simple animations using CSS3 code, which is easier and less open to exploitation by malicious users than using a program.

Likewise, browser programs update over time, for much the same reasons as operating systems must update: to keep up with innovations and to thwart people trying to get around security. What this means to you is that websites created in the past might not display correctly on new browsers, and websites created now might not display correctly in old browsers. Even different current browsers might display a website differently; a developer might need to put extra code into a site just to make sure that it displays similarly from one browser to another. It may be necessary to preserve old versions of browsers or to find ways to emulate old browsers.

Sometimes developers make a note on websites that the site is best used in a certain browser. This means that the functions and the display of a site will work and look best in the noted browser. You can use this as a cue to determine how the site should best be preserved and displayed. For an old website, copyright dates (when available) can be used to guess which browsers might work best.

At the introduction of this section, it was noted that databases can make the preservation of websites more complex. This is true for many websites that are intended for users to participate in. For example, if you do things like make a post on a social networking site or indicate that you liked a post, that information can be stored in a database. If a website requires a database, any archived version of that site will be only partially complete and functional without the corresponding database. It is possible to take screenshots (images of a site on a user's screen) and thus store things like user posts, but this is problematic in some ways. For example, a screenshot is going to be more difficult to make searchable than text, as text lends itself easy to keyword searches, as explained in previous chapters. To truly archive a website, it may be necessary to get the cooperation of its owner.

Some websites never change, some change occasionally, and some change daily. An important consideration is when and how often you want to store a copy of a site. You may want to track changes over time, or you may be trying to preserve specific information on a site.

As mentioned before, things like websites are how modern society is creating and sharing data; they are the way history is being made at the moment. Another and even more prevalent method of creating modern history, however, is e-mail.

E-mail

E-mail has actually been around for a pretty long time now. People have been using computers to send messages to one another for about as long as computers have been able to network with one another—since around the early 1960s. Messaging capabilities were developed alongside the early Internet.

In some ways, e-mail works a lot like websites do. That is, a website has an address that your computer (the client) can use to request information from another computer (the server). An e-mail has an address, too. Suppose that you are writing a message to your friend Jane Smith. Her e-mail address is `jsmith@genericemail.com`. The second part,

“genericmail.com,” indicates which server needs to receive this e-mail. In this case, it is the server for “genericmail.” The first part indicates who you want to receive it on that server. In this case, it is Jane Smith, who has decided that her address will be jsmith. This is roughly how the software you use to send your e-mail will know where that e-mail file needs to go and how the software on the server will know who should read it. There is a little more to it than this, but what is more likely to be important to you is the content of an e-mail itself, not how it got to where it was going.

E-mail is, in essence, a type of text file. Typically, though, you read the message on a specialized piece of software known as an e-mail client. Much like how a website needs a browser to be easily interpreted and experienced as intended, so an e-mail needs a client. It should be noted that there are e-mail clients designed to be run locally on a computer as well as web-based e-mail clients. Local ones are typically easier to get e-mail files from; it can be a challenge to get the files from a web-based client, as some do not allow the files to be downloaded.

There are actually a number of types of files for e-mail, including some that are specific to certain types of e-mail clients, which is not ideal for archiving. For instance, the EMLX file type is intended for use with Apple Mail, although it is possible to use other programs to open this kind of file. Like many other file types, there are also e-mail file types that are now obsolete.

A popular type of file in 2020 is the EML file format. This is a type used by the e-mail client Microsoft Outlook Express, but can be opened by a variety of other e-mail clients, which makes this file type rather useful. An interesting and useful feature of this file type is that, if you change the EML extension to TXT or HTML (you can do this by choosing to view the file extensions and renaming the file), you can easily open and read the e-mail in an ordinary text editor, like any other type of text file. It’s possible to do this by simply choosing to open the file using a text editor rather than the e-mail client, too (Library of Congress, 2018a).

The MBOX format is also popular. This is a little harder to define than an EML file type. MBOX can refer to several file formats that are related to one another in how they were developed: MBOXO, MBOXRD, MBOXCL, and MBOXCL2. Again, they are all related to one another and are essentially just plain text files, like an EML file. Something unusual about MBOX type files is that they can store the entire contents of a folder from an e-mail client rather than one message at a time; new messages are simply separated by a line. A complicating factor for this file type, however, is the fact that, while they are related, the four MBOX file types are not compatible with one another—that is, software that works for one type will not necessarily work for another (Library of Congress, 2019).

Key Points

- Fully archiving software may also require storing equipment, compatible operating systems, or appropriate emulators.
- Source code can be read by a person, and it is ideal to store source code when archiving software. Source code requires a compiler, however.
- Databases are files with plain text data that has been stored in such a way that it can be searched in a meaningful way by a computer.
- Storing websites is possible using software that works similarly to a web crawler as used by search engines. Websites are complex to store due to a number of factors.

- E-mail is a common method of online communication that has existed for decades. E-mails can be stored as plain text using a number of different file types.

The past few chapters have covered a number of types of information that you may want to archive. Next, you will need to learn about how data is stored so that you will be able to make decisions about how to best store information for your archive. In the next chapter, you will learn about floppy disks, an older method of data storage that you may encounter if you need to archive materials that are born digital or have already been digitized. You will also learn about optical media, which includes CDs, DVDs, and Blu-ray disks. This is another older method of data storage that you may need to work with, and may still have some merits for your project.

References

- Library of Congress. 2018a. "Email (Electronic Mail Format)." <https://www.loc.gov/preservation/digital/formats/fdd/fdd000388.shtml>.
- Library of Congress. 2018b. "Library of Congress Recommended Formats Statement 2018–2019." <https://www.loc.gov/preservation/resources/rfs/RFS%202018-2019.pdf>.
- Library of Congress. 2019. "MBOX Email Format." <https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml>.
- Owens, Trevor. 2012. "Life-Saving: The National Software Reference Library." Preserving.exe. http://www.digitalpreservation.gov/multimedia/documents/PreservingEXE_report_final101813.pdf.
- USGS. n.d. "What Are the Differences between Data, a Dataset, and a Database?" Accessed November 11, 2019. https://www.usgs.gov/faqs/what-are-differences-between-data-a-dataset-and-a-database?qt-news_science_products=0#qt-news_science_products.



Floppy Disks and Optical Media

IN THIS CHAPTER

- ▷ What is a floppy disk?
- ▷ How does a computer encode data on a floppy disk?
- ▷ How might an archive need to handle floppy disks?
- ▷ What is an optical disk?
- ▷ How does a computer encode data on an optical disk?
- ▷ How do CDs, DVDs, and Blu-ray disks differ from one another?
- ▷ How should optical disks be stored?

One of the primary features of a computer is the ability to execute calculations. An additional, desirable feature, however, is the ability to store the results of those calculations. Early computers had a variety of options for how this could be done. Punch cards, for instance, were stiff paper cards with the binary encoding stored as holes or solid spots in the card. Paper tape had a similar principle, but had the holes on a spool of paper that could be wound and unwound. Magnetic tape was another possibility; this is actually still in use and will be discussed in the next chapter. These were fine for that time, because computers needed an expert to run them, and so it was less important if a method of data storage was a bit cumbersome.

In the 1970s, however, computers designed for home use started becoming available. A normal user needed a more convenient way of storing data and loading new programs onto a computer. Early solutions involved the well-established paper tape as a storage method as well as cassette tapes (a convenient version of magnetic tape) (IBM, n.d.).

The solution that took off, however, was the humble floppy disk. A simple square with a readable/writable medium inside, floppy disks were simple, cheap, and easy to use. The technology persisted for decades, and during that time, three different common sizes emerged: 8-inch, 5.25-inch, and 3.5-inch floppies.

If you search for information about floppy disks online, you are very likely to find a lot of articles declaring the floppy “dead.” After all, a typical 3.5-inch floppy held a mere 1.44 MB of data, or 1.44 million bytes. This might be enough to store a couple of small photographs. Compare this to modern technology, which more typically offers data storage in gigabytes or even terabytes, or billions and trillions of bytes, respectively. Modern storage devices can hold thousands of large photographs or videos, which are large and complex data files.

So, why is this chapter discussing floppy disks at all? After all, it’s obsolete technology. Laughable, even, if you were to read these online articles. If you conduct an online search for what to do with old floppies, you are likely to find some suggestions that they will make interesting novelty coasters. Certainly this chapter will not suggest that you store your valuable archival data on floppy disks.

Are floppies dead? Well—not exactly. There are actually a lot of industries that still need this technology. For instance, in the 1990s, a lot of machines were designed to be updated using floppy disks. It was assumed that this technology would persist for the lifetime of the machine, but this is not what ultimately happened: The machines still function just fine, but the floppy disks that they need to operate have become obsolete. Even the U.S. Air Force is still using floppies; the computers in US nuclear silos use 8-inch floppy disks, although there are plans to change this (Jones, 2015).

More likely to be of importance to you, however, is the vast amount of data that may exist on floppy disks of one type or another. At the height of their popularity, about five billion floppies were sold per year (IBM, n.d.). Even at only 1.44 MB per disk, that’s a lot of data.

It’s entirely possible that your archive will need to retrieve unique data from a floppy disk. As an example, in chapter 3, you learned about an incident in which early digital artwork created by artist Andy Warhol was recovered from floppy disks. This is not an isolated incident. For decades, people stored their personal data on floppy disks, and it’s impossible to know what interesting information might still exist on these formats and nowhere else.

For instance, Gene Roddenberry, creator of the Star Trek franchise, passed away in 1991. He left behind around 200 floppy disks, most in the 5.25-inch format. Most of the data was able to be recovered from these disks, but it was a challenge, as they were written on computers using an obsolete operating system (Mah Ung, 2015).

As another example, in 2014 Princeton University acquired the papers of author and Nobel Laureate Toni Morrison. These “papers” included some items that were not paper—about 150 floppy disks in the 5.25- and 3.5-inch sizes. From these disks, however, the Princeton University Archives were able to retrieve a variety of interesting information, such as correspondences, royalty information, and early drafts of her novel *Beloved*. What is particularly relevant about this case is that many of Morrison’s paper documents were destroyed in a fire. These digital “papers” offer another way to view information for future research (Colon-Marrero and Hughes, 2015).

It is quite possible that you, too, might need to retrieve unique data from floppy disks. It is also very possible that you have floppy disks from your own archive, the results of an earlier effort to digitally preserve material. Therefore, your goal would be to transfer these

already-existing files to a more recent form of data storage in order to protect the data from hardware obsolescence.

It is also possible that you would like to preserve floppy disks as physical artifacts for your archive, either for potential future study or for interested patrons who wish to view the collection. After all, there are many technological artifacts that your patrons might find interesting. As an example, the late fantasy writer Terry Pratchett, who sold millions of books translated into 37 languages, requested that his hard drive be demolished by a steamroller after he had passed away. This request was executed and the flattened hard drive put on display in the Salisbury Museum in 2017 so that his fans could view it (BBC, 2017). Therefore, this chapter will also cover ways to properly store a floppy disk (steamrolling is not recommended). But first, you may be wondering: What exactly is a floppy disk? And how does it work?

How Floppy Disks Work

If you've ever handled a 3.5-inch floppy disk, you may wonder why they are referred to as "floppy" disks. A 3.5-inch floppy is not floppy at all; it is made from rigid plastic. However, the older 5.25- and 8-inch floppy disks were indeed floppy.

"Floppy" is also a decent descriptor of what you find inside a floppy disk if you open it up. The outside of a floppy disk is square. The inside, however, is not. The inside is a round plastic disk coated with magnetic material—often iron oxide. Floppy disks are actually a type of magnetic media, discussed in more depth in the next chapter. Essentially, binary ones and zeroes are coded as small magnetized areas, either north to south or south to north (Gregersen, n.d.).

The data on these disks is arranged into "tracks" that go around the disk, much like tracks on a vinyl record. Also like a record, these "tracks" are read and written to with a tiny electromagnet (Gregersen, n.d.). The disk itself is spun on a spindle to help the electromagnet reach the correct track and read the data at that point on the disk. For the 5.25- and 8-inch sizes, the disk has a hole in the center for this spindle (IBM, 1977). The 3.5-inch version has a metal disk connected to the plastic film; this metal disk is instead rotated to spin the film.

The very first versions of the floppy disk consisted only of this magnetically coated plastic film, but it soon became clear that contamination was a major issue and caused a high error rate when reading the disk. The casing for a floppy disk is actually lined with fabric that cleans the disk when it is being used (Brandel, 1999). The 8- and 5.25-inch versions had openings that allowed the reader to access the disk inside. These types of disks came with paper envelopes to further protect the disk. In the 3.5-inch version, the hole was covered with a little metal shutter, which used a spring to automatically close the hole when not in use.

While this chapter is addressing the 8-, 5.25-, and 3.5-inch sizes, it should be noted that floppy disks actually came in a variety of sizes; it is simply that these were the most common and popular sizes. There were also several formats for floppy disks. For example, double-density disks were formatted at twice the density of a normal-density disk (IBM, 1977). High-density disks hold more data than a double-density disk. There are also such things as extra high-density disks, but they are uncommon. How exactly the density is increased varies from one format to the next—for instance, the coating is different on high-density 5.25- and 3.5-inch disks, which is one factor in improving the density.

Table 7.1. Common Floppy Disk Features

CHARACTERISTICS	INDICATES
8 inches, three notches	This floppy is read-only. Covering the third notch (closest to corner) makes the disk writable.
5.25 inches, notch in the corner (when write hole is facing down)	This floppy is writable. Covering the notch makes the disk read-only.
5.25 inches, notches in both corners (when write hole is facing down)	Some computer users cut an extra hole in the other side of a single-sided disk to create a “floppy” disk that could be read on both sides. The disk may have information on both sides.
3.25 inches, tab in upper left (when shutter faces down)	Moving the tab up to obscure the hole in this corner makes the floppy writable. Moving the tab down makes the floppy read-only.
3.25 inches, hole in upper right (when shutter faces down)	The disk capacity is likely 1.44 MB.
3.25 inches, no hole in upper right (when shutter faces down)	The disk capacity is likely 720 KB.
8 or 5.25 inches, one small hole near center hole	The disk is “soft sectored.” The hole is used to detect the start of the written data.
8 or 5.25 inches, multiple small holes	The disk is “hard sectored.” The holes are used to detect the start of each sector of data.
5.25 inches, ring around center hole	The disk capacity is likely 360 KB.
5.25 inches, no ring around center hole	The disk capacity is likely 1.2 MB.

You can tell something about a floppy and its capacity without ever putting it into a machine. Table 7.1 explains some common features you can look for.

Floppies could be single-sided or double-sided. As the name implies, a single-sided floppy was intended to be written to and read on only one side of the magnetic film, whereas a double-sided floppy used both sides of the disk. A small hole was used to indicate to the floppy reader whether a disk was intended to be single- or double-sided (IBM, 1977).

Now that you understand a little about how a floppy disk actually works, how can you get data from one if you need to?

Retrieving Data from Floppy Disks

Fortunately, although desktop computers have not been made with built-in floppy drives for some time now, it is not very difficult to get an external floppy disk reader, at least for the 3.5-inch version. These devices can read a floppy disk and plug into a USB port on a computer. This is a convenient way to use a modern machine to retrieve old data. Finding a reader for a 5.25- or 8-inch floppy, however, can be more of a challenge. In addition, the different densities available (e.g., a double-density versus a high-density floppy) all require their own unique readers. While some newer floppy drives can read older disks, old floppy drives cannot read newer disks with more data (Schmidt, 2000).

When it comes to recovering the data on a floppy disk, you have a couple of choices. The most straightforward of these is to simply copy the files onto a more modern storage device. However, you have another choice. *Disk imaging* is a process that exactly reproduces the data on a disk. When you use disk imaging, you can get some additional data, such as deleted file information and unused space on the disk. In essence, it tells you the exact condition of the disk when it was read for the “image” file. It is also possible to use disk imaging technology to reproduce other essential information, such as metadata (information about the file, not what is in the file) and file modification dates (Colon-Marrero and Hughes, 2015).

If you don't know what is on a floppy disk, take precautions before attempting to read it. A floppy disk could contain files that are malicious in nature and may potentially damage other files or equipment. Even if you do know what to expect on a floppy disk, taking precautions by default can be a good policy (Gialanella, 2018). This can apply to other forms of storage, as well.

Another possibility for retrieving data is to have a computer like the one originally used to write files to the disk itself. This has both convenient and inconvenient aspects to it. It can be inconvenient in that locating, operating, and storing computers for the sole purpose of retrieving data from a disk may be expensive and cumbersome. However, it should be noted that the goal of retrieving data and the goal of actually reading that data in a way that a human can understand are two different goals.

Today, there are a handful of commonly used operating systems, such as Windows 10, macOS, and Linux. A few common mobile-oriented operating systems exist, as well, such as Android or iOS. When thinking of word processing programs, Microsoft Word comes to mind.

The limited number of systems is relatively new, however. Floppy disks were in common use for more than thirty years and are still used to some extent today. It is entirely possible to encounter files that were made using programs and operating systems that have not existed for decades. In these cases, finding a computer capable of running these programs may be problematic.

It should also be noted that there are companies that can assist with both retrieving data from floppies and determining what the information is. It may be more practical, depending upon the extent of your collection and your resources, to hire such a company to retrieve and evaluate data in your collection.

It is most likely that retrieving data from a floppy disk is your primary objective. However, you may also want to try to preserve the disks themselves as an artifact for your collection. In this case, there are some general precautions that you can take.

Preserving Floppy Disks

Floppies are quite delicate. As mentioned earlier, they are highly vulnerable to contamination, to the point that floppies are designed to clean themselves during normal operation. Cleaning a floppy manually isn't really practical, as most normal cleaning methods could deteriorate the magnetic surface (Ahl, 1983). A potential contamination issue that you can solve, however, is a contaminated read/write head in the floppy drive itself. This can be cleaned using a specialized cleaning disk (Schmidt, 2000).

As a magnetic storage medium, floppies are highly vulnerable to magnetic fields, which can disrupt the data. A magnet as weak as a refrigerator magnet or a speaker can potentially damage the data on a disk (Schmidt, 2000). Like many archival materials, heat and sunlight are potentially damaging, as well (IBM, 1977).

The actually “floppy” floppies, the 8- and 5.25-inch versions, are particularly delicate. Bending, folding, or placing heavy objects on these types of floppies can disrupt the data. Even labeling one using a ball-point pen can damage the disk; felt pens were recommended. Such disks should also be stored in their protective envelopes and handled by their tops, where the label is normally placed (IBM, 1977). Disks should not be stored on top of one another, as this can also cause warping (Ahl, 1983).

It is difficult to say how long a floppy disk can last. They are, as stated, very delicate, but it is quite possible to find floppies that are in good condition and still readable. It really depends upon storage conditions and, perhaps, luck.

Floppies are now considered long out-of-date, and this chapter will cover one more storage method that is becoming vulnerable to obsolescence: optical media.

Optical Media

Floppies were really a very convenient form of data storage for a typical home computer user. The 3.5-inch version was compact, sturdy, and could be read and written to with an average computer at the time. However, floppies were limited in how much data could be stored on the disk and were rather fragile.

Optical media, such as CDs, DVDs, and Blu-ray disks, offered a more practical solution for users. They were a bit more durable and lacked the contamination issues of floppies, they were more reliable, and, perhaps more importantly, they could store much, much more data than a floppy disk. This is owing to the fact that optical media is read using a laser, which is able to focus on an area of data on the disk with much finer precision than the magnetic heads used to read and write floppies. This means that more data can be written to and read from an area on an optical disk than on the same amount of area on a floppy (Tikkanen, n.d.).

The technology for optical media has actually been around for quite some time. Development of optical technology is often attributed to a scientist, James Russell, starting in the mid-1960s. Although you probably think of optical media as being a round plastic disk, Russell’s initial prototypes were actually rectangular glass plates. The key idea behind optical media is storing and retrieving data using a laser, which was possible with these plates. Additionally, Russell originally intended to use his new recording medium specifically with music in mind, as he was a music fan and was searching for a way to play music without wearing out the recording medium, which can happen to records and tapes (Dudley, 2004).

CDs as a method of playing music data became available in the early 1980s, and by 1985, they were available as a method of data storage for software, too. The 1990s saw the invention of a writable CD, finally making this medium available for users to record their own data. In 1995, the CD-RW, a rewritable format was introduced (Philips, n.d.). Before these innovations, CDs could only be produced with their data pre-written—that is, the manufacturer produced a CD to specifically play music, contain a software program, and so on. Writable media made it possible for anyone to record data on a CD.

There is a form of commercial optical media that existed before the more familiar CD—the LaserDisc, intended for storing movies. Although it had many merits, this was ultimately an unpopular form of media. Still, LaserDiscs were created (and coveted by movie buffs) for several decades, and it is quite possible for an archive to need to access and move the contents of this storage method to a more current form of media.

Various types of optical media are easy to use and familiar to most people, and this makes them appealing as a method of storage. However, optical media of all types has started to lose popularity, so, as with the floppy, it may be helpful for you to learn more about this method of storage so that you can prevent data from being lost.

Optical Media Construction

CDs, DVDs, and Blu-ray disks are the common formats for optical media, and they are all similar in some key respects. All three have the same general construction, but the differences in their construction make a big impact on the storage capacity of each type of media.

A CD is the least complex of these three types of media. A regular, manufactured CD is composed of several layers, like a sandwich. It has a label (usually), followed by a layer of lacquer. The bottom of the CD is clear polycarbonate, which is a plastic (Byers, 2003). In the middle is a layer of a reflective metallic alloy; distinctive of CDs, this is the shiny silvery part that you see through the back. A major function of the lacquer and the polycarbonate is to protect this layer, since the CD will not function if the metallic layer is damaged.

DVDs, or Digital Versatile Disks, are constructed in basically the same way, but instead of a single metallic layer, DVDs have two. DVDs can store more information than a CD and can store information on both sides of the disk. Each side can have either one or two recordable layers of material, whereas a CD always has only one and is single-sided (Byers, 2003). Blu-ray disks have a similar construction to DVDs, but the equipment used to read a Blu-ray disk is different from that needed to read a DVD.

Earlier, you learned that a floppy disk contains a piece of film read by a magnetic reader, a little like a record player, but using magnets. A CD works similarly to a record, as well. Though you can't see it, that all-important metallic layer in a manufactured CD is not smooth. It actually has many tiny little grooves, much like a record. Those grooves are called *pits*, and the spaces between them are called *lands*. They are formed around the disk in an even spiral, moving from the inner hole to the outer edge. Rather than using a needle on these grooves, though, a CD drive has a little laser inside. The laser moves along this spiral track, exactly as a record player's needle. These grooves contain information, which can then be read by a computer, as a record player plays a record.

When a CD (or any other kind of optical media) is manufactured, the polycarbonate bottom is the part imprinted with the little bumps, and the shiny metal layer is placed over it, kind of like how a very fine layer of metal is attached to another item for gilding (Andrews, 2007). You may be wondering why the metal part is important, then, if it's not the part imprinted with these little data-containing pits.

The little pits and lands don't do anything themselves. What they can do, though, is change how the light reflects off this metal surface. When you put a disk into the CD drive of a computer, the disk begins to spin. As it spins, a laser inside the drive runs

over these grooves along the spiral. They reflect some of the light back from the laser by bouncing it off the metallic layer. Lands reflect a lot of light, and pits reflect a little light. Another device within a CD drive that detects light reads these bursts of reflected light and converts them into ones and zeroes for the computer, thus communicating the information that is encoded on the disk. With optical media, a land represents a one, and a pit represents a zero (Andrews, 2007). Because the encoding is in binary, you could potentially store any kind of digital information on an optical disk. Though CDs are often known for storing music or software and DVDs and Blu-ray disks are known for storing movies, digital files of any type are all suitable for storage on optical media.

There are several types of CDs and DVDs as well as Blu-ray disks, and they all have some different qualities. For instance, “burned” or writable optical media works a little bit differently from the kind with the data imprinted during manufacturing, but the principle is the same.

OPTICAL MEDIA ACRONYMS

ROM	Read Only Memory, a manufactured disk that can't be altered
DA	Digital Audio, audio disk that can't be altered
R	Writable, can be written to once by the user
RW	Rewritable, can be written and erased multiple times

CD-ROM and CD-DA

There are basically two types of commercially produced CDs: CD-ROMs and CD-DAs. These work exactly as the basic CD described above does: they contain a metallic layer, which is usually aluminum, with grooves in the polycarbonate to encode the information (Byers, 2003). CDs of any kind record information on one side of the disk only—that is, only the clear polycarbonate side has data.

CD-ROM stands for Compact Disk-Read-Only Memory. You might notice that the ending of this, Read-Only Memory, is the same acronym and stands for the same words as those used to describe a ROM chip. In both instances, the device, the disk or the chip, can only be *read* by a computer. The computer is not able to change any of the information on the device, which is a good thing, because then it can't be accidentally erased.

CD-DA stands for Compact Disk-Digital Audio. This is the kind of CD used for commercially produced music. It is exactly like a CD-ROM, but with one major difference: CD-DAs have some additional information used for timing, which is important for music playback (Dale and Lewis, 2013).

CD-R

CD-R stands for CD-recordable; these are blank CDs that have no information encoded until the user “burns” information onto the disk. Recordable CDs also have their information encoded in a spiral in the CD; in fact, blank recordable CDs have this spiral pre-stamped into the CD (Byers, 2003). However, they lack the pits and lands of a regular CD-ROM; the metal layer is smooth and blank. You already know that the information

is encoded on a CD-ROM during the manufacturing process; it is physically embedded into the polycarbonate bottom before the other layers are put on top. To record information yourself, though, you don't have to make CD sandwiches in some sort of miniature factory in your computer (though that might be fun). So how do CD-Rs encode information?

The back of a CD-R is shiny and metallic, but it's not silver in color, as a CD-ROM. The coloration that you do see is photosensitive dye. This dye reacts to a laser in the CD drive that, operating at a higher power than necessary for simply reading disks, is able to heat the disk and create little darkened areas of dye (Optical Storage Technology Association, 2001). The pre-stamped spiral exists for the purpose of guiding the laser while it writes information to the CD (Byers, 2003). This process is known as *burning*.

The series of dark and shiny areas is what the laser will read rather than pits and lands (Sandstrom et al., 2001). They are read in essentially the same way as a manufactured CD, with shiny parts reading as ones and dark, less reflective parts reading as zeroes. To the computer, there is no difference between these two types of CDs because they work in the same way, by either reflecting a lot or a little bit of light when a laser passes over an area on the CD.

You might notice that there are several different possible colors on the backs of CD-Rs, which vary by the manufacturing company. This is because there are several different kinds of photosensitive dyes that can be used effectively, and these dyes are different colors. All of the various dyes work, but some are more stable and durable than others. This will be discussed further later in the chapter.

The basic structure of a CD-R is a lot like a CD-ROM. The bottom is a layer of polycarbonate, the middle is a layer of a metal, and the top has a layer of lacquer and some sort of label, usually designed for the user to write on. However, there is a layer of dye between the polycarbonate and the metal. The metal is typically different, as well, and is usually gold, silver, or a silver alloy. Gold is the best choice of these, as it's very stable and resistant to corrosion. Aluminum can't be used for the reflective layer as for a regular CD-ROM because it reacts with the dye on the disk (Byers, 2003).

CD-RW

CD-Rs are not the only type of media available that can be used to record data. CD-RWs are rewritable, with CD-RW standing, rather logically, for CD-Rewritable. They are more convenient than CD-Rs in this respect: a CD-R can only be used once and the information is permanently encoded into the photosensitive dye. A CD-RW, on the other hand, can be altered and rewritten, deleting old information and adding new information. The process (and the data) is not permanent.

CD-RWs are pretty similar to CD-ROMs and CD-Rs in construction. They have a layer of lacquer and polycarbonate and, like a CD-ROM, there is a layer of aluminum in the middle. CD-RWs also have two dielectric, or insulating, layers, which are made from zinc sulfide and silicon dioxide. Rather than a dye layer, as CD-Rs have, CD-RWs instead have a layer of something known as a *phase-change metal*, which is sandwiched between those two dielectric layers. This layer is made from a combination of indium, silver, tellurium, and antimony, and is the part that contains the information on the disk (Optical Storage Technology Association, 2001).

This phase-change metal is able to exist in two different states—crystalline or amorphous (Sandstrom et al., 2001). This can be a little difficult to visualize. Imagine that you

have an ice cube and a similar amount of water. The ice cube is solid because the water molecules inside it are more organized. They have a structure. The water, though made of the same material, lacks structure. The molecules are all largely independent, so they can move around easily. The phase-change metal is like that—the crystalline state is very organized, and the same metal in the amorphous state is very disorganized at a molecular level.

As in the process of burning a CD-R, the laser inside a CD drive operates at a higher power than normal and “burns” a CD-RW. Instead of getting spots of darkened dye, however, that phase-change layer becomes amorphous, rather than crystalline, under the heat of the laser (Optical Storage Technology Association, 2001). The amorphous spots are dark, nonreflective zeroes, and the crystalline spots are shiny, reflective ones. What is unique about this is that the process is reversible up to about 1,000 times (Sandstrom et al., 2001).

CD-RWs do have drawbacks, though. Older machines are not able to read these disks, and though you *can* record audio information on them, these disks can’t be read by regular audio CD players (CD-Rs can). These types of CDs have significantly lower reflectivity than other CDs, which means that the reader for the laser doesn’t get as much

OPTICAL MEDIA TYPES

CD:

- Single layer of data on a side
- Most common media type and least expensive
- Smallest data capacity

DVD:

- Can hold multiple layers of data on a single side
- Large data capacity

Blu-Ray:

- Can hold multiple layers of data
- Largest data capacity, most compact data
- Most expensive
- Requires specialized reader

light bouncing back off the CD, and so any device that is not designed for reading CD-RWs specifically cannot do so (Sandstrom et al., 2001).

DVDs

DVD-ROMs, DVD-Rs, and DVD-RWs work in essentially the same way as their CD counterparts, with the disks being embedded with little lands or having photosensitive dyes or phase-change metals to encode the information. They also look alike and are the

same physical size. However, DVDs store a lot more information than CDs and can range from around 4.7 GB to 17 GB, depending upon how the disk is configured. Remember, a normal CD can hold 650–700 MB of data. A 17 GB DVD has an equivalent amount of around 17,000 MB of data, or around 24 times as much as a regular CD.

DVDs get so much information onto the disk because those little pits and lands are configured a bit differently and are more densely arranged. DVDs are also capable of having two layers of information on one side, and can also be double-sided, essentially quadrupling the amount of possible storage space, whereas the information on a CD can only be on one side of the disk and in one layer. This variation in construction is what accounts for the rather wide range of DVD storage capacities (Fuller and Larson, 2008).

Unlike a CD, in which the laser beam goes through the polycarbonate and reflects off the metal layer, the initial layer in a dual-layer DVD is only semi-reflective. It's typically made from silicon, gold, or a silver alloy and can allow some of the light to pass through and to a fully reflective coating on the other side, which is normally made from aluminum (Byers, 2003).

Blu-Ray

A Blu-ray disk is very like a DVD, but has several important differences. One side of a Blu-ray disk can hold up to 27 GB of information, about one and a half times the maximum capacity of a DVD (Fuller and Larson, 2008). Like DVDs, Blu-ray disks can be two-layered, having about 50 GB of storage available, and some companies are producing disks with four layers total and between 100 and 128 GB of storage space. Blu-ray disks are, as mentioned, very similar to DVDs and CDs, and Blu-ray players are often capable of reading CD and DVD disks, though players or drives designed for CDs and DVDs can't read Blu-ray disks (Overton, 2012).

A Blu-ray disk, as the name suggests, also uses a blue laser to read the contents of the disk, as opposed to a red laser for normal CDs and DVDs (Dale and Lewis, 2013). This laser is what enables so much more information to be encoded. The red laser in a normal DVD or CD player or drive has a wavelength of 650 nanometers. A blue laser has a 405 nanometer wavelength. This finer wavelength allows the laser to be more precise. The pits and lands on an optical disk are measured in microns, which is a unit of length about a thousandth of a millimeter, or one millionth of a meter (0.000039 inches). The diameter of the hairs on your head can also be measured in microns. The size of a single pit or land on a Blu-ray disk is .32 microns, as opposed to .74 microns for a DVD; so, it's less than half the size of a pit or land on a DVD. Essentially, all of the information is the same and is encoded in the same way as the other technologies, but is extremely compact, putting more information into the same amount of space (Overton, 2012).

Like DVDs and CDs, Blu-ray disks are writable and rewritable. A normal, manufactured Blu-ray disk is known as a BD-ROM; a writable disk is a BD-R; and a rewritable disk is a BD-RE. The capacity for these disks is the same as for manufactured Blu-ray disks (Overton, 2012).

Optical Media for Storage

Optical media is losing popularity, as there are more stable forms of data storage available now and the cheap, convenient CD holds a very small amount of data relative to other current forms of data storage. As with floppies, it is unlikely that you would want to store

data on new optical disks of any kind, and many computers are now being manufactured without a drive to read optical disks at all.

However, optical media is still a valid form of data storage, and its low expense might make it appealing for some situations. For instance, it might be desirable to copy information to a disk that is not intended to be archived; it is intended to be shared with a patron or with another archive. While it is possible to send files over the Internet, as well, it can be more desirable to use physical media for large files (and it can be more secure, if that is important).

Additionally, like floppies, you may need to address this form of data storage in the future, or even now. Many people used CD-Rs and CD-ROMs to store their own data, (some still do) and you may need to attempt to retrieve this data in the future. Because CDs were cheap and convenient, many archives may have used them to store or back up data, as well. In fact, the previous edition of this book recommended it as an option for some situations, and it could still potentially be a convenient choice for backing up a collection for a small or specialized archive.

Optical media is more stable and more durable than delicate floppies, but there are still some problems facing optical media as a storage method. If you need to archive optical media, retrieve data before it disintegrates, or use it as a backup yourself, there are some aspects of this storage medium that you should keep in mind.

Optical Media Storage and Longevity

Many manufacturers of recordable optical media claim a lifespan of decades for their product, and disks may indeed last for decades under ideal conditions. In a study conducted by the Library of Congress, samples from the collection were evaluated for errors after fifteen years of storage under ideal conditions. The tested disks showed about a 4 percent error rate, with lower-quality disks and disks that had been damaged during handling more likely to have data failure. In an accelerated aging study, it was predicted that about 70 percent of the test disks had a longevity rate of about 100 years, while around 4 percent of the disks would fail within ten years (Library of Congress, n.d.). There are many factors involved that determine how fast information degrades, and there are several problems that optical media is vulnerable to.

Disk Rot

As you know, optical disks work by reflecting or not reflecting a laser back to a sensor. When the laser shines off a metallic layer inside the disk, then a computer can interpret this as a one, and when this does not happen, this is interpreted as a zero. If the metallic layer becomes corroded, however, this leads to a condition known as “disk rot” or “laser rot.”

These phrases are general in nature and can describe any sort of deterioration that is causing the shiny part of the disk to reflect poorly. Remember, the top and bottom layers of an optical disk are designed to protect the delicate metal interior layer. If exposed to contamination, this layer can become corroded and will no longer properly reflect the laser, making the disk unreadable.

Dye Stability

CD-Rs as well as DVD-Rs and BD-Rs have an extra problem unique to their construction. The shiny metal layer is essential, but there are no physical pits and flats on the disk.

Again, these are instead encoded depending upon whether or not a laser, through heat, has created a dark spot on the dye layer of the disk. These dyes are therefore sensitive to heat and light, and the stability of the dyes is variable. Different manufacturers use different dyes. Phthalocyanine, cyanine, and azo dyes are some of the commonly found ones. Of these three, disks with phthalocyanine dye are the most stable (Iraci, 2019).

Again, how you store a disk greatly enhances the stability and longevity of optical media, and may be useful to note if you want to preserve optical disks as a physical object of note or use disks as a backup.

Storage Conditions

There are a number of steps that you can take that will preserve the life of any kind of optical media.

Temperature

Heat is very damaging to optical media and can lead to a variety of problems. The optimal temperature range for optical media is between 39 degrees and 68 degrees Fahrenheit (Byers, 2003). Sunlight is also damaging to optical media—particularly burned media—and so storage in an area away from windows is best (Fuller and Larson, 2003). Remember, the dye on a writable or rewritable disk is sensitive to the heat and light from a laser, and so exposure to heat and light can corrupt the data written to the disk.

Humidity

The polycarbonate base of optical media can absorb moisture. This means that getting an optical media disk wet can be damaging, and also means that the humidity at which you store optical media has an impact on its longevity. A relative humidity of 20–50 percent with fluctuations smaller than 10 percent is optimal (Iraci, 2019).

Storage

Most manufactured CDs come in a plastic casing known as a jewel case, which is important for protecting the disk inside from damage. Optical disks should always be stored inside their cases, and never stacked on top of one another, inside or outside the case. Likewise, they should not be stacked on objects or have objects stacked on them (Fuller and Larson, 2003). Optical disks should always be stored upright inside their cases, just as you would treat a book (Byers, 2003).

There are also sleeves available for storing optical media. These will protect the disks from dirt and scratches. However, they don't offer the same degree of protection that a jewel case does. It's also difficult or impossible to get the disk out of the sleeve without touching the back, and taking a disk out of a sleeve and putting it back multiple times is likely to start deteriorating the polycarbonate, adding scratches.

Handling

Always try to pick up disks by the edges, with your fingers between the center hole and outer edge, and avoid touching the shiny side (Iraci, 2019). In general, touch optical disks

as little as possible, only doing so to take a disk from its casing to put it into a reader, and to take it back out to put the disk away.

Cleaning

Scratches, fingerprints, and smudges of any kind will all obscure the laser from reaching the data in the middle of the disk. The cleaner the disk is, the better. If a disk needs to be cleaned, you can use a soft, lint-free, cotton cloth to wipe away dust and dirt. If more thorough cleaning is needed, you can use a small amount of dishwashing detergent followed by distilled water. The direction in which you wipe off a disk is important, as well, wiping in a straight line from the middle to the outside being better for the disk than wiping in a circle around the shiny side (Iraci, 2019).

If the disk is still dirty, cleaning it with a commercial optical disk-cleaning solution or some isopropyl alcohol and a soft, lint-free, cotton cloth is the best way to remove the dirt (Fuller and Larson, 2008).

Labeling

Though it may seem like a harmless activity, labeling optical media is something that must be done carefully. Writing on the disk is preferable over using an adhesive label. You should never, of course, write on the shiny side, since it will interfere with reading by the laser. Always write on the label side. Felt-tip markers are safest. Markers must not be solvent based; water-based markers are best, as the solvents can deteriorate the lacquer on the label side. You should use a pen that doesn't need to be pressed hard onto the disk. Ballpoint pens or pencils are capable of actually distorting the information on a disk, because the label is over the lacquer layer and the lacquer layer is much thinner and closer to the metallic layer than the polycarbonate layer is. Labeling double-sided DVDs is tricky in that both sides are the shiny, readable side; writing around the center hole is fairly safe (Byers, 2003).

It's often impossible to tell whether or not a disk's data is becoming corrupted unless the damage is extreme, as in the case of a delaminated disk where the layers have simply come apart, or when disk rot is very apparent. It is possible to purchase software and equipment to test errors, software to assist with recovery, and even services that can do this type of testing for you (Iraci, 2019).

IDEAL STORAGE AND HANDLING OF OPTICAL DISKS

Storage Temperature: 39–68°F (4–20°C)

Humidity: 20–50% RH

Storage: Upright, in jewel cases or other hard casing

Handling: Touch label side or outer edges only

Cleaning:

- Remove dust with air
- Wipe dirt with a soft, lint-free cotton cloth
- Remove additional dirt with an optical disk-cleaning solution or isopropyl alcohol

Key Points

- Floppy disks were an early form of practical data storage for home computer users.
- Floppy disks were used for decades and may still hold information of interest to archives and their patrons.
- CDs, DVDs, and Blu-ray disks are all different kinds of optical media, in which binary information is encoded using tiny spots that are either highly reflective or not very reflective.
- Optical media offers many conveniences in that it's fast, inexpensive, requires no special equipment, and is likely to be familiar to both staff and patrons.
- Optical media is vulnerable to problems like delamination, laser rot, and corruption of dyes.

While it's a very familiar and simple form of storage, optical media is only one way to store digital information for your archive. Magnetic storage, which has already been discussed somewhat, is one of the oldest methods of data storage and continues to have many appealing qualities for an archive.

References

- Ahl, David H. 1983. "Floppy Disk Handling and Storage." *Creative Computing*. https://www.atarimagazines.com/creative/v9n12/205_Floppy_disk_handling_and_.php.
- Andrews, Jean. 2006. *A+ Guide to Managing and Maintaining Your PC*. Boston: Course Technology, Cengage Learning.
- BBC. 2017. "Terry Pratchett's Unpublished Works Crushed by Steamroller." <https://www.bbc.com/news/uk-england-dorset-41093066>.
- Brandel, Mary. 1999. "1971: IBM Fashions the Floppy." CNN. <http://www.cnn.com/TECH/computing/9907/08/1971.idg/>.
- Byers, Fred R. 2003. *Care and Handling of CDs and DVDs—A Guide for Librarians and Archivists*. National Institute of Standards and Technology Special Publication 500-252. Washington, DC: Council on Library and Information Resources, and Gaithersburg, MD: National Institute of Standards and Technology. <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub121.pdf>.
- Colon-Marrero, Elena, and Allison Hughes. 2015. "Toni Morrison's Born-Digital Material." Mudd Manuscript Library Blog. <https://blogs.princeton.edu/mudd/2015/08/toni-morrison-born-digital-material/>.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Dudley, Brier. 2004. "Scientist's Invention Was Let Go for a Song." *Seattle Times*. http://old.seattletimes.com/html/business/technology/2002103322_cdman29.html.
- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology Comprehensive*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Gialanella, Leigh Anne. 2018. "Disk Imaging for Preservation: Part 1." Bits and Pieces. <https://www.lib.umich.edu/blogs/bits-and-pieces/disk-imaging-preservation-part-1>.
- Gregersen, Erik. n.d. "Floppy Disk." *Encyclopedia Britannica*. Accessed June 23, 2019. <https://www.britannica.com/technology/floppy-disk>.
- IBM. n.d. "The Floppy Disk." Accessed June 23, 2019. <https://www.ibm.com/ibm/history/ibm100/us/en/icons/floppy/>.

- IBM. 1977. *The IBM Diskette General Information Manual*. Bitsavers.org. http://www.bitsavers.org/pdf/ibm/floppy/GA21-9182-3_Diskette_General_Information_Manual_Sep77.pdf.
- Iraci, Joe. 2019. "Longevity of Recordable CDs and DVDs—Canadian Conservation Institute (CCI) Notes 19/1." Government of Canada. <https://www.canada.ca/en/conservation-institute/services/conservation-preservation-publications/canadian-conservation-institute-notes/longevity-recordable-cds-dvds.html>.
- Jones, Brad. 2015. "Think the Floppy Disk Is Dead? Think Again! Here's Why It Still Stands between Us and a Nuclear Apocalypse." Digital Trends. <https://www.digitaltrends.com/computing/why-do-floppy-disks-still-exist-the-world-isnt-ready-to-move-on/>.
- Lake, Susan, and Karen Bean. 2008. *The Business of Technology: Digital Multimedia*. 2nd ed. Mason, OH: South-Western Cengage Learning.
- Library of Congress. n.d. "CD-ROM Longevity Research." Accessed June 30, 2019. https://www.loc.gov/preservation/scientists/projects/cd_longevity.html.
- Mah Ung, Gordon. 2015. "How Star Trek Creator Gene Roddenberry's Words Were Freed from Old Floppy Disks." PC World. <https://www.pcworld.com/article/3018315/star-trek-creators-lost-words-recovered-from-old-floppies.html>.
- Optical Storage Technology Association. 2001. "Understanding CD-R & CD-RW Disk Construction and Manufacturing." <http://www.osta.org/technology/cdqa15.htm>.
- Overton, Gail. 2012. "Can New Techniques Continue to Densify Optical Data Storage Capacity?" *Laser Focus World* 48, no. 10: 39–42. EBSCOhost.
- Perenson, Melissa J. 2012. "Blu-Ray: Frequently Asked Questions." PCWorld. <http://www.techhive.com/article/128205/article.html>.
- Philips. n.d. "The History of the CD—The CD Family." Accessed June 28, 2019. <https://www.philips.com/a-w/research/technologies/cd/cd-family.html>.
- Sandstrom, Chad R., Gordon Rudd, and Robert DeMoulin. 2001. "How Do Rewritable CDs Work?" *Scientific American*. <http://www.scientificamerican.com/article/how-do-rewriteable-cds-wo/>.
- Tikkanen, Amy. n.d. "Optical Media." *Encyclopedia Britannica*. Accessed June 28, 2019. <https://www.britannica.com/technology/optical-storage>.



Magnetic Tape

IN THIS CHAPTER

- ▷ How are magnets useful in computer technology?
- ▷ How is magnetic data encoded on tape, and how is the tape constructed?
- ▷ What is the typical storage capacity of magnetic tape, and how does it compare to other options?
- ▷ What are the benefits and drawbacks of using magnetic tape for archival storage?
- ▷ What is the optimal way to store magnetic tape?

When you think of a magnet, you probably don't consider its possibilities for anything other than its main quality: that a magnet is attracted to certain types of metals as well as to other magnets. This is an extremely useful quality, but a magnet's characteristics on a smaller level are even more remarkable.

Magnets have actually been used as a form of data storage for more than a hundred years. While you probably think of things like records when considering early forms of audio storage, magnetic storage has been used for recording audio since 1898, when a device called the telegraphone was patented by an engineer named Valdemar Poulsen (Library and Archives Canada, 2015).

For decades, this was a major way of recording sound. It was more reliable than its other major competitor at the time of its invention, the wax cylinder, and eventually the technology was adapted so that people could purchase their own recorders and record their own sounds on steel wire. Though this form of recording may be unfamiliar to you, it was in use until the 1970s (Strongman, 2016).

The steel wire method of recording was a remarkable invention, but it had some problems. The wire was heavy, and the range of pitches it could record was limited. In the 1930s, development began on a product that could perform the same function as steel wire, but better. This product, developed by German researchers, was a celluloid tape

coated with iron oxide. This invention had a variety of advantages over traditional steel wire, including the fact that it was much lighter and more easily magnetized (Library and Archives Canada, 2015). More important, though, was the fact that it recorded sound with much higher fidelity than wire, and once the technology had become more developed and the price went down, tape quickly overtook wire in popularity (Strongman, 2016).

You may be wondering at this point why this book is discussing recording sounds on wires and tapes. Like some other technologies discussed so far, while magnetic storage was originally designed for another purpose (recording sound), it was soon recognized as having potential for storing digital information, too. After all, magnets have a built-in binary value, just as computers do, in their polarity—either south to north, or north to south—and in 1951, a type of magnetic tape intended for storing computer data specifically was developed (Coughlin, 2012).

You might recall that in the previous chapter, floppy disks were described as a form of magnetic storage. The next chapter will cover hard drives, which are yet another form of magnetic storage. Magnetic materials are really quite useful when it comes to data storage, but tape in particular has some unique characteristics that make it suitable for an archive.

The previous chapter discussed two forms of data storage that a typical computer user could use to store their own data. Because of this, you are most likely already at least somewhat familiar with these forms of data storage: floppy disks and optical media.

Magnetic tape is also available for typical computer users, but for several reasons, it's not very popular in modern times as a form of data storage. As a result, most people are not so familiar with this form of data storage. There are, however, some forms of magnetic tape that were aimed at a general audience. Cassette tapes, designed to play music, are a form of magnetic storage. There have also been software programs released on cassettes designed for this purpose; these devices look very similar to the typical audio cassette, but are coded to run a software program.

The fact that magnetic tape is not familiar to most people is not an indicator that it has become obsolete or that it is not of value. It's really that magnetic tape is impractical for an average user. One of magnetic tape's biggest assets is its price: no other form of storage quite beats the value of magnetic tape. However, because specialized equipment is required, a person would need to store a lot of data in order to get the cost-saving benefits. An archive, on the other hand, might indeed need to store a lot of data.

Magnetic tape is also one of the oldest methods of data storage, and this is another potential advantage for you. Though modern forms of magnetic tape are not quite like earlier versions, there are many similarities, and so the strengths and limitations as well as the potential lifespan of magnetic tape are well understood at this point.

Even if your archive does not decide to use magnetic tape as a form of storing *new* data, magnetic tape has been used as a method of data storage for such a long time that it's in your interest to be familiar with it should your archive ever need to deal with archival tape. Again, it was a major method of storing many types of information for decades and persists as a major method of data storage today.

If you are not familiar with this method of storage, then magnetic tape is essentially what it sounds like: a piece of tape with a magnetic coating. But how is this used to store data?

Every method of data storage for modern computers uses binary encoding. Optical disks, as described in the previous chapter, encode the required binary ones and zeroes as either strong or weak light signals, respectively. However, this is merely one of many ways in which binary information can be encoded. For instance, punch cards and paper tape used holes in a piece of paper to code data. All that is needed to effectively encode binary data is something that can be interpreted as having an either/or condition. For instance, a light bulb is either on or it's off. With optical media, the either/or condition is the strength of the light: Is the light bright—yes or no? Within the computer itself, the electronic impulses are either high, or they're low. They're never high *and* low, and they're never sort of high or sort of low.

Magnets are ideal for encoding binary information for just this reason, lending themselves nicely to an either/or condition and offering some different options for either/or conditions, as well. As you probably remember from basic science classes or even just playing with magnets as a child, magnets have poles—north and south—a situation referred to as *polarity*. The like poles repel one another and opposite poles attract one another. So, a magnet can be encoded by using its poles. Is the pole north, or is it south? Magnetizable materials in general can also be categorized by whether or not they are actually magnetized. Just because something is magnetizable doesn't mean that it's magnetized, and magnetizable materials have the potential for both states. For instance, when you were little, you may have run a needle across a magnet to make a compass. The needle is made of *magnetizable* material (probably steel), but isn't *magnetized* until you run it across the magnet. Since electricity can magnetize certain metals, this is perfect from the standpoint of a computer: a jolt of electricity can transform a material from one state to the other—that is, from magnetized to non-magnetized, and vice versa.

Magnets are common in everyday life, found in all kinds of things like toys and refrigerators and headphones, but you may not think much about how they actually work since their everyday function is so familiar. There are two types of magnets: *permanent magnets*, and *electromagnets*. Both are necessary for magnetic tape (and computers in general).

Permanent magnets are magnetic because of their physical structure. As you probably know, also from basic science, the universe is formed of materials known as elements. Each element has some special properties due to its form and behavior on an atomic level, or the level of a single unit of an element, the atom. Not every element is magnetizable. Elements capable of generating a magnetic field have a peculiar property that is unique to them. To think of it in simple terms, this property allows the element to generate a magnetic field and produces the phenomenon of magnetic poles—north and south—even at the atomic level. What this means to you is that magnets can be exceedingly small, which is important for encoding data, since with computers, smaller is almost always better. Technically, a single atom of a magnetizable material could be considered a magnet. Iron, nickel, and cobalt are some common examples of elements with this unique property (“Magnet and Magnetism,” 2006).

Commercial magnets are often a combination of elements—for instance, a common Alnico magnet is made from aluminum, nickel, and cobalt. In a permanent magnet, the fields of the atoms of the magnetizable elements making up the magnet line up according to their poles, and thus have an overall magnetic field. Permanent magnets need to be orderly, even on an atomic level. If a permanent magnet becomes demagnetized, then the

fields of the atoms in the magnet are randomly oriented, and the magnet has no overall pull or influence on other materials. There are a number of ways that this can happen, such as storing the magnet at improper temperatures or subjecting the magnet to the influence of another magnetic field, which can disrupt the field of the first magnet.

In an *electromagnet*, the magnetic field only exists when electricity is running through the wires that make up the magnet. Any wire conducting electricity generates a magnetic field that is just like a permanent magnet, but this field is extremely weak, so you'd probably never realize that a magnet was present (it would be awkward if your paper clips were attracted to your desk lamp, for instance). To make the field stronger, you can bend the wire into a loop ("Magnet and Magnetism," 2006). The more loops you have in the wire, the stronger the magnetic field becomes. Electromagnets have poles, just as the permanent kind, though the poles can be reversed by reversing the flow of electricity (while nothing changes the polarity of a permanent magnet in this way). Both types of magnets—permanent magnets and electromagnets—are essential for creating magnetic storage and for reading it.

Encoding Magnetic Data

A piece of magnetic tape consists of sets of data placed along tracks. You can almost picture the tracks as being like a racetrack in structure. Binary ones and zeroes lie along each track. They are read by a read/write head, with one head per track, similar to how there is one runner per track. So, if a tape had nine tracks, then the drive for the tape would have nine read/write heads (Wiehler, 1979).

The reading head and writing head, of course, have different functions (though both are often part of the same device, rather than two separate heads) (Bycer, 1965). Within each head is a little device with a small electromagnet inside. This electromagnet is sensitive to magnetic fields.

A simple bar magnet has a north and a south pole. Imagine that within those tracks on the tape are sets of bar magnets. These magnets are all parallel to one another, but can be oriented north-south or south-north. Though a simplification, this is essentially what is really happening on a piece of magnetic tape, except that these bar magnets are very, very small. Typically, a person thinks of a magnet as being a big chunk of metal or an alloy, but this isn't necessarily the case. The magnets on the tape are, in fact, made from powder, but still have the function of a permanent magnet because the fields of the atoms in the powder are oriented to have an overall magnetic field. It's so weak and the spots of magnetism are so small, though, that you'd never have any way of noticing this yourself.

These minuscule magnets encode the binary data based on how their poles are oriented toward one another on the tape. Remember, the reader head is sensitive to magnetic fields. When it runs along a track, the electromagnet in the head is able to sense the orientation of the fields below, and then converts the information about the fields it senses into electrical signals that the computer can interpret as binary code (Weihler, 1979).

The writing head, in contrast, generates a field that can change the orientation of the poles on the tiny tape magnets. A magnetic field can influence the properties of another field, magnetizing or demagnetizing it, which is the phenomenon in effect here. The writing head can create binary information on a new tape by magnetizing the tape below or changing the information already there by altering the orientation of the poles. In this way, the reader head encodes information onto the tape (Weihler, 1979).

There are three basic types of magnetic tapes: reel-to-reel (which involves tape moving between two large reels), cassettes, and cartridges. The latter two are smaller than a reel-to-reel tape and are housed within a plastic casing.

It's unlikely that you will encounter new data being put onto a reel-to-reel setup, but this was a common way to store data for decades. In a reel-to-reel setup, the reel on which the tape is stored is fitted onto a spindle for playback or recording, and the free end is threaded through the machine. An empty reel takes up the tape as the tape plays. If you've ever used a microfilm reader, the process is similar to this, with a reel of microfilm fitting onto a spindle and the loose end of the film running through the reader and onto an empty reel. This kind of tape needs to be rewound to get back to the beginning of the tape.

A cassette, as you may know from seeing audio cassettes, is like the reel-to-reel setup but contains both reels within a protective case. Cassettes are put into a tape drive, and the tape runs from one head to the other. Like a reel-to-reel setup, cassettes need to be rewound to get to the data at the beginning. A cartridge works a little differently from a cassette in that it only has one reel. A cartridge is designed to play the data within in an endless loop, with the tape continually being wound onto the single reel ("Magnetic Recording," 2006). The advantage of this is that, unlike reel-to-reel or cassette tape, cartridges don't need to be rewound. Once the tape plays through, it's ready to be played once again (Gifford, 1977).

Regardless of the type, all magnetic tapes have the same basic construction. A piece of magnetic tape essentially has two layers. The top is a layer of a magnetizable material. This is often some kind of iron oxide, chromium dioxide, or pure iron. Iron oxide tends to be more stable, while chromium dioxide and iron tend to be higher quality (Northeast Document Conservation Center, n.d.). This is also known as the pigment layer or the pigment.

This pigment layer is suspended within a polymer binder. While these are the only materials necessary for the top layer, the binder may contain other materials to facilitate tape reading and recording, such as a lubricant to help the tape move more easily through a device or a cleaning agent for the heads on the device used to read or write to the tape (Van Bogart, 1995). Although it's not essential to data storage, the lubricant on a tape can be very important to proper playback, and tapes that lose their lubricant can have problems with playback. Binder coatings may also be on the back of a tape (Northeast Document Conservation Center, n.d.).

This magnetizable layer is, of course, the important part, because it is all that is needed to store the data. However, magnetic tape needs to be rolled up onto a reel, and since the magnetic material is essentially a powder, it is not suitable for this kind of treatment. The bottom layer is essential for making the tape function.

The bottom layer is a piece of film. In the past, this part was made from cellulose acetate, which is a plastic material created from a preparation of cellulose fibers (a plant material). However, this material has several problems, including its tendency to become brittle due to moisture in the air and its susceptibility to a condition known as *vinegar syndrome*, which is also caused by moisture. Vinegar syndrome, as the name implies, causes tape to smell strongly like vinegar. This is because, in the tape, the moisture is creating damaging acetic acid (which is essentially vinegar), which causes the tape to become soft or even dissolve into a powder or a slimy substance. If you ever handle archival magnetic

tapes, you may encounter this as a problem. Because of the problems created by vinegar syndrome, modern tapes are made from polyester, which is far more stable (Northeast Document Conservation Center, n.d.). There may also be a third layer on magnetic tapes, a back coat, the purpose of which is to control static or friction as the tape moves through a machine (Van Bogart, 1995).

Reading Tapes

Again, a magnetic tape basically looks like what the name implies—a long piece of plastic tape.

Think about what a music cassette tape looks like. There are two reels and tape inside the case. Magnetic tape in a cassette for archiving works exactly the same way, with two reels that move the tape between them as they rotate. With a cassette tape, the reels move in one direction to play the music on one side, then the user flips the tape to play the music on the other side. Of course, you can rewind a tape to replay a section of music before it gets to the end.

Magnetic tapes for archiving work a little differently in this respect. While a music album or audio recording has a logical sequence of information, playing from the beginning of a performance to the end, this is not necessarily the case with a magnetic tape for computer data. For example, some people use tapes to make backups of their hard drives. If you do something like this, then the software that you use will write the files on the hard drive to the tape in the order that the computer decides. A person at a music company decides the logical order for music on a musical cassette tape and makes it easy for humans to use. This may not be what happens with a data tape (though you can potentially store like data on the same tapes to improve the logic).

Both magnetic tapes for music and magnetic tapes for data use a method of storage known as *sequential access*. This means that the heads need to go through all of the data that precedes the desired data on the tape, and the tape must also physically move to the desired location under the heads. If you had a favorite song on a music tape, for instance, you'd have to fast forward or rewind in order to listen to the song again, which takes time and can be tedious.

Hard drives and floppy disks, however, use *direct access*. The physical location of specific data is less important, because the head is located on an arm that can move up and down the disk as it spins, seeking out the requested information. In contrast, the reader heads on a tape drive are fixed in place. Optical disks, such as CDs and DVDs, are another example of a method of data storage that uses direct access. Because the physical location of the data is less important and the arm can seek it out rather than go through all the preceding data, direct access is very quick in comparison to sequential access.

Typical computers do not come with built-in tape readers, so a computer needs a device to interpret the data on the magnetic tape and to communicate with the machine. As an example, a tape cartridge requires a tape drive to both write to and read the tape. Such a drive can connect to the computer in several ways, but for most purposes, a drive that connects via a USB port will likely be the most convenient. There are also drives that can be installed directly into a computer; if you don't need to be able to move the drive from one computer to another and have a dedicated computer that only makes tape backups, this is an appealing feature. When you purchase a drive, you need to keep the format and size of tape that you want to use in mind. There are a number of standard formats and

sizes available, and your drive must match these specifications. As an example, though a VHS tape and a cassette tape are similar in many ways and even look somewhat similar, you can't play a cassette tape in a VHS player. However, while a tape drive may only be able to write with one format, oftentimes, drives can read several different formats, which is helpful if you want to be able to read tapes from different archives (Andrews, 2006).

Magnetic Tape for Archiving

As time goes on, the variety of ways to store data keeps increasing. Engineers keep striving for devices that are smaller and smaller while offering more convenience, greater data safety, and higher storage capacity. Also, as time goes on, many devices become obsolete—some very quickly.

So, since magnetic tape is such an old method of storage, you might suppose that it's obsolete, or will soon become obsolete. This is not the case at all. Like other methods of data storage, tape keeps improving, year after year. Though other methods of storage are competing with magnetic tape as the best method for data storage, tape continues to be a major contender, and tape's best features have yet to be imitated in any other storage method. Still, magnetic tapes do have drawbacks, and so it's important to know about both benefits and drawbacks before you decide if tape is the right choice for your archive's storage needs.

ADVANTAGES AND DISADVANTAGES OF TAPE STORAGE

Advantages:

- Very inexpensive per GB of storage space
- Not dependent upon a specific reader manufacturer
- Very large storage capacity
- Old and well-understood method of data storage
- Easily moved off-site
- Doesn't need a power source for effective storage

Disadvantages:

- Magnetic tape readers are an expensive specialty item.
- Storage may be expensive.
- Magnetic tape is unfamiliar to many people.
- Magnetic tape is somewhat delicate.
- Reading and writing data to magnetic tape is slow.

Advantages

One of the most appealing features of magnetic tape lies in its price. Compared to other storage methods, magnetic tape is the clear winner when it comes to storing a lot of data

inexpensively, with tape costing less than a penny per gigabyte of storage space (Graham-Rowe, 2010). It's the medium of choice for many companies that need to back up a lot of data year after year, and it's appealing for archives for just the same reason. As the years go by, there will be more and more data that needs to be saved, and storing this data can become quite expensive when using other storage methods.

Magnetic tape comes in several standard sizes, just like optical disks, like CDs and DVDs, come in a standard size. This is extremely important, as it makes it possible to use magnetic tapes and readers that were produced by different manufacturers, and prevents your data from becoming obsolete because a particular company no longer makes tape in a certain size (Wiehler, 1979). For example, all musical tape cartridges are the same size and will play in any player produced by any company, regardless of who manufactured the tape.

Capacities for magnetic tape vary, and the amount of data that can be stored on a tape keeps increasing as technology improves, allowing more data to be read without simply adding more tape to a cartridge. The very first commercially available magnetic tapes held a mere 1.1 MB of data. As of 2018, it was possible to obtain a single tape cartridge that would hold 15 TB, or terabytes, of data (Lantz, 2018).

In fact, the low cost and high capacity of tape are so appealing that it is actually used as a method of backing up data for large tech companies. As an example, in 2011, Google updated the software for its online e-mail product Gmail. Unfortunately, an error in the update deleted saved e-mails for 40,000 accounts. Unable to restore the data from data center hard drives, Google eventually used a tape backup to restore these lost e-mails (Lantz, 2018).

The future may offer even *more* storage capacity. A 2017 breakthrough in tape technology led to the development of magnetic tape capable of storing 330 TB of data (Anthony, 2017). There are also possibilities for magnetic storage that have not yet been developed into a practical application. For instance, it is possible to store magnetic data as a single molecule; this technology is not currently in use, but may lead to storage materials with extremely high data capacities in the future (University of Manchester, 2017).

Though innovations to technology are helpful, the fact that this medium has been in use for such a long time can be an asset to you, as well. Because tape has been around for decades, its limitations are well understood. For instance, researchers discovered a problem that occurs with tape during the manufacturing process. The plastic is formed in large sheets, which are cut to the correct size for a reel. If the blades for cutting the tape become dull, they can cause tiny rips in the ends of a tape that are invisible to the naked eye but can shorten the lifespan of the tape or even destroy data; information like this can be used for quality control and can improve the final product. Because tape has been studied for so long, issues like this are known and can be addressed (Lawrence, 2003). Similarly, while more recent technologies haven't existed long enough for a definitive estimate on lifespan, tape has been around for long enough that lifespan estimates are fairly accurate.

Along with storing a lot of data, tape offers some other benefits. For instance, tape is easy to take off-site. If you have the funds for it, you can construct a room or building designed to store tape at its optimal temperature. If this room is not part of your library or archive, then if anything should happen to your main building, like flooding or other natural disasters, your tape backup will be safe. You could even make several backups of your data and easily store them in different locations. It's also handy if you want to share information with another archive. For instance, sending a terabyte's worth of information to someone online is a challenge and will take some time. Shipping a reel of tape may be quicker, easier, and safer from the point of view of data security.

Tape doesn't need a continual source of power to store its data, in contrast to the RAM memory chips in a computer, nor does tape require the Internet to access information. The fact that you can store tape away from a computer and that it needs a special machine to read the data encoded on it is a potential advantage to you for data security. Theft of magnetic tape is highly unappealing because the data is hard to access, since it requires a special reader to retrieve the information.

Disadvantages

Though tape has a lot to offer for an archive, for certain situations, it's not the ideal choice. If you *don't* need to store very large amounts of data, then magnetic tape may not be the best choice for you. While the tape itself is very inexpensive, computers don't come with readers for magnetic tape. You'll have to purchase extra equipment, like a tape drive and software, to both read and write to the tape. However, as mentioned before, it should be noted that there are tape drives aimed at a typical computer user for backing up personal data. Computers using the Windows operating system often come with software that is designed to write a hard drive to a tape drive for the purpose of backup, and when you purchase a tape drive, it often comes bundled with useful software, which can make things easier for your archive (Andrews, 2006).

Storing the tape can be an additional expense, especially if you want a robotic or automated system of tape storage and retrieval, which is sometimes used in large facilities. These devices don't require a human user to search for and retrieve tape from your collection. Magnetic tape is also rather delicate, and maintaining a temperature- and humidity-controlled environment is another expense to your library and will be a constant expense rather than a one-time thing.

A tape library is a device that can store tapes and retrieve their data using an automated method. They come in a variety of sizes and can be useful for larger archives and projects. There are also similar, smaller devices known as "autoloaders," which can read one tape at a time.

Magnetic tape is also an unfamiliar medium in comparison to others discussed in this book. A cassette tape is often used as the example in this chapter because this is the kind of magnetic tape that most people are familiar with—and cassette tapes are typically used only for audio recordings and are considered largely obsolete. Magnetic tape in general is not an obsolete form of storage, but your staff may have trouble adjusting and learning to use this less familiar medium.

Magnetic tape is not invulnerable. While the plastic substrate that the magnetic coating adheres to is tough and durable, there are limitations in the construction of tape. Dust or dirt on the tape will disrupt the data and cause problems with reading data. The plastic is vulnerable to warping, sticking, and other problems.

In addition, writing to and reading data from a tape is slow, much slower than other methods of data storage and recording. A magnetic tape, as mentioned earlier, is sequential in nature. This means that, every time you want to retrieve data from a magnetic tape, the head on the tape drive has to go through every bit of data that comes before the information you want, physically turning the reels and looking for the location of your requested information. Though improvements are being made, this is comparatively slow. This means that tape is best used for long-term backup—something that you don't plan

on needing again soon and exists merely to keep information safe. It is not a good choice for quick access to the content.

Though magnetic tape has been in use for decades and is one of the oldest and best understood methods of storing data, it is not invulnerable to obsolescence. Though tapes do come in standard size and formats, it's possible for those sizes and formats to eventually become replaced by better versions. As mentioned in the previous chapter, floppy disks originally came in 8-inch size, which were replaced by 5.25-inch size, and then later were replaced by 3.5-inch size. All of these are floppy disks and use the same basic technology and can be read on a variety of computer brands. However, it's a challenge to find a device that will read one of the larger floppy disks today. Likewise, the equipment necessary for recording and reading data from older tapes may eventually become unavailable.

As advances are made with tape, equipment necessary to operate the old tape becomes obsolete and companies stop making it for practical reasons. Though there are standards with tape, once a type of tape becomes outdated, there is no guarantee that modern equipment will be able to operate old styles of tape, or be backward compatible. The data on older reels of magnetic tape may be permanently lost due to an inability to retrieve the data stored on them, just as the data on computer punch cards can't be retrieved because there are no longer computers that can interpret the data.

What this means to you is that a tape archive cannot be static. You must plan for needing to move tapes from older cartridges to new ones before it becomes impossible to read the old ones. Some companies do make equipment that converts one tape format to another; this will be another expense for your archive and could require people to learn to use the new equipment and to transfer information from one kind of tape to another (Bigourdan et al., 2006).

In addition, if something cheaper and more convenient than tape is ever developed, you may have a problem on your hands, since ordinary computers don't come with a way to read tape. In contrast, if you had a collection of information on floppies before they became outdated and wanted to move them to, say, a CD, there was a large window of time in which average desktop computers had both floppy drives and CD drives, meaning that you would have needed no special equipment for the transfer. However, since so many institutions use magnetic tape for mass storage, it's unlikely that tape will be dropped from use as quickly as the floppy disk was.

Storing Magnetic Tape

Like any method of data storage, keeping the item that the data is stored on safe is essential to long-term preservation. Tape has a number of enemies that must be combated, and, in general, having a location designed for the tape that has the proper conditions is the best plan.

Humidity

Magnetic tape is vulnerable to water, and high humidity is the most dangerous condition to tape. While there are a number of ways that a tape can fail, the polymer that binds the magnetizable substance to the substrate is the part most likely to fail on a tape, and it can degrade if exposed to moisture. This can result in tape that is brittle, soft, or sticky

and unplayable, or the magnetic powders may come off the polyester substrate. It's also possible for the magnetic coating to oxidize if exposed to moisture or, essentially, for your tape to get rusty (Bigourdan et al., 2006). An additional and rather distasteful problem is fungal growth, which can occur on tapes exposed to humidity. High humidity for a tape is anything above 65 percent relative humidity (Orio et al., 2009). Storing tape at around 40 percent relative humidity is best (Bigourdan et al., 2006).

Temperature

Temperature is another important condition. High temperatures can cause the plastic to warp, distorting the data on the face of the tape. It can also increase the rate at which tape decays, particularly if there is also high humidity present (Orio et al., 2009). Optimal storage temperature for tape is around 20 degrees Celsius, or about 68 degrees Fahrenheit (Bigourdan et al., 2006). Temperatures that are too low can be harmful as well; tape must be kept above freezing temperatures. In addition, if the tapes and their drive are not in the same area, then the tape needs to acclimate to the new temperature before playing or recording (Van Bogart, 1995). It's important to keep tape away from any sources of heat, such as radiators or heating units, and to keep it out of direct sunlight and away from windows, even ones that aren't functional, since glass doesn't insulate well.

Pollutants

Dust and dirt are also enemies of magnetic tape. While a laser can sometimes read through dirt on a CD, this is a challenge for the read/write head for magnetic tape, since the read/write head is in close contact with the tape. The data on a tape is so tightly compacted that a single particle of smoke from a cigarette is enough to obscure data from the head for modern high-density tapes. Any smog or other chemicals in the air can also disturb the data or cause chemical deterioration of the tape (Orio et al., 2009). Therefore, the cleanliness of the air and the concentration of airborne pollutants is a factor to consider when choosing and maintaining a storage area for your tapes (Van Bogart, 1995).

Magnetic Fields

Since magnetism is key to both reading from the tape and writing to it, magnetic fields are another problem. As with floppy disks, even a relatively weak magnetic field, like one on a microphone or a headset, can be powerful enough to weaken the strength of data written on tape (Orio et al., 2009). Remember, magnets and electricity can influence other magnets, and if the atoms making up a magnet, like magnetic tape, become randomized, then the magnetization is lost and so is the binary code stored on the tape. Magnetic fields interfering with the data on tapes is generally not a major issue, but it's good to take precautions. Tapes should not be stored near any electronic equipment or machines that might generate a strong magnetic field (Van Bogart, 1995).

Other Problems

Though not quite within the abilities of an archive to address and combat, distortions, deformities, and other problems with the physical qualities of the tape itself is another

issue in regard to data preservation (Orio et al., 2009). This can be avoided by storing the tapes properly. Tapes shouldn't be stored flat; the reel that holds the tape should be perpendicular to the shelf (Van Bogart, 1995). This is similar to how CDs and DVDs should be stored, since flat storage can cause warping in both media and lead to an inability to retrieve the data.

While all this might sound somewhat complex, the ideal tape storage conditions aren't much different from ideal room temperature conditions and are within the capabilities of most archives to handle with minimal equipment. If you had a small building with a fairly small collection, for instance, you could simply dedicate a closet to storing your collection of tapes and just be sure that it stays cool and dry inside.

IDEAL STORAGE AND HANDLING OF MAGNETIC TAPE

Temperature: 68°F (20°C), do not freeze

Humidity: 40–65% RH

Storage: Upright, not flat

Storage Location:

- Away from windows and sunlight
- Away from airborne dust, smoke, chemicals, and general pollution
- Away from magnetic fields and electronic equipment

Transportation:

- Do not subject to harsh temperatures and moisture.
- Wind tape tightly and pack well with bubble wrap.
- Do not subject to handheld metal detectors.

While tape can be transported, either to an off-site location for safety or to another archive, it's important to remember to maintain optimal conditions during transport. The temperature should never exceed 100 degrees Fahrenheit and the tape must not be exposed to water. It's best if the tape can be transported in the same position that it's stored, upright rather than flat. Ensuring that tapes are properly wound and that they are protected by packing materials like bubble wrap, which will absorb shocks, will help to protect tapes. Some detectors, like those used in airports, can erase tapes. Walk-through metal detectors and X-ray scanners aren't an issue, but hand-held metal detectors generate a powerful enough magnetic field to erase a tape (Bogart, 2009).

It's estimated that, when magnetic tape is stored under typical room conditions, it will last anywhere between 10 and 30 years. Cooler, drier conditions can increase this life expectancy (Bigourdan et al., 2006).

As part of your routine for storing magnetic tapes, you should examine them periodically for damage. In general, there are two ways of going about this; usually both are done at the same time. You can physically examine tape to determine the extent, if any, of decay, or a computer can also play through the data to determine if there is any decay.

Your recording device will also require care. Remember, the read/write heads on a player are extremely sensitive and come in very close contact with the tape. Be vigilant and clean them when necessary. For a tape drive designed for a personal computer, there are such things as cleaning cartridges, which can be inserted, instead of a data cartridge, for the purpose of cleaning the drive (similar devices exist for floppy disks, as well). The drive will run the tape through, and the tape will help clean the heads. There is also such a thing as a head cleaning spray, which is a little like compressed air used for cleaning the insides of computers and keyboards (Andrews, 2006). This gets any dust or particles that might interfere with reading the tapes out of the heads.

Key Points

- Magnetic tape is one of the oldest methods of long-term, mass data storage and continues to be going strong today.
- Magnetic tape is durable and inexpensive and its vulnerabilities and estimated longevity are well understood, making it one of the best choices for a large archive.
- Magnetic tape does require controlled storage conditions for optimal function and longevity, and it needs specialized equipment to read and write to the tape, which adds to the overall expense. It's typically not the best choice for a small project or library.

Tape is a good choice for mass storage, but it is not your only choice for large amounts of data storage. Hard drives, another form of magnetic storage, have been in use for some time and may still be in use for the foreseeable future.

References

- Andrews, Jean. 2006. *A+ Guide to Managing and Maintaining Your PC*. Boston: Course Technology, Cengage Learning.
- Anthony, Sebastian. 2017. "IBM and Sony Cram up to 330 Terabytes into Tiny Tape Cartridge." *Ars Technica*. <https://arstechnica.com/information-technology/2017/08/ibm-and-sony-cram-up-to-330tb-into-tiny-tape-cartridge/>.
- Bigourdan, Jean-Louis, James M. Reilly, Karen Santoro, and Gene Salesin. 2006. "The Preservation of Magnetic Tape Collections: A Perspective." Image Permanence Institute. <https://www.imagepermanenceinstitute.org/imaging/research/magnetic-tape>.
- Bycer, Bernard B. 1965. *Digital Magnetic Tape Recording: Principles and Computer Applications*. New York: Hayden Book Company.
- Coughlin, Tom. 2012. "Magnetic Tape Turns 60." *Forbes*. <https://www.forbes.com/sites/tom-coughlin/2012/05/17/magnetic-tape-turns-60/#3bb5af4c611e>.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology Comprehensive*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Gifford, F. 1977. *Tape: A Radio News Handbook*. New York: Hastings House.
- Graham-Rowe, Duncan. 2010. "New Life for Magnetic Tape." *MIT Technology Review*. <http://www.technologyreview.com/news/417218/new-life-for-magnetic-tape/>.
- Ken, Lawrence. 2003. "Old Tape Gets New Edge." *Machine Design* 75, no. 11: 52. EBSCOhost.

- Lantz, Mark. 2018. "Why the Future of Data Storage Is (Still) Magnetic Tape." *IEEE Spectrum*. <https://spectrum.ieee.org/computing/hardware/why-the-future-of-data-storage-is-still-magnetic-tape>.
- Library and Archives Canada. 2015. "Virtual Gramophone: Canadian Historical Sound Recordings. History: The Tape Recorder." <http://www.collectionscanada.gc.ca/gramophone/028011-3021.3-e.html>.
- "Magnet and Magnetism." 2006. In *Encyclopedia Americana*. Danbury, CT: Scholastic Library Publishing.
- "Magnetic Recording." 2006. In *Encyclopedia Americana*. Danbury, CT: Scholastic Library Publishing.
- Northeast Document Conservation Center. n.d. "Inherent Vice: Magnetic Media." Accessed July 19, 2019. <https://www.nedcc.org/preservation101/session-6/6inherent-vice-magnetic-media-topics>.
- Orio, Nicola, Lauro Snidaro, Sergio Canazza, and Gian Luca Foresti. 2009. "Methodologies and Tools for Audio Digital Archives." *International Journal on Digital Libraries* 10, no. 4: 201–20. EBSCOhost.
- Strongman, Phil. 2016. "Forgotten Audio Formats: Wire Recording." *Ars Technica*. <https://arstechnica.com/information-technology/2016/11/wire-recording-forgotten-audio-format/>.
- University of Manchester. 2017. "Major Leap towards Storing Data at the Molecular Level." <https://www.manchester.ac.uk/discover/news/major-leap-towards-storing-data-at-the-molecular-level/>.
- Van Bogart, John W. C. 1995. "Magnetic Tape Storage and Handling." National Media Laboratory. <https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2017/02/pub54.pdf>.
- Wiehler, Gerhard. 1979. *Magnetic Peripheral Data Storage*. London: Heyden & Son.



Hard Disk Drives

IN THIS CHAPTER

- ▷ What is the purpose of a hard disk drive, and how does one work?
- ▷ What are the major features of a hard disk drive?
- ▷ What are the benefits and drawbacks of using hard disk drives for archival storage?
- ▷ How should hard disk drives be stored?

In the 1950s, if you wanted to run a computer program or store data, you had some options. Punch cards were heavy paper cards with holes put into them to designate the data, with the computer reading the presence or absence of a hole, usually either by detecting an electrical current through the hole or by detecting the presence or absence of light through the hole (like optical media). Similarly, paper tape was a long strip of heavy paper that, again, used holes to designate the data. This type of storage was in use long before the invention of any modern computer; punched paper was initially used to control patterns for weaving on mechanical looms starting in the 1700s.

Magnetic tape had also been invented at this point and could be used for storing computer data. As stated in the previous chapter, magnetic tape is a strip of plastic coated in magnetic material. Data is encoded using the orientation of the magnetic powder on the tape (south to north or north to south).

These methods of data storage all suffered from some serious problems, however. Magnetic tape and paper tape both use a method of finding and retrieving data known as *sequential access*. This means that every time you want to access data, the tape's reader needs to go through all of the preceding data on the tape in order to locate the data that you want to access. This is pretty slow and isn't a convenient method of storage if you want to access data quickly. It's also inconvenient for storing data that you want to change often. This sort of storage is really best for backups.

Punch cards had their own unique problems. As with magnetic tape and paper tape, it was important to run punch cards in a certain order. If you wanted to change or update data, you needed to select the correct card from a series of cards, update it correctly, and put it back in the right order. As an early method of storage, punch cards were handy in many ways, but they were highly vulnerable to human error (Press, 2016).

What was needed was a method of storage that eliminated the problems of both sequential access and human error. The 1950s saw a lot of amazing advancements in computer technology, and in 1956, the IBM computer company released a product that was quite revolutionary: a computer with a built-in hard disk drive.

The IBM 305 RAMAC computer came with a device called the 350 Disk Storage Unit. It was a series of 50 platters with a magnetic coating that rotated at 1,200 rpm (rotations per minute). It could only hold a few megabytes of data, but at that time, this was an enormous amount of storage (*Wired*, 2014). This type of computer became obsolete pretty quickly, since computers soon after moved from using vacuum tubes as the method of computing data to the more efficient transistor, which was becoming available at around this time (Press, 2016).

The hard disk drive, however, remained revolutionary. Although things are changing at the moment, at the time, built-in hard disk drives became a standard feature for computers of all kinds. Modern hard disk drives are a little different from the initial models. For one thing, they're smaller—a few inches in diameter as opposed to the 350's 24 inches (Press, 2016). They also hold a lot more data—hundreds of gigabytes as opposed to a few megabytes; and there are fewer platters—normally one to four as opposed to 50.

But the basic operation of early hard disk drives and modern ones is exactly the same, and the design allowed for a concept known as *random access*. This means that a computer program locates the data on the disk, and *where* it is actually stored on the disk doesn't matter very much. This greatly speeds up data storage and access time. Floppy disks, as covered in chapter 7, also use this method of access, and are actually very similar to hard drive disks in their operation in many ways.

The terms “hard drive” and “hard disk drive” are synonymous. The term “hard disk drive” is a little more accurate, but “hard drive” is the more commonly used term. It can also be abbreviated as HDD.

Like past computer users, people still need a way to conveniently store data on their computer. RAM chips, as discussed in chapter 2, are helpful for temporarily storing values from the CPU. However, they lose their memory as soon as they lose power, which is very inconvenient for long-term storage. So, the hard drive serves as a form of long-term storage for a computer. It can store the operating system data as well as any data that the user wishes to save, and it does not need a continual source of power to do it the way that a RAM chip does. This is so essential that modern computers *always* have some form of long-term storage as part of their physical components.

This makes hard drives a little different from other methods of data storage discussed so far in that many computers come with one already installed. This chapter qualifies this statement with the word “many” in that there is another potential option for long-term storage: a solid-state drive, or SSD. SSDs will be discussed in more depth in the next chapter.

Another way in which a hard drive is different from the other methods of data storage discussed so far (floppy disks, optical disks, and magnetic tape) is that it is typ-

ically installed *inside* a computer and is not designed to be removed when used in this way. Therefore, you may want to know more about this method of data storage not as a long-term method of storage for your archive, but to learn more about improving your equipment.

If hard drives are used for regular, everyday data storage, you might be wondering—why bother with things like optical disks or magnetic tape for archiving? If a hard drive is perfectly serviceable, stores a significant amount of data, *and* is found already installed on many personal computers, then what is the point of using something else?

The truth of the matter is that you *can* use a hard drive disk to store your archival data. There are some perks to doing so, as well. As with everything else, though, there are problems with relying on hard drives. To understand what their vulnerabilities are, it helps to understand a bit more about how hard drives function.

Hard Drive Operation

The hard drive on a typical desktop computer is pretty easy to find. It looks like a little box with some small vents in the sides that is attached by a cord to the motherboard. It typically has a sticker with helpful information on it, such as model information and the drive's storage capacity.

This little box contains some of the few parts of a computer that actually, physically move, as you'd normally expect from a machine, rather than simply creating pulses of electricity. There are quite a few things inside the casing, like a motor and circuits to buffer information to and from the motherboard. Buffering, by the way, refers to temporarily storing data so that it can be easily retrieved, processed, and so on. The important part of a hard drive, the part that holds the data, is contained on a series of disks, one stacked atop another. One to four disks is common for a hard drive. These disks are referred to as *platters*. The disks are circular and made from aluminum, glass, or ceramic—resembling, both in form and function, small records. The platters are also coated in extremely thin layers of several other substances that enable the hard drive to function, such as a layer of a magnetizable material to encode data. The entire disk is very thin and fine, in spite of all the coatings.

Like magnetic tape, covered in the previous chapter, hard drives encode the binary ones and zeros using minuscule magnetic fields, with the polarity of the fields signaling either a one or a zero. This is the reason for the layer of magnetizable material. With magnetic tape, there is a read/write head that is able to use pulses of electricity to change the polarity of a binary one or zero. Hard drives have read/write heads, too, and can do the same thing to change the data.

Exactly how the read/write head works is slightly different from a magnetic tape reader, though. In a tape reader, the magnetic tape rotates and the tape moves under the read/write head. In a hard drive, the platters with the information encoded on them are lined up on a spindle, one over another. When the spindle turns, so do the disks. Working a little as magnetic tape does, the disks have to physically rotate to the right location where the data is encoded. Unlike magnetic tape, though, the read/write head is not fixed. A little like a record player's setup, the head is attached to an arm, which can move up and down the face of a platter in order to search for the data requested by the user. This approach is much quicker than magnetic tape's method of seeking information.

The read/write head for magnetic tape is in very close contact with the tape as it turns, which is normal and desirable. On a hard drive, everything is very compact and close together, so the read/write head is also very close to the platters, but must not touch them. The spinning of the disks creates a tiny cushion of air, which the head rests upon while it reads the data. This is important, because if the read/write head touches the disks, it can cause physical scratches and damage to the platter.

As you learned in the previous chapters, the data on storage media is written in an organized way. For optical disks, the data is written in little pits and lands that are lined up along a spiral on the disk; this spiral is so important that it's physically imprinted on writable CDs to help write the information more precisely for the laser to read. On a piece of magnetic tape, the data is organized in tracks, with one track running under each read/write head, appearing just a little like the tracks on a racetrack.

The platters on a hard drive have the data physically organized, as well, but it's slightly more complicated. If the disks were square, then the logical physical organization would probably be a grid pattern, but since they're circular, it's a little more like a spiderweb.

A hard drive is divided into concentric circles of data, kind of like a dartboard. Data is written along these little circles, known as *tracks*. The data is divided further, however, into *sectors*. Imagine the hard drive being sliced up like a pie—only each “slice” is a sector. This is a little like how sectors actually work.

The way data is organized on a hard drive disk and the way it is organized on a floppy disk are actually very similar.

Typically, each sector can hold 512 bytes, so one sector is pretty small (a very thin slice of pie). Remember, at least eight bytes are needed to make a letter, so each sector could hold 64 letters, which is not very much data.

When a file gets written to a disk, it selects part of a track that has been divided into sectors. Again, think of it as being kind of like a dartboard—the data gets written to one of the sections on the board. The computer can write data in segments called *clusters*, which are segments of track that are adjacent to one another on the disk. Clusters are the minimum amount of space that you can use at a time.

Computers can keep track of the exact location of where the data is written—sort of like using a table of contents for a book, only with data files. These files “tell” the computer where to look for data, and thus the read/write head can move and the disk can rotate to the correct physical location for the data. For Microsoft products, an FAT or File Allocation Table is an older method of doing this; the newer method is to use an NTFS, or New Technology File System. For Apple products, the HFS or Hierarchical File System and HFS+ are the older methods, and Apple File System is the newest method.

Although the ideal situation is for all of the data for a file to be written in one area on a disk, data can be written to different places on a disk and not be stored all in one location. For instance, if you were playing a game, the save files of your actual game play might be scattered about the disk and may not be stored alongside the data for the game program itself. This phenomenon is known as *fragmentation*. As you learned in chapter 8, this would be highly inconvenient on something like magnetic tape, since the tape must physically rotate under the read/write head to reach the desired data, and it would take even longer to access all of the data that you needed if different parts of the same file were physically located on different sections of the tape.

The platter in a hard drive must rotate to the correct position, as well, but this is significantly less important to access time than it would be with magnetic tape. The arm holding the read/write head moves up and down the disk as it rotates, seeking out the desired information. This, in combination with the high rotation speed of the disk (the rotation speed on a typical hard drive is currently between 5,400 and 7,200 rpm, or rotations per minute), makes the access speed quite rapid in comparison to magnetic tape. As mentioned in the previous chapter, this method of seeking out information is known as *direct access*, as opposed to sequential access, the method used by tape.

This access method isn't perfect—having parts of a file scattered about in multiple locations *does* reduce the speed of access in comparison to having a file in one physical location on a disk. Most of the time, a computer will allocate data to a logical spot that makes for optimal access speed, and it's possible to improve the distribution of your files by defragmentation. *Defragmenting* a computer means that a software program will move data around on the hard drive so that it is more logically configured. The program will search for files that are fragmented, or in multiple locations, and rearrange the order of the data written to the disk so that parts of files are next to each other on the disk.

All hard drives operate in the same general fashion and have the same basic parts, but there are some differences among them.

Types of Hard Drives

If you want to use hard drives to store your archival data, it's important to be aware of the typical features of a drive and how to choose the optimal one for your archive. You may also be interested in this information so that you can choose a good hard drive for your computer, for daily use rather than long-term storage, since hard drives are easily used for both functions.

QUALITIES OF A HARD DRIVE

- Installation: internal versus external
- Size: desktop versus laptop
- Connections: SATA, PATA, USB, eSATA
- Rotation speed: 5,400–10,000 rpm
- Cache memory size
- Storage capacity

Internal versus External

There are many factors involved in choosing a hard drive, and the first of these is deciding whether you want an internal hard drive or an external one. The difference between them is pretty obvious: an internal hard drive goes inside a computer and an external hard drive attaches from the outside. Note that there are two basic types of hard drive that are designed to be used externally; “external” will be the term used here for the sake of simplicity.

You may wonder what the point of getting an internal hard drive would be, since your computer already has a hard drive inside it. Your computer can, in fact, have more than one hard drive, and people who need to store a lot of data (for example, someone who digitally edits video files) frequently install two or more in their computer. Many desktops even come with a space inside the case to hold additional hard drives so that this is easier to do. This applies to laptops, too: some laptops have empty space inside that is designed to accommodate an extra hard drive. If you want to be able to store more data, then an internal hard drive has appeal.

You can also have an external hard drive rather than an internal hard drive. External hard drives plug into a port on your computer. This has a couple of advantages for you. First, it requires no expertise at all to attach an external hard drive to a computer, whereas it requires some technical knowledge to correctly install a new internal hard drive. An external hard drive is also ideal if you want to store the hard drive someplace else; once installed, internal hard drives are best left inside the computer, but you can remove an external hard drive and put it elsewhere whenever you want. If you want to physically archive the object that holds your data, then an external hard drive is probably what you need.

It is possible to convert an internal hard drive to an external one using a hard drive enclosure. This has many potential uses, including continuing to use a hard drive for storage when the rest of the computer has failed or retrieving data from a computer that is no longer functioning.

It should be noted that there are some significant benefits to internal hard drives—for one, they are cheaper than external hard drives per gigabyte of storage. External hard drives also store and transfer data more slowly than internal ones, since internal hard drives are more directly connected to the rest of the computer (an external hard drive must connect via a port and a cable). It *is* possible to use an external hard drive for the same purpose as an internal hard drive and some instances in which it is better, but for the most part, it's best to get an internal hard drive when upgrading a computer and an external one if you need to move or disconnect the data.

Size

There are generally two sizes for modern internal hard drives—the 3.5-inch desktop version and a 2.5-inch version that is designed for use in laptops. The number refers to the size of the platters, not the casing. Larger drives usually hold more data than smaller drives and typically spin more rapidly. Choosing a size in this respect is not a big decision: get the size appropriate for the computer (a big one for a desktop and a small one for a laptop).

There are two basic sizes for external hard drives: portable and desktop. Portable hard drives are small, as you might expect from the name. Desktop hard drives are physically larger.

Portable hard drives often hold less data and spin more slowly (which results in a slower access time) than desktop hard drives. The advantage of portable hard drives is, of course, portability: they're designed to be compact and easily moved around. This may or may not be important to you, and if portability is not a factor in your situation, then the desktop version is most likely the better choice as far as usefulness to your archive.

The hard drives discussed here are aimed more at typical computer users. There is also such a thing as an enterprise hard drive. These are designed more for heavy-use applications, such as servers or workstations.

However, portable hard drives do have a subtle advantage over desktop hard drives. The disks in a hard drive need a power source to spin and read the data. With a portable drive, when you connect the drive to the computer, it's usually able to draw enough power from the computer via the connection port to function. Desktop hard drives sometimes need to be plugged into an additional power source, which can make things a bit inconvenient, since you'll have to be near an electric outlet (which may already have a computer plugged in) and then plug and unplug the device if you want to move it.

Connections

Internal hard drives plug into your computer using a cable inside the computer that connects it to the motherboard (motherboards were discussed in chapter 2). There are several different types of plugs or interfaces that have been used, and it is necessary to use a hard drive that is appropriate for the connector that has been installed. It may be possible to install a new connector, and sometimes computers have more than one type. Computers that are quite old might need some upgrading to recognize a new hard drive, since they are not designed to be able to access the amounts of data that a hard drive can store in modern times.

You are most likely to find hard drives using Parallel ATA (PATA), Small Computer System Interface (SCSI), or Serial ATA (SATA) connections. The first two, PATA and SCSI, have largely been replaced by SATA connections, but you may still encounter these types of hard drives, especially in older computers.

As mentioned earlier, external hard drives plug into a port on the outside of your computer. There are a couple of different ports that an external hard drive can connect to, with different types of USB ports and eSATA ports being the most common. USB ports were covered in more detail in chapter 2, since they can have many functions and connect to a variety of devices.

USB ports are very convenient in that they're found on nearly all modern computers, and in that computers usually have several of these ports—meaning that, if you need to use multiple devices that connect to a USB port at the same time, you won't be using your only USB port for the hard drive. The convenience offered by USB ports makes them very appealing, but they're not always the fastest way to transfer data.

An eSATA port is very like the connector for an internal hard drive, but connects external hard drives and allows for a rapid data transfer rate. However, these kinds of ports are less common than USB ports, and if your computer doesn't have one, installing an additional port requires some computer expertise. However, if your computer already has such a port or installing the necessary port and adaptor card is not an obstacle, this may be an appealing option.

Rotation Speed

The speed at which a hard drive rotates is one of the factors you might use in deciding which hard drive is best for your archive. The faster it rotates (the rpm), the higher the data access speed. Rotation speeds of 5,400 or 7,200 rpm are typical, but they can be slower or much faster—up to 15,000 rpm. If you want the hard drive for long-term storage without accessing it on a regular basis, the speed is likely to be less important to you, and slower speeds typically mean a less expensive disk drive, so you can save a little money by choosing a lower rotation speed. If you're looking for an extra hard drive for a computer that you'll be using for processing data, then higher speed may be a better choice.

Cache Memory

Like the CPU, the hard drive needs a place to temporarily store and process data before recording it. This is known as *cache memory*. The bigger the cache memory, the more efficient the hard drive is. However, if you're using the hard drive for long-term storage and don't plan on accessing the data often, this is another area in which you can save a little money, since it will matter less to you how quickly the hard drive operates. If you're adding a hard drive to a computer for function, a large cache might be something to consider. Similarly, when purchasing a hard drive, the manufacturer may list the access time, which refers to how quickly it can find a file. Again, a quick access time might not matter to you for storage, but might for daily function.

Storage Capacity

While you can omit some features to save money with a hard drive by not getting the fastest, most powerful device out there, the number for which bigger is always better is the hard drive's capacity. This will be noted in gigabytes (GB) or terabytes (TB). Remember, a terabyte is about a trillion bytes, whereas a gigabyte is about a billion bytes, so a 200 GB hard drive is a fraction of the size of a 1 TB hard drive.

Hard Drives for Archiving

Hard drives aren't really designed for the long-term data storage that an archive typically needs. They're designed to be able to change their data quickly and to be updated frequently, with the user writing new files or erasing them regularly. In fact, a hard drive can start losing its data if not used this way. However, you can certainly use them to store data long-term if you desire, and there are some advantages to using hard drives for your archive's main method of data storage.

ADVANTAGES AND DISADVANTAGES OF HARD DISK DRIVE STORAGE

Advantages:

- Large storage capacity
- Fairly inexpensive
- Do not require special equipment to read data
- Simple to use

Disadvantages:

- Short lifespan
- Difficult to assess condition
- Prone to reading and writing errors, as well as virus attack
- Easy to break
- Needs refreshing

Advantages

In many ways, using a hard drive for storage is economical. It's not as cumbersome as using optical disks, such as CDs, since it could take hundreds of CDs to store the same amount of data as a single hard drive—even a hard drive that is fairly small by today's standards. Optical disks are pretty inexpensive per gigabyte compared to hard drives, but convenience is important to the efficiency of your archive, and efficiency is always a cost saver: no one wants to take the time it might involve to find a particular file on a particular disk among hundreds when you could find the file in seconds on a hard drive via a keyword search.

Hard drives aren't as inexpensive per gigabyte of data as magnetic tape, either, but if you don't need to save huge amounts of data, hard drives make for a nice compromise because you don't have to buy a reader for them the way you would with magnetic tape. In other words, hard drives can be efficient if you need to store more data than would be convenient on optical media (a very small archive), but less than the amount that would make using magnetic tape worthwhile (a very large archive).

Using external hard drives is pretty simple, too—always an advantage—because your staff won't need much training to learn how to use one. As mentioned earlier, these drives simply plug into a port. Installing an internal hard drive, if you're interested in that, is a little more complex, but is still something that you can often do yourself if you're careful. Using an installed internal hard drive, of course, requires no expertise whatsoever.

Another way in which you can use hard drives is as part of a backup system. A RAID array (Redundant Array of Inexpensive or Independent Disks) is a series of hard drives configured to work as a single unit. There are a few ways to set up a RAID array, but this is potentially a way to use multiple hard drives to back up data. When configured as a backup system, the same data is written across multiple hard drives so that if one drive fails, the same data is on the other hard drives. Note that an entire RAID array can fail in various ways, so such a device shouldn't have the only copy of data files.

Optical disks, like CDs and DVDs, have an advantage in that they're not a unique storage method: you can go into many stores, even ones that don't sell electronics specifically, and buy perfectly serviceable blank disks. This is an advantage that they share with hard drives. External hard drives are pretty common, and many stores, such as those offering office equipment, sell external hard drives that will be perfectly serviceable, although you'll certainly want to get the best quality one available if your budget allows it. Magnetic tape, on the other hand, is more of a specialty item that's harder to find.

If you want to make your collection available online and not simply store the data passively, data stored on hard drives lends itself pretty well for this purpose. You can use multiple hard drives and store data within a server computer by connecting these hard drives, which then can be configured to allow access to the data through a network. This means that multiple computers could access the same hard drives. It should be noted, however, that this is also possible to do with other methods of data storage, but the access time may not be as good.

Disadvantages

Hard drives are highly mechanical in nature. While the important parts are the platters that contain the data and the arm with the read/write head that writes and accesses the

data, a hard drive consists of many parts. It has a motor, bearings, lubricants, and more. The more parts there are, the more that can go wrong. The motor can go bad; the lubricants can evaporate; the air intakes can fail, letting dust into the casing and causing damage to the platters as they spin, and many other problems can occur.

Essentially, hard drive failure is inevitable, and hard drives often have a short lifespan relative to other storage methods. However, this doesn't mean that you should automatically rule out hard drives as a storage medium, since all methods of storage will wear out or decay eventually—it's a matter of whether or not the benefits are worth the risks that are involved.

Hard drives can fail without warning. They'll be working one moment—and not at all the next. There are a number of ways that failures can happen. For instance, the platters, which rotate at extremely high speeds, can bump into one another. This is not supposed to happen during normal use, but, of course, accidents happen. It's also possible for the head, which reads and writes to the disks, to bump into the disks. Normally, there is a small cushion of air between the disks and the read/write head, which protects them from each other, but if this cushion is disrupted somehow while the disks are still spinning, the lack of cushioning can create physical damage to the disk below, grinding away the data. These particular problems can generate audio cues that indicate something is wrong, since hard drives make little sound when working properly, but hearing these cues may mean that it's already too late and your data has been lost due to physical damage on the platters. It's also possible for parts of a hard drive to go bad, but the drive itself to still be functional, or for sections of data to be lost, but not all of them.

Assessing a hard drive's condition can be difficult in comparison to other methods of data storage discussed so far. The platters and all of the mechanisms required for operation are concealed within a protective case, and opening the case exposes the drive to the environment. This can damage the drive. With a CD, for example, you can visually examine the disk for things like cracks or discoloration, or with a magnetic tape, look at it for rust.

Hard drives are susceptible to a phenomenon known as "bit rot." This phrase refers to the decay of the binary coding in a storage medium, and can mean different things when referring to different types of storage media. All methods of media storage are vulnerable to this in some way. With hard drives, the bits of magnetic encoding are so small that they can be erased by temperature fluctuations.

Similarly, sectors on hard drives can go bad for a variety of reasons. Modern hard drives are even manufactured with this in mind and have sectors that are held in reserve for whenever sectors on the hard drive start going bad. Damage to sectors can be caused by small errors in writing the data to the hard drive, which leads to an inability to read the data, or can be caused by mechanical issues, like dust in the hard drive. Errors with writing can be repaired by erasing the disk (filling it with binary zeroes); physical damage can't be fixed so easily. It's also possible for viruses to attack a hard drive and create false readings, making it appear as though a hard drive has bad sectors (another problem that can often be fixed).

Hard drives don't withstand physical abuse well; a fairly short drop could severely damage a hard drive. In contrast, if you were to drop a CD on a rug, for instance, it might bounce and still be readable. Trying this, however, is not recommended.

Because hard drives are quite common and have been in use for some time, they do have some of the same advantages as magnetic tape in that they are well understood. However, because they aren't really designed for archival storage and are supposed to be

used for everyday data storage, their efficacy as a long-term storage device has not been explored as well as that of magnetic tape or even optical media.

Hard Drives for Archival Storage

As for magnetic tape and optical disks, the environmental conditions under which you store hard drives can make a difference in how long they last.

Temperature

Hard drives can actually operate safely at a wide range of temperatures, and a high temperature doesn't have a lot of impact on how long a drive lasts unless it's very hot—more than 125 degrees Fahrenheit or about 52 degrees Celsius. However, computers *can* operate at temperatures that exceed this, so what you need to do to ensure that your hard drive is safe during operation is to make sure that your computer is not overheating and that your hard drive has good ventilation and airflow to cool it adequately (Jacobi, 2007).

However, it's best, in general, to keep electronics cool. Very cold temperatures are not very good for hard drives, though, and a hard drive that is cold needs to acclimate to room temperature before use (Western Digital, 2002).

With old hard drives, freezing the drive was a possible way to get stuck disks free (the metal inside would constrict enough to let the disk move freely). This is not advisable with newer hard drives, though, as freezing can lead to the formation of ice crystals that can damage and corrode the disks (Hachman, 2016).

Humidity

It was long thought that high temperatures were a major cause of hard drive failure. While extreme temperatures are bad for hard drives, humidity is a much better predictor of hard drive failure. Keeping the humidity low is best to prevent hard drive failures (Harris, 2016).

Location

Though you should take care with choosing your storage location, hard drives are fairly resistant to magnetic fields—unlike magnetic tapes, which are vulnerable—though both are forms of magnetic data storage. When you store a hard drive, it's typically best to keep it flat (the disk inside should lie horizontally), as opposed to optical disks and magnetic tapes, which are best stored vertically, in book fashion. Don't stack hard drives on top of one another or put anything heavy on top of a hard drive. If you were to store multiple hard drives, for instance, you would have to put them next to one another on a shelf, or have a storage setup in which each drive has its own shelf or slot.

Handling

Hard drives are designed to be used inside a computer and not moved around, and are somewhat delicate in that respect. Handle hard drives carefully; be sure to never drop

or shake a hard drive, and take precautions against static electricity. The case will help protect it, but it's always a good idea to be cautious. Don't open up the casing, as this will expose the hard drive to things like dust or static electricity.

In the case of an external hard drive, it's important that the data has finished saving to a hard drive before you remove it. Failure to do so can cause damage to the drive itself, and though this is uncommon, it can be serious and result in permanent damage. Computers come with a program designed to "eject" a drive like an external hard drive, which ensures that the data has completed writing before you remove the drive. This same program is useful for devices using flash memory, too, which will be covered in the following chapter. When you use an external hard drive, you should always take care before removing the drive.

Use

Actively used hard drives are thought to last around ten years. Under archival storage conditions, however, a hard drive may last longer—up to thirty years. Like magnetic tapes, this set of conditions is not much different from room temperature conditions, and is within a pretty standard range for storing archival materials in general, regardless of whether or not they are electronic in nature (Williams et al., 2008).

IDEAL STORAGE AND HANDLING OF HARD DRIVES

Operating Temperature: Below 125°F (52°C), above freezing

Humidity: Low

Storage:

- Store with platters horizontal to storage surface.
- Do not put objects on top of hard drives.

Handling:

- Handle carefully; do not shake or drop.
- Prevent static during transport.

Hard drives are a bit different from optical disks and magnetic tape in that they're not only designed to be written and rewritten to over and over again, they actually work best when used in this way. They will lose data or generate errors over time if left alone, untouched. With the other media covered so far (particularly magnetic tapes), the less they are used, the better. Not so with hard drives. It's in your best interest to plan for this if you use hard drives for long-term storage. You will need to plan to either move the data from one disk to another or at least refresh the disk periodically.

It's also a good idea to check the disk more often than this to detect errors (e.g., every year or so). Though not perfect (it's possible for software to miss problems), some software programs can alert you to errors and problems; they can signal that you should take action to save the recorded data. There are a few ways to do this. One is to use Self-Mon-

itoring, Analysis and Reporting Technology, or SMART, which is something that comes standard on modern hard drives. This is a feature that analyzes the physical attributes of a disk and can alert you to issues like problems in the motor. Your computer may also come with programs, such as CHKDSK, that look for data errors or bad sectors. If you purchase an external hard drive, it may come with diagnostic software, a feature that may have appeal to you when you're trying to make a decision.

Key Points

- Hard drives are a form of magnetic storage and have been a major component in desktop- and laptop-style computers for decades. They offer a convenient, inexpensive form of long-term storage.
- Internal, external, and portable hard drives are available and come in two main sizes.
- A computer can have multiple hard drives to increase storage space.
- Hard drives are rather delicate and need to be used in order to retain their data; if used externally, they will require careful handling and periodic refreshing.
- Hard drives are still relevant due to their convenience and low expense, but may become obsolete in the future due to advancements in their major competitor, the solid-state drive.

While hard drives are the storage medium of choice for many computers, it's not the only method of storing data long-term within the device itself. The solid-state drive, using a technology known as *flash memory*, is replacing the hard drive in many computers. While flash memory technology is one of the most expensive choices for data storage, it has many qualities that make it a desirable storage method for archiving and can be used for both internal and external storage.

References

- Hachman, Mark. 2016. "That Old 'Freezer Trick' to Save a Hard Drive Doesn't Work Anymore." *PC World*. <https://www.pcworld.com/article/3035017/that-old-freezer-trick-to-save-a-hard-drive-doesnt-work-anymore.html>.
- Harris, Robin. 2016. "Heat Doesn't Kill Hard Drives. Here's What Does." ZDNet. <https://www.zdnet.com/article/heat-doesnt-kill-hard-drives-heres-what-does/>.
- Jacobi, Jon L. 2007. "Hard-Drive Failures Surprisingly Frequent." *PC World*. <http://www.pcworld.com/article/131168/article.html>.
- Press, Gil. 2016. "IBM Gave Birth to Disk Drives 60 Years Ago: This Week in Tech History." *Forbes*. <https://www.forbes.com/sites/gilpress/2016/09/12/ibm-gave-birth-to-disk-drives-60-years-ago-this-week-in-tech-history/#2f3c4daf5185>.
- Western Digital. 2002. *3.5 Inch Hard Drive Handling Guide*. <http://products.wdc.com/library/other/2579-001027.pdf>.
- Williams, Paul, David S. H. Rosenthal, Mema Roussopoulos, and Steve Georgis. 2008. "Predicting Archival Life of Removable Hard Drive Disks." LOCKSS. https://web.stanford.edu/group/lockss/resources/2008-06_Predicting_Archival_Life_of_Removable_Hard_Disk_Drives.pdf.
- Wired*. 2014. "Tech Time Warp of the Week: The World's First Hard Drive, 1956." <https://www.wired.com/2014/01/tech-time-warp-ibm-ramac/>.



Flash Memory

IN THIS CHAPTER

- ▷ What is flash memory, and where is it commonly used?
- ▷ How does flash memory work?
- ▷ What kinds of devices use flash memory, and what are the differences among them?
- ▷ What are the benefits and drawbacks to using flash memory for archiving?
- ▷ What are the features to look for when purchasing flash memory?
- ▷ How should flash memory devices be stored?

In the previous chapter, you learned about the hard disk drive, a revolutionary data storage method that allows for rapid data access and is capable of storing a large amount of data. The merits of hard drives are so significant that they continue to be an important invention today and will continue to be for the foreseeable future.

Hard drives have problems, though. They are limited in how small they can be, and this is important for devices that are designed to be portable, like smartphones. Although improvements in the technology have enabled them to go from the original 24-inch disks to ones that are only a few inches across, this is still pretty big for storing data on a portable device, and hard drives aren't really getting smaller. Instead, enabling hard drives to store more data in the same amount of space is currently a more practical goal for engineers.

Hard drives are also very delicate. They are often made from fragile materials, such as ceramic or glass. The reader arm that detects data on the disk can actually damage the disk if things go wrong. Although the technology has improved quite a bit over time, hard drives can still easily be damaged by dropping or shaking.

For a very long time, hard drives were the primary data storage method for personal computers, along with RAM and ROM chips, which each have their own limitations. ROM chips, as you might remember from chapter 2, can also store data for a long time

and are necessary for starting up a computer, but are not designed to be changed the way a hard drive can be (some types of ROM chips cannot be changed at all after manufacture). RAM chips, necessary for storing data as programs run, can change their data and offer high access speed, but lose their data when there is no electricity available. A data storage method that did not have these limitations was needed.

In the 1970s, the Japanese electronics company Toshiba was concentrating their efforts on the development of RAM chips. Again, RAM chips only retain their data as long as there is power available. This is why, if a computer is accidentally turned off while in use, any data that was not saved to the hard drive is lost. Some programs make temporary backups to prevent this type of data loss, but that data is on the computer's long-term storage. As in the past, RAM chips today still don't retain data without electricity.

One of the engineers for Toshiba's RAM chips, Dr. Fujio Masuoka, was assigned the task of improving this technology, which he did with the development of one megabit DRAM (a specific type of RAM chip). However, Masuoka wanted to create a better method of storage—something that would retain data without power, unlike a RAM chip, and lacked the limitations of other types of long-term storage, such as floppies and magnetic tapes (Gregersen, n.d.).

Working without permission from Toshiba, the technology he eventually developed was called *flash memory*, and while it has its own limitations, it is essentially what Masuoka had in mind. This method of data storage retains data without power, and it allows for fast random access, like a hard drive, but without the delicacy and size limitations. This allows for devices using flash memory to be quite portable.

Flash memory is the newest of the data storage methods discussed in this book, with Masuoka presenting NAND-type flash memory at the International Electron Devices Meeting in 1986 (Katz, 2012). What NAND-type flash memory is will be explained later in this chapter.

Items using flash memory storage are extremely prevalent in 2020, and this type of data storage is essential to the function of a number of everyday devices. Chances are good that you own a device (or even several devices) that depend upon flash memory for their operation. Smartphones, tablets, digital cameras, MP3 players, and nearly any other type of modern electronic device that needs to be portable and to have memory that can be changed uses flash memory.

Unlike optical disks, floppy disks, magnetic tape, or hard drives, flash memory doesn't refer to a specific device or physical method of storage. While flash memory is, of course, physical in nature (all data storage is), it's better to think of it as a *method* that several different, but related, devices use to store electronic data.

There are a lot of different devices that use flash memory, but to make things a little easier, consider flash devices as coming in three basic formats. One is the *solid-state device*, also known as a *solid-state disk*. These terms refer to the same thing and have the same acronym, SSD, which can help you when making a purchase.

Another is the *flash drive*, also known by several other names, like thumb drive, jump drive, pen drive, or memory stick. These are small and portable and have a connector attached that typically plugs into a USB port to transfer data. They are far less delicate than optical disks, and are useful for moving data from one computer to another or for small amounts of external storage.

The third type is the secure digital, or SD card, which is typically used for temporary storage in cameras and similar devices. Like flash drives, these are small and highly portable.

These three devices are slightly different from each other and have different sizes, both physically and in terms of storage. The way that these devices generally operate and how they are useful to you is the same, but there are some important differences that you should be aware of when it comes to archiving. It is most likely that the SSD, which has the highest storage capacity of the three, is the most relevant to you.

Hard drives and flash memory are going to be compared a lot in this chapter because they are comparable to one another in a lot of ways, having similar uses and storage capacities. While a hard drive won't do for small devices, hard drives and flash memory can both be easily used as the long-term data storage method for larger devices, such as laptops and desktop computers. They also are both well suited for external data storage, and so it is possible that you may need to make a decision between these two methods of data storage.

While flash memory can be used for the same purposes as a hard drive, that's where the similarities end. Flash memory is a technology unlike any other type of storage method, and how it works is a little more complex than the other methods discussed so far.

How Flash Memory Works

The way that flash memory works is harder to visualize than any of the other methods explored so far. This is in part because it's not like any familiar forms of technology, such as a record or a cassette tape. It is most similar to other devices that use transistors. As mentioned in the previous chapter, some early computers were transistor based. RAM chips also use transistors. Though transistors play an important part in modern electronics, chances are good that this technology doesn't seem very familiar to you.

A *transistor* is a device that can both conduct electricity and resist it. Because it's necessary for a transistor to be adaptive in regard to electricity, they're made of semiconductive materials that are neither good nor poor conductors of electricity; that is, they *can* conduct electricity, but aren't very efficient at it. To contrast, think of the copper in an electrical wire, which conducts electricity very well, or of rubber, which doesn't conduct electricity at all and can instead insulate a conductive material. In modern times, the semiconductive material of choice is typically silicon.

RAM chips operate in part by using a system of transistors and capacitors; capacitors are devices designed to store electricity. A RAM chip's capacitors can store a tiny amount of electricity; the binary values are stored as either the presence or absence of electricity. The transistor controls whether or not electricity is in the capacitor; remember, a transistor is capable of both conducting and resisting electricity, so it can change the value of the capacitor. The simpler and more common type of RAM, DRAM, has one capacitor and one transistor; each pair forms one memory *cell*.

However, as stated, a RAM chip needs electricity, and the reason for this is that the capacitors will lose their charge over time, and so the value they hold needs to be refreshed. Flash memory actually works extremely similarly to the way a RAM chip works. The big difference between them, however, is the fact that a flash memory cell *can* store a value without electricity. The electrons that will leak out of a RAM chip after a short time will stay in a flash memory cell and enable it to retain a value for long periods of time.

Flash technology works by recording whether or not electrons can flow through its transistors. If the transistor can conduct a current, then it reads as a one, and if it can't,

then it reads as a zero. In a device using flash memory, “one” is the default value; erased cells have a value of “one,” as well. When you want to save data to the device, then it has to change some of those ones to zeroes using a pulse of electricity (Hruska, 2019). This is a bit unusual, since the default value is typically zero for other storage devices. Each unit in a device using flash memory (one of those miniscule transistors), is referred to as a *cell*.

Each cell has four basic parts: a floating gate, a control gate, a source, and a drain. The source and drain move electricity through the cell. While there are two gates, the floating gate is the one that stores electrons, which changes how conductive the cell is, thus making the entire cell read as a one or a zero based on the level of voltage detected running through the cell. The floating gate is surrounded by insulators, which help keep the electrons in place. This is necessary, because otherwise, electrons would leak out of the gate and erase the cell’s binary value.

When a charge is applied to a cell, electrons are injected into the floating gate, making it be a binary zero. It’s erased, or changed back to a one, using another pulse of electricity. Ideally, if left alone, these electrons stay trapped in there indefinitely, and can store information without the need for a constant power source, just as optical disks, magnetic tape, or a hard drive do not need a power source.

Flash memory is more complex than other kinds of memory in that it can potentially have more than a yes/no condition. That is, a cell can potentially hold more than one value. In contrast, an optical disk has only two conditions: brightly reflected light or less brightly reflected light. Data storage typically has two conditions because it mimics the two possibilities in binary—a one or a zero—and binary is the essential language of computers.

FLASH MEMORY CELLS

- Single-level cells store one bit per cell and are the fastest, most accurate type.
- Multilevel cells store two bits per cell and are the most common type, with median qualities.
- Triple-level cells store three bits per cell and are the slowest, but most compact, type.

A *single-level cell*, or SLC, is like this, too. It can only hold one bit of data at a time, either a one or a zero, and the circuitry within the cell is only able to detect two thresholds for the voltage of an electrical current. However, there is such a thing as a *multilevel cell*, or MLC, which can hold two bits. The device is designed to detect four different levels of voltage for a cell, and thus a value of 00, 01, 10, or 11, depending upon the level of charge stored in the cell. This gets a little complicated. In an SLC, the cell can detect either conductivity or a lack of conductivity. In an MLC, the cell can detect no conductivity, a little conductivity, more conductivity, and a lot of conductivity and can assign different pairs of binary values to each state rather than just one binary value (Hruska, 2019).

There are also *three-level cells* (also known as *triple-level cells*), or TLCs, which can hold three bits per cell and have eight potential threshold values, and *quad-level cells* (QLCs), four bits with sixteen potential values. There are some benefits and drawbacks

to each storage method. For instance, a cell that can contain more than one value needs a more precise application of electricity in order to write the correct value to the cell. This results in an increase in time for writing to the cell, so while single-level cells (SLCs) can't store as much information per cell as cells with more levels, writing to them is quicker (Cornwell, 2012).

Differences between Flash Devices

Flash drives, Secure Digital cards, and solid-state drives all use the same technology and store information in the same general method. However, types of devices are not *exactly* the same, so it's important to know what those differences are.

TYPES OF FLASH MEMORY DEVICES

- Flash drives: small and portable, useful for transferring data
- Solid-state drives: larger with a large storage capacity; comparable to a hard drive
- Secure Digital (SD) cards: very small and compact, normally used for additional storage in small devices like cameras

Flash drives are probably the version of flash memory that you'll be most familiar with. They are also sometimes called thumb drives, and both of these names are helpfully descriptive. "Flash" can refer to the storage technology as well as the speed with which it operates. These devices are, as mentioned earlier, designed to be put into the USB port of a computer, and a computer can begin reading the information encoded on the drive quite rapidly. "Thumb" refers to the size of the drive; these devices are often about the size of a thumb or smaller. This makes them highly portable and convenient for transferring information.

Solid-state drives or solid-state disks (SSDs) are another type of flash memory device. The term "solid-state drive" is also descriptive, referring to the fact that these devices have no moving parts, in contrast to the other storage methods explored so far. However, flash drives don't have moving parts, either, so these names don't quite describe the difference between them (although solid-state storage is far less portable than a flash drive). SSDs are physically larger than flash drives and are somewhat less portable and more delicate than flash drives. SSDs have the most memory of your options, but are also the most expensive.

Like hard drives, SSDs can be both internal (designed to be installed in the computer) or external (designed to be easily removed). They can also be used as a replacement for a traditional platter-style hard drive, and there are many reasons that installing an internal SSD and/or replacing a hard drive with one might be appealing. Because flash memory is so quick, it is possible to improve the speed at which your computer stores and retrieves files by replacing a hard drive with an SSD.

Secure Digital, or SD cards, are very small, thin cards. There are several kinds of SD cards, such as the miniSD, microSD, miniSDHC, and more. They are all slightly different sizes, and SD cards can range from around the size of a postage stamp to about the size of a fingernail.

As far as portability, capacity, and capabilities go, SD cards and flash drives are very similar to one another. However, there is one big difference that will be important to you. While some computers have a built-in SD slot, not all do, and most only have one SD slot, which will not accommodate every type of SD card. Flash drives, on the other hand, plug into any USB port. SD cards are more commonly used in devices like digital cameras, which can then connect to a computer via a cord that plugs into a USB port, as opposed to using an SD card as a method of general data storage. However, it is possible to buy SD card readers, which can plug into the computer as a peripheral, or to purchase adapters to make a slot designed for a large SD card accommodate a smaller one.

Solid-state drives (SSDs) are bigger and heavier than flash drives and Secure Digital (SD) cards, which makes the other two devices the optimal choices for true portability. If you want to be able to transfer files between computers, the ease of use, portability, and lower expense of flash drives is going to appeal to you. If you're looking at flash technology as a method of long-term information storage, then an SSD is more likely to be what you need.

Flash Memory for Archiving

Like any other storage technology, solid-state storage has some benefits, as well as some drawbacks, for you.

ADVANTAGES AND DISADVANTAGES OF FLASH MEMORY STORAGE

Advantages:

- High durability
- Works well in less-than-ideal conditions
- Does not require special equipment to read data
- Simple to use

Disadvantages:

- Difficult to assess condition
- Cells will eventually burn out
- Needs refreshing
- Expensive
- Prone to loss or theft

Advantages

One of the greatest benefits of solid-state technology is the lack of moving parts. With magnetic tape, for example, simply reading the tape will cause it to wear down over time and can eventually break it. Stress caused by moving the tape between reels is an issue. Regular hard drives are susceptible to a number of mechanical issues due to a high number of moving parts. Even a CD can potentially shatter inside the reader that spins the disk.

Solid-state storage has none of these problems. The lack of moving parts makes this technology extremely durable. You can even drop flash drives and still read the data. That's impossible with a hard drive and definitely inadvisable for optical disks and magnetic tape. Nothing will change the values encoded in the transistors but a change in their voltages. While it is certainly possible for devices using flash technology to fail—and they *can* physically break—they will withstand a lot of abuse in comparison to your other options. There are also flash memory devices that are specifically designed to withstand a number of extreme environmental conditions.

Along with general physical durability, flash memory is typically very tolerant of less-than-ideal operating conditions. Even a device not designed for this can withstand some harsh conditions. As a real-life example, in 2008, a couple accidentally dropped a digital camera into the ocean from a cruise ship while on vacation. It was later retrieved by chance in a fishing net, whereupon the fisherman who found it was able to retrieve some of the images inside and posted them online so that the camera's owners could be identified, which they eventually were (BBC News, 2011). This means that the device was able to withstand not only water, but also corrosive salt water, and both withstood it for some time and retained enough information for complete files to be retrievable by an average computer, which is a pretty amazing feat.

While water *could* warp an optical disk or rust a magnetic tape or hard drive, oftentimes flash devices will still work after exposure to water (assuming you dry them completely before operation). These devices can sometimes operate at very high humidity—80 percent or higher—and at extreme temperatures, in some cases below freezing temperatures *and* above 150 degrees Fahrenheit (Kingston Technology Corporation, 2012). If you lack the ability to store devices for your data in a humidity- and temperature-regulated environment, this is a highly appealing quality of flash memory, and if your archive happens to be someplace very humid, swapping SSDs for hard drives could greatly improve your general computer function and not just be helpful for archiving.

In addition, if you are particularly concerned about the prospect of a natural disaster, like an earthquake or a flood, the durability of flash memory devices may appeal to you, particularly that of devices designed for extreme conditions.

Along with durability, speed is the other major appeal of flash memory. No moving parts means that you don't have to wait long to access your data. Though it's only a tiny amount of time, you do have to wait for the platters in a hard drive or an optical media disk to rotate to the right location, and for the arm or the laser to seek out the information you want. Magnetic tape has to physically move to the right spot on the reel for information access. Though it doesn't offer *instantaneous* information access, flash memory is very quick in comparison to your other choices.

Unlike magnetic tape, which requires special equipment to access the information, it's typical for devices using flash memory to plug into a USB port; these ports were covered in chapter 2. USB ports are standard on most modern computers and

provide an easy way to use external devices to communicate with the computer. In essence, all you need to do is plug the device into the port and the computer will automatically access the information.

Like optical media, this technology is fairly familiar, and flash drives in particular are very common. This means that, also as with optical media, you won't need much in the way of special training for your staff if using removable media (as opposed to SSDs designed to be installed into a computer).

Though it may not matter to you as far as its ability to store information goes, flash memory also uses less power than your other options, since it does not physically have to move parts. This is another appealing aspect, and is important to applications that might require storing a lot of data (such as a data center).

While flash memory's benefits are numerous, it's not a foolproof method of data storage, and there are many reasons to not use it or situations in which it's not a good choice for your archive.

Disadvantages

Flash memory does have some drawbacks. There's no warning that you can use to detect whether the drive is going bad other than any warnings the device itself can provide you. With tape, you may be able to visually detect that there is stress on the tape, smell funny odors, or otherwise see signs of decomposition. With optical media, you can sometimes see cracks or scratches or notice warping. Normal hard drives can give you audio cues that something is wrong (though audio cues may mean that the damage is already done, as well). The nature of solid-state drives, since their insides are hidden and nothing moves, conceals problems from the user.

There is currently such a thing as a hybrid hard drive. This is a hard drive that also contains a small SSD. Files that don't need to be accessed often go on the hard drive part, and files that require rapid access or are accessed more often go on the SSD. What goes onto which portion is managed by a software program. This is something that might be considered by someone who wants the rapid access of an SSD but the low expense of a hard drive.

Flash memory devices do wear out, and this is inevitable. Those cells can only be written to and erased so many times before they don't work anymore. The high-voltage pulses used to erase the drive will also eventually ruin it, burning out the cells and making it impossible to keep electrons in the floating gate. Electrons can also become trapped, creating false readings or measures of resistance in the cell (Cornwell, 2012). However, it should be noted that, with technology continuing to advance, your archive may decide to purchase new devices for memory storage before this even becomes an issue with a device using flash memory, especially if you don't write and rewrite to the device often, since, while the cells do burn out, they can be erased and written to thousands upon thousands of times before this happens (Ngo, 2017).

Similar to the problem of burned-out cells, electrons can leak from the cells from disuse, effectively erasing the device. Flash memory *doesn't* need a constant source of electricity to function, but it *does* need refreshing periodically, just as a traditional hard drive. While a manufacturer may state how many times a device can be written to, stating how long a device will retain data is less common (Cornwell, 2012).

Flash memory can also become very pricey in comparison to other methods of storage. A simple flash drive is relatively cheap and convenient, but larger storage devices can be quite expensive, especially in comparison to hard drives and magnetic tape. It's very likely that the price of flash memory will continue to decrease over time, but it will probably remain the most expensive of your options for some time.

While flash memory is a highly desirable form of data storage, it's important to remember that this can easily change in the future as engineers continue to try improving on storage technology, creating faster memory that lasts longer. A 2018 development is 3D XPoint memory, which has similarities to RAM technology and flash memory, but stores much more data than a RAM chip, is nonvolatile (retains data without power), and is faster and more durable than flash memory (Bright, 2018). It's unknown what will become of this technology in the future, but it or another innovation could overtake flash memory if the benefits are significant.

Although flash memory offers speedy access and uses less power, hard drives are still commonly used in operations that process a lot of data (such as a data center) because the price per GB of data is so much lower.

If you purchase several flash devices, you may have a hard time determining how to distinguish one from another. For example, optical disks and the largely obsolete floppy are both designed to be labeled (a floppy has a sticker for writing information, and writable optical disks often have lines printed on the face for this purpose). An external SSD is large enough to label with a sticker, but flash drives and SD cards are quite small and may be physically identical. This can be particularly problematic with SD cards, which are tiny and have no room at all for labeling. Flash drives do often have a ring, hook, or similar object on one end, which enables them to be attached to a lanyard, key ring, and so on, which may help you overcome this issue.

Another issue that is a result of human error is losing the device. If you use an external SSD, then this probably won't be an issue because the device will be too large to easily forget about. However, flash drives are well known for their portability and can be forgotten while still in the USB port, put into a pocket and carried off, or accidentally dropped and lost in that way. If you use SD cards, often used for cameras, and take them out of the camera, they are even easier to lose due to their tiny size. As noted earlier, many SD cards are about the size of a postage stamp or smaller.

Similarly, because this technology is expensive and desirable, you may have a problem with theft. This is particularly true with solid-state drives, since they are very expensive. Again, flash drives and SD cards are much more common and will be less vulnerable to theft, but it's still a possibility. In addition, because they are so common, someone could carry off a flash drive or SD card by accident, not realizing that they have the wrong device or simply forgetting that they have it.

If your archive contains information that you would prefer to restrict access to, this portability and ease of access can pose an additional problem. For example, if someone wanted to get the information that you have stored on a tape by taking a tape, then it would be problematic for them because computers don't come with tape drives. However, the vast majority of computers have a USB drive that will be compatible with external SSDs and flash drives. Though there are ways to protect your data, it will lack the advantage that tape has in this regard.

Flash memory as a *concept* has been around for quite a while, but flash memory devices as they exist today are pretty recent. This is another possible drawback. It's difficult to say whether this type of memory will perform well for years to come or not. As another difficulty, this technology is quite popular for its convenience and quick data access speed, so innovations are being made all the time to take advantage of a market longing for ever better flash memory devices. The product you purchase today could be outstripped by one you could buy next year, or even six months from now, so it's also difficult to determine how good your device is for archiving in comparison to one that is older or a newer device. As with any other method of data storage, the lifespan of this type of device varies and is subject to how it is handled and stored, as well as how often it is used.

How do you decide if flash memory is the best option for you? The expense of flash memory and the rapid changes in the technology may make it a poor choice for your main data storage, especially if you need to store a lot of data. However, because it's so durable and withstands less-than-ideal conditions so well, it's an excellent choice as a secondary backup for your data, especially if, as mentioned earlier, you're concerned about protecting your archive against a disaster of some kind. It's also a good option if you can't maintain ideal archival storage conditions, since it won't degrade as readily as some of your other choices.

Flash memory is also a very good option if you need to store more data than could be stored on a simple optical disk, but less than the amount that would start making the low price of hard drives or magnetic tape appealing.

Flash memory devices, and flash drives in particular, are also very good for sharing data, since these devices are noted for their rugged nature and portability, whether you need to share data with other departments, in the case of a university archive, or between archives or libraries. It can even be a great way to move information over shorter distances, from computer to computer.

Purchasing Devices

There are several terms that you might encounter regarding the construction of a device using flash memory. As mentioned earlier in the chapter, the terms single-level cell (SLC), multilevel cell (MLC), triple-level cell (TLC), and quad-level cell (QLC) all refer to how much data can be stored in a single cell. An SLC can store one bit per cell. However, it is possible to store more than one bit per cell. MLCs store two bits, TLCs store three, and QLCs store four, and they do this by having more than one possible state (other than "on" or "off") in a single cell. Of the three, SLCs are much faster, use less power and generate less heat, last longer, and are more durable in general than the other kinds. MLCs, however, are less expensive and can store much more data in the same amount of physical space as an SLC (Hruska, 2019). TLCs and QLCs store the most data for the size of the device and are generally the cheapest option.

Flash memory devices will wear out if written to enough times. The number of times a flash device can be written to is expressed by the manufacturer as its endurance. Put simply, a device's endurance is the total amount of data that can be written to the device over its lifespan. For example, if you have a 5 GB video written to a drive and erase it, then add a new 5 GB video, that is 10 GB of total memory written to the device. It is not uncommon for current SSDs to have an endurance rating that amounts to terabytes of

data. Larger devices with more memory will have a higher endurance rating than devices with less memory (Ngo, 2017).

As another consideration, *reading* a device does not affect the cells in the same way writing does, and so drives that are not going to be written to often will last much longer. In addition, you may be interested in features that alert the user about the state of the drive and whether the cells are still in good condition.

Related to a flash drive's lifespan, the term "wear leveling" refers to a method of using the cells in a device evenly. That is, a certain block of cells won't get written to and erased more than the other cells, thus wearing out that block before the others. There are two types: dynamic and static.

Two more terms that you might see regarding flash memory are NOR or NAND. Unlike many other computer terms, these are not acronyms. What they refer to is the exact construction of the cells in flash memory, as there are several ways to construct flash memory cells. The type of cell is named after a logic gate, as the transistors resemble computer logic gates, such as AND, OR, or NOT gates (NAND and NOR mean "not AND" and "not OR," respectively, and essentially have the opposite function of AND and OR gates). Exactly what logic gates are and how they work is beyond the scope of this book, but basically, logic gates exist within the computer to make calculations.

There are four basic kinds of flash memory cells: NAND, NOR, DINOR (divided bit-line NOR), and AND. So, which type is best for you? While there are such things as DINOR and AND cells, these are less common than NAND and NOR cells, so you'll most likely be looking at NAND or NOR technology. NAND is typically optimal for personal data storage; the biggest advantages of this type of storage are that it's less expensive than your other options and the cells are very compact, making the device smaller and more portable. NOR flash memory is typically used in devices in which the flash memory is embedded into the device, such as cellphones (Kingston Technology Corporation, 2012). NOR technology has a quicker access time than NAND technology; both DINOR and AND cells are attempts to retain NOR's quick access time while reducing the area of the cell to one comparable to that of a NAND cell (Integrated Circuit Engineering Corporation, 2002). So in essence, NAND is the cheapest, smallest type, but NOR has a quicker access time. The other varieties have mixtures of these qualities, but again, it is unlikely that you will encounter DINOR or AND types.

Using and Storing Devices

When using optical disks and magnetic tape, the storage conditions make a huge impact on how long the item will last with all of the data intact. This is less true with flash memory. While cool, dry conditions are best for electronics, flash media can work perfectly well when stored under less-than-ideal circumstances. Flash memory is not susceptible to many of the vulnerabilities of other storage mediums. As mentioned previously, it's often possible to recover data from a flash device that has gotten wet, so long as it's thoroughly dried first. Unlike tape, flash memory is not vulnerable to magnets or magnetic fields.

Static electricity can be a problem with flash devices. Again, the cells in flash devices use precise charges to encode the data, so anything that can potentially disrupt this is a problem. You may want to store and transport flash devices with this in mind (Kingston Technology Corporation, n.d.).

There are currently no reports of X-ray scanners damaging flash devices, but you may want to take precautions anyway if transporting devices. In addition, radiation scanning, which is done by the U.S. Postal Service, can damage a flash device, so keep this in mind if it is ever necessary to mail one (Kingston Technology Corporation, n.d.).

It's possible to corrupt the data on a flash drive. One of the common ways that this happens is when the user removes the drive while the computer is still writing information to the drive (this problem is more prevalent on older machines). If the user takes the device out of the computer too soon, the computer may not have completed writing the files to the drive. To be sure that the transfer is complete, it's best to use a software wizard designed to safely remove the hardware. This wizard will stop the computer's interaction with the device and inform the user as to when the file transfer is complete. Removing the device too soon can result in either an incomplete file transfer or damage or data corruption to the device.

As mentioned before, if you use flash drives or SD cards to move or store information, it's possible to lose the device through simple human error. If you're concerned about the safety of the data on the device or preventing others from accessing the information, you may be interested in encryption software, which will make it very difficult to access the information on the drive. Some major companies that manufacture these devices offer such software included with the device.

Key Points

- Flash memory is a more recent type of memory that functions similarly to the way a RAM chip does.
- Flash memory is one of the fastest, most convenient methods of data storage.
- Desktop and laptop computers can use internal flash memory devices as their main storage method. In the future, this technology may completely replace traditional hard drives.
- External storage devices like flash drives are compatible with a wide range of computers, both new and older, and typically plug into a USB port.
- Flash memory is a highly desirable technology that will withstand difficult storage conditions and harsh handling.
- Flash memory does wear out over time and is quite expensive. It's also prone to loss or theft.

In the following chapter, you will learn about a method of data storage unlike any previously explored in this book, a method that doesn't require you to use any space or storage devices, or even to perform checks or maintenance: cloud storage.

References

- BBC News. 2011. "South African Owners of Camera Found in Sea Traced." http://news.bbc.co.uk/2/hi/uk_news/england/8510314.stm.
- Bright, Peter. 2018. "Intel at Last Announces Optane Memory: DDR4 That Never Forgets." *Ars Technica*. <https://arstechnica.com/gadgets/2018/05/intel-finally-announces-ddr4-memory-made-from-persistent-3d-xpoint/>.

- Cornwell, Michael. 2012. "Anatomy of a Solid-State Drive." *Queue* no. 10. <http://queue.acm.org/detail.cfm?id=2385276>.
- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology Comprehensive*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Gregersen, Eric. n.d. "Flash Memory." *Encyclopedia Britannica*. Accessed February 29, 2020. <https://www.britannica.com/technology/flash-memory>.
- Hruska, Joel. 2019. "How Do SSDs Work?" ExtremeTech. <https://www.extremetech.com/extreme/210492-extremetech-explains-how-do-ssds-work>.
- Integrated Circuit Engineering Corporation. 2002. "Flash Memory Technology." <http://smithsonianchips.si.edu/ice/cd/MEMORY97/SEC10.PDF>.
- Katz, Jeff. 2012. "Oral History of Fujio Masuoka." Computer History Museum. <http://archive.computerhistory.org/resources/access/text/2013/01/102746492-05-01-acc.pdf>.
- Kingston Technology Corporation. n.d. "Caring for Your Flash Memory." Accessed August 31, 2019. <https://www.kingston.com/us/usb-flash-drives/caring-for-your-flash-memory>.
- Kingston Technology Corporation. 2012. "Flash Memory Guide." <http://media.kingston.com/pdfs/FlashMemGuide.pdf>.
- Maini, Anil K. 2007. *Digital Electronics: Principles, Devices, and Applications*. West Sussex, UK: John Wiley & Sons.
- Ngo, Dong. 2017. "This Is How SSDs Work and What You Can Do to Make Yours Last Longer." CNET. <https://www.cnet.com/how-to/how-ssds-solid-state-drives-work-increase-lifespan/>.



Cloud Computing

IN THIS CHAPTER

- ▷ What is the Internet? How does it relate to computer networks?
- ▷ How does the Internet work?
- ▷ What is cloud computing? What are the different models for cloud computing?
- ▷ What are the benefits and drawbacks of using a cloud computing service for archival storage?
- ▷ What are some things to look for in the contract for a cloud computing service?

In modern times, a computer is so inexpensive that millions of people literally carry one in their pockets, but this has been the case for a relatively short amount of time. In the 1950s and 1960s, when commercial computers started coming into use, computers were extraordinarily expensive. They were also extremely large and required a lot of maintenance. In general, only large organizations would have the ability to buy, maintain, and store such a piece of equipment. Generally, only large universities, government organizations, and large commercial businesses would have been able to own one.

However, then, like now, lots of businesses and organizations would have wanted to use a computer. It was useful for calculations, but also for maintaining business data, such as payroll and inventory. To meet this demand, it was also possible for companies to rent time on a mainframe computer, which can have multiple simultaneous users, essentially a time-sharing device (Ranger, 2018).

But how did companies access such a computer? It was possible, at the time, for a user to interact with a mainframe computer using a terminal (Arms, 2015). This would have been essentially a keyboard and a screen (some early ones did not even have the screen), which communicated back and forth using telephone lines or radio waves (Edwards, 2016).

Computers have come a long way since then and are now small, inexpensive, and ubiquitous. Renting time on a mainframe computer and communicating through phone lines using a terminal is unthinkable.

Only, it actually isn't different at all. The modern concept of cloud computing isn't modern at all, and it's nearly *identical* to this early situation, only instead of terminals and phone lines, you can use a fully functioning computer connected to the Internet, usually through much speedier means than telephone lines.

A lot of companies are offering services that allow you to use software "in the cloud" or encourage you to store data "in the cloud." This sounds very nice. Clouds are usually pleasant imagery, and so this phrase sounds pretty good to most people. It makes it sound as though your valuable data is hovering about the earth in a cloud, just waiting for you to pluck it back down from any computer, anywhere.

The reality is less mystical and abstract. It's difficult to give cloud computing a definition that applies to every situation, but basically it refers to a service that allows a client to access and use a different computer over the Internet. Generally this is for additional processing power or to save data to another machine, just as was done with those early mainframe computers.

Cloud computing offers some great benefits to an archive, but it also has some drawbacks that must be considered. Both will be explained in this chapter.

For modern cloud computing to function, a connection to the Internet is typically a necessary component. So, in order to understand what cloud computing is and what it has to offer for your archive, it's therefore essential to know this: What exactly *is* the Internet, anyway?

The Internet

What is the Internet? It's easy to describe things you can do on the Internet, such as view maps or videos, send e-mail, or shop. But what exactly is this technology? How does information reach your computer, and where does it come from?

Networks

Think of the Internet as a way of moving information from one computer to another. This book has already covered some ways that you can do this: You can burn information from one computer to a CD-R, for instance, or put it on a thumb drive, then put the CD into the CD drive of another computer or put the thumb drive into the USB port of another computer. That is very easy to do.

Another fairly easy thing to do is to connect two computers together so that they can directly communicate with each other—no storage medium required. This involves a router and connecting the devices with cables (a physical connection) or wirelessly (using a technology called *Wi-Fi* and sending data using radio frequencies). If you have a couple of computers connected together like this, then what you've made is a Local Area Network, or LAN. This is a group of computers, all in one relatively small area, that are connected for the purpose of sharing information.

Suppose that you want to communicate with a computer that's farther away, though? Say, in another building? If you make a bigger network that spans a distance greater than a single room or a small building, then what you've built is a Wide Area Network, or

WAN. There's also such a thing as a Metropolitan Area Network, or MAN, which spans a very large area, like a city or area of a city, but this is essentially just a really big WAN.

The Internet is also a WAN. Basically, the only difference between a WAN and the Internet is size. A WAN spans a few buildings, or maybe a city, as in the case of a MAN. The Internet is a WAN that spans the entire Earth. What this means to you is that the Internet consists of a collection of computers all around the world, which are connected to one another and are able to directly relay information from one computer to another. There are a lot of ways to transmit this information, but fiber-optic cables play a major role in the modern Internet, as they transfer information very quickly.

There are several companies whose computers, routers, and physical connections form the major part of this relay system, known as the backbone of the Internet. Many of these companies are probably familiar to you, such as AT&T, Verizon, and Sprint. The Internet does not belong to any single entity and would still exist, at least in part, if any of these companies stopped running their part of this network. The Internet has a lot of redundancies, which is good, because technical problems can occur with even the best and most well-run equipment, so the Internet can still function even if part of the network isn't working.

The World Wide Web

Again, the idea of connecting computers to share information has been around for a very long time now. However, it used to be that only someone with the knowledge and training to access a network and share information in this way could do so and actually find any information, as compared to today, when even a child can access the Internet with little trouble. Why have things changed?

The answer is the World Wide Web, also just called "the web." While often used interchangeably, the web and the Internet are two different concepts. The Internet, as explained, is a series of connections, while the web is basically a way to access information through those connections.

The web is essentially an easier way to access content through the Internet. Before its invention, it was necessary to know exactly where information that you wanted was (which computer it was on) and how to access it. Web technology makes this process much easier.

So, the Internet is the series of connections between computers that allows them to communicate with one another. Any type of information can be transferred via the Internet. The World Wide Web refers specifically to the web pages and websites that can be accessed via the Internet, which are designed to make accessing information very easy.

Without the World Wide Web, the Internet still exists and can be useful. Without the Internet, the World Wide Web is just a bunch of encoded HTML documents that sit on computers around the world with no way to share them. The web needs the Internet to function, but the Internet can do all sorts of things without the web.

Using the Internet

You most likely know how to do all sorts of things using the Internet and the web. It is, after all, designed to be intuitive. However, you may not know exactly what happens when you use the Internet or casually surf the web.

Information accessed via the Internet is not “out there” floating somewhere, intangible. It’s kept on a server somewhere, the ones and zeroes physically encoded on a computer. The Internet functions using server and client computers. When you request information—say, you want to view a website or download something—you’re digitally making a request for the information to be sent to your computer through the network. When this happens, your computer (the one making the request) is the *client* computer. The computer that distributes the information is then the *server* computer.

The term “server” might bring up images in your mind of big rooms filled with huge, incomprehensible computers, images that are probably due to the influence of Hollywood. There are, in fact, server computers that are like this. However, many servers are much smaller than this, similar in size and price to an ordinary desktop computer. It’s also not necessary to have a special type of computer to work as a server, and an ordinary desktop computer could be configured to be a server computer. It’s better, though, to have a computer specifically designed to function as a server. If you’d like to set up a network inside your library or host your own website for public access to your archive rather than have a company host it, this is something you can do, although it is far easier to pay a company to do it for you.

The data needed to make the web function is stored on server computers all over the world. Sometimes people refer to the act of viewing a website as “visiting.” This makes it seem as though the web is a place that you can visit and browse around for things, or that the server computers are like stalls in a market from which you can pick and choose. Though it makes for a fun visual, the reality is actually the opposite of this situation. Remember, server computers send data to the client when the client sends a request. So, what you’re really doing when you “visit” a website is asking a computer somewhere to send a copy of the site to you, more like ordering a package than going to a store.

How does your computer know where to send the request, though? There are many, many servers connected to the Internet, and the web page that you want when you send a request is on one of them. Try opening up a web browser—such as Edge, Chrome, or Firefox (any browser will do)—and visiting a site. Again, it doesn’t matter which site. Look at the top of your browser; there will be a string of letters, characters, and possibly numbers. This is the *Uniform Resource Locator*, or URL. Think of it like a mailing address. You send a request, and this string identifies exactly which web page you want and where it is; that is, exactly which server contains the web page. When you click a link or type an address into the search bar of your browser, you’re making a request of a server computer somewhere around the world.

This URL system is a way of “addressing” computers, just like houses have addresses to avoid confusion. Remember, though, that computers like to work in numbers; it’s humans who find words easy. Every URL has a corresponding IP, or Internet Protocol, address. The IP address is a series of numbers that the computer actually uses to locate another computer. Computers have to translate one to the other, which is beneficial to people, since it’s easier to remember a string of words instead of a string of numbers. Your computer has an IP address, too, if you’re connected to the Internet. After all, the server has to know where the information is going in order to send it to you.

As you now know, the Internet is essentially a big network, and you can request information from server computers attached to this network. In a LAN, computers can be directly connected to one another with cables. You don’t have a direct connection to any particular computer on the Internet, though. For instance, the search engine Google has several data centers in California (and in many other locations, but this is just an exam-

ple). If you live in Maine, then there isn't a cable running from your computer all the way to a server in California. That would be pretty inefficient.

Instead, when you send a request through the Internet for a particular web page, it moves through several different computers. A device called a *router* handles the requests from different computers and sends them on to other computers. It's not very precise; the router sends the request in the "general" correct direction, so your request for a web page might go on a bit of a roundabout journey on its way to the desired computer. Think of it as being a little like traveling via airplane. It would be impractical to have an airplane that goes to every place in the world at every airport, so you may need to fly between several airports before you get to one that is closest to your desired destination. Your request may need to travel through several other computers making up the network known as the Internet before it reaches the desired server computer, which can then send the information you wanted to your computer.

Ultimately, you need to have a correct URL in order for a server computer to get your request for a web page. A generic URL might look like this: `http://www.awebpage.com`. All URLs have some elements that are the same. The HTTP at the beginning stands for Hypertext Transfer Protocol. There are several different protocols that computers can use. Think of it as computers using a common language. On the Internet, all computers "speak" the same language to avoid confusion and to facilitate the exchange of information; this is the Hypertext Transfer Protocol.

The .com at the end of the name is a top-level domain name or TLD; this indicates the type of organization that owns the web page. The .com TLD name is pretty generic, and can belong to almost any person or organization. For instance, a business could have a .com TLD, as the commercial website Amazon.com does. It could be used for a personal website, too, and your archive could also use a .com TLD if you had a website. The TLD names .org and .net are unrestricted, too. The .org is typically used for nonprofit organizations; for example, the charity organization the American Red Cross has the web page `http://www.redcross.org`. In contrast, the TLDs .edu or .gov have restricted use. Only educational organizations in the United States, such as universities and public school systems, may use the .edu TLD, and only United States government websites may use .gov: federal, state, and local governments are all allowed to use this TLD. There are many other TLDs, as well, but these are the most common ones.

Defining Cloud Computing

You now know what the Internet and the World Wide Web are, as well as the basics of how they work. So what exactly is cloud computing?

As explained at the beginning of the chapter, the concept behind cloud computing has been around for some time now. Even today, a mainframe computer (which has a lot of processing power) or a server might be connected to multiple smaller computers (wired or wirelessly) in order for those other computers to share data or to use the superior processing power of the mainframe or server computer. This can be very efficient. However, this kind of setup is often only done within a small area, such as a single building.

Cloud computing is like this model in that someone using cloud computing is sharing a more powerful computer with others. However, you're sharing that computer with lots of other people around the world (you may or may not know who, depending on your

setup), and the distance between your computer and the one providing the service might be quite great—even on the other side of the planet.

The term “cloud” makes it seem as though the information stored in the cloud is nebulous, floating about in the air like the water molecules of a cloud. It drifts along, waiting for someone to pluck the information out of the air from anywhere in the world.

The reality of the situation is quite a bit less glamorous. What generally happens in cloud computing is that you, the user, agree to allow someone else, usually a company, to store your information on their server computers, typically for a fee. Cloud computing can also be used for processing data; you can use software programs “in the cloud.” You can request your data from the computer storing that data in the client/server model described earlier, with whatever computer you use at the time, requesting the information from their servers, and those servers send it to you via the Internet. It really works a lot like how you get web pages. However, access to your information is typically restricted to you or to a small group that is allowed access.

The data that is stored “in the cloud” typically resides on physical hard drives or solid-state drives. Hard drives are still popular because they are cheaper, but SSDs are increasingly used because they are fast and require less electricity.

Cloud computing is not easy to define in precise terms, and there isn’t a pretty, tidy definition available at the moment, unfortunately. The National Institute of Standards and Technology, or NIST, which is a part of the U.S. Department of Commerce and works to define national standards, defines cloud computing as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (NIST, 2018).

While this is, in fact, a definition (and a pretty concise and accurate one, considering what it’s trying to describe), it’s a bit clumsy and difficult for an average person to understand. An easier way to go about this is to think of cloud computing in terms of what it can do. NIST also defines the features that any cloud computing service has:

On-demand self-service. What this means to you is that you don’t have to interact with an actual person to access your data. This is all done automatically whenever you want to access the service, no humans involved (NIST, 2018).

Broad network access. This means that the cloud service is accessed over a network and can be accessed via a wide range of devices (an important concept in an age of smartphones and tablets) (NIST, 2018).

Resource pooling and rapid elasticity. These both refer to the flexibility with which a service can meet consumer demand (NIST, 2018). For instance, suppose you bought a one-terabyte hard drive to store data, but you really only need to store 500 GB of data. The money you spent on the rest of that terabyte could be considered wasted because you paid for something you aren’t using. With a cloud computing service, however, you can pay according to *exactly* how much storage space you need. You can also get varying levels of computer processing power.

Measured service. This refers to the ability of the service to control and meter the service, changing how much storage, processing, and bandwidth is available to a user depending upon their needs and usage (NIST, 2018). For instance, imagine that you and another archive are using the same service (imagine that it’s only these two using the servers for the sake of simplicity). You have a photo-editing program stored on your

service, and in the mornings, you use the processing power offered by a cloud storage service to use the program more optimally than you could with your computer, since the server computer can make the necessary calculations faster than your computer (in this scenario). The other archive is using the service for a photo-editing program in the afternoons. The service can make more processing power available to your archive in the morning and less in the afternoon, and make that extra processing power available to the other archive when they are using the photo-editing program.

Cloud Computing Models

There are a number of models for cloud computing. One of the common types is a *public* cloud. These are clouds, formed by servers, that are available to anyone who subscribes to the service. For instance, the company Amazon offers its users cloud storage services. Because any of its users can access their information on a server and the same servers are used for many different users, this is a public cloud.

FOUR MODELS OF CLOUD COMPUTING

- Public
- Private
- Community
- Hybrid

A *private* cloud, in contrast, is a cloud designed for one specific group or an organization. It's only accessible to members of that group or organization (ideally), making the data more secure. It also gets around many of the legal issues that can arise when using a public cloud, which will be discussed later in this chapter. A private cloud doesn't have to be connected to the Internet to function (you could use an ordinary LAN or WAN), and so there is some debate as to whether a private cloud really counts as cloud computing (Corrado and Moulaison, 2011). While the benefits of extra security and dodging issues with service contracts is highly appealing, a private cloud is probably going to be impractical for you unless your archive is quite large.

A *community* cloud is like a private cloud in some ways, but it's shared among several organizations that are similar or have similar needs. For example, suppose that you and two other archives wanted to get some of the benefits of a cloud computing service (such as backups in different locations). If your archive and the other two archives each had a server computer that contained the data for all three archives and each was able to access any of the other servers, you'd have a community cloud. There are other ways you could go about it, as well. The organizations that operate the cloud may also manage it, or they may use a third party for this part. A community cloud can divide the operation cost among the parties involved, and unlike using a public cloud, in which you must agree to the terms of service set by the company, your archive can have a say in how the cloud is run, how resources are distributed, and how security is maintained (Corrado and Moulaison, 2011).

There's another option, known as a *hybrid* cloud. This is less simple to define, as a hybrid cloud is essentially a cloud service that doesn't fit neatly into the other categories and is instead some combination of the other cloud types (Dale and Lewis, 2013).

As with some of the other storage methods and concepts explored in this book, chances are good that you're already using cloud computing, but you may not realize it. For example, if you have an account with Google and use their e-mail, documents, calendars, or many other services, you're using cloud computing. Other companies that provide e-mail services but don't leave a permanent copy of the e-mail on your computer, like Yahoo Mail, are working "in the cloud." You must access your data from the company's servers, since it is not physically on your computer.

Along with general setup models, there are also several models for exactly what type of service a cloud computing service is offering. The most common of these are IaaS (infrastructure as a service), PaaS (platform as a service), and SaaS (software as a service).

With IaaS, you're essentially renting computers, servers, operating systems, and so on. This is very like the original model of renting a mainframe computer; the cloud computing service is providing the hardware and software for you to use. One of the benefits is that, using this service, you don't have to set up and maintain expensive hardware to get the benefits of having that hardware.

PaaS is typically intended for an organization developing their own software; using the cloud service, they can both use the software provided by that service to create software and use the cloud service to additionally run that software in the cloud.

Some cloud computing companies are offering software and associated databases as their service. SaaS refers to services like this, in which you are using a company's software "in the cloud." Again, services like Dropbox or Google Docs are SaaS-type cloud computing services.

Cloud Computing for Archival Storage

Cloud computing offers benefits for archives of all sizes, and can be quite cost-effective in some situations. As with all methods of data storage, though, cloud computing has some drawbacks, which are a little more complex than those explored in the book so far. This is because these drawbacks have less to do with physical problems or technical issues and more to do with legal and security issues, which are somewhat more abstract.

ADVANTAGES AND DISADVANTAGES OF CLOUD COMPUTING STORAGE

Advantages:

- Multiple copies of your data (backups)
- Very cost-effective (hardware, maintenance, electricity)
- Potentially time-saving

Disadvantages:

- Potential disputes with data ownership and rights
- Access dependent upon the Internet and an external company
- Potential problems with security

Advantages

One of the best things that you can do to protect your data is to keep copies of it on multiple devices and in multiple locations. This protects it from hardware obsolescence as well as any environmental issues that may arise, such as a flood or a fire in the archive. Cloud computing lends itself nicely to this. In most cases, cloud computing services offer backups of your data (though this is something you should investigate when choosing a company, since not all do), which helps keep the data safe. In the best-case scenario, these backups are kept on different servers in different physical locations, which protects your data from a disaster occurring in one location, as there will be a backup copy at a completely different and unaffected area. Ensuring that the data is saved on an up-to-date storage device is also, theoretically, part of what you pay for, so this is no longer part of your considerations when planning for your storage.

Cloud computing can potentially save your archive money if you're on a tight budget. You don't have to buy a physical object with cloud computing. You pay for what you use, both in processing power and storage space. While using cloud computing to store your data may be the use that first comes to mind for you, you can actually use cloud computing to gain a boost to the processing power of your archive. For example, in chapter 2 you learned how the CPU of a computer is a major component that controls how fast a computer can operate. Suppose that you don't have the ability to upgrade your computers right now. You could potentially access the processing power of the server computer, which you use as part of the cloud computing service, to run complex programs for your archive, making the fact that the computers you actually own are somewhat slow less relevant (so long as you have a good Internet connection, of course).

You also don't have to pay someone to monitor a server or mainframe computer in your archive or any other equipment along those lines with a cloud service. Someone else does this for you. You don't need to hire someone to come to your archive to maintain your equipment or to understand the complexities of computing. The hardware, the software, the storage space, the maintenance, the technical assistance—this is all part of what you pay for, which may make this a particularly economical choice for you if you have a small staff or a limited budget to make your archive run.

As another bonus that you might not initially consider, using your own server computer requires power. You can potentially save money on electricity by using a cloud computing service, since providing the energy to run the server computers is the job of the company. You also save valuable space by not purchasing a physical item.

Cloud computing offers the benefits of a server or a mainframe computer, as mentioned earlier. You can keep any kind of data in a cloud, including software programs. If you do this, then you get some significant benefits. For instance, you may have a particular program installed on only one computer in your archive, and so if you want to use that program, you must use that particular computer; this is a problem if more than one person needs to use the program or if something happens to that computer. Or, you may want a certain software program on every computer in the archive, and so you'll need to install the software on every computer, a time-consuming process.

If you use software in the cloud, then you'll eliminate the need for tedious tasks like updating or installing software on every computer in your archive, and get around problems like software existing on a single computer only. The updated software will automatically be available to everyone. The same principle can apply to your collection: you don't need to install a copy of your digital collection on every computer, since every

computer can access it from the cloud. This is potentially a huge time-saver for you. You should be aware of the terms of service of software before doing something like this, though, since the terms of service may not legally allow for software to be uploaded to a server in this manner.

If you store your data in the cloud, this has advantages, too. You don't need to find a physical object to retrieve your information. Your data is accessible from a single location, and everyone in your archive can access it at the same time. This can make finding and retrieving data much quicker and simpler. Cloud storage can also make it easier to share your archive with others, whether you want to share it with other archives, patrons, or both. Because the data is online, you can potentially grant access to it to whomever you choose without the need for that person to physically come to your archive or be sent a physical object.

For all the wonderful things that cloud computing can do for your archive, though, there are some drawbacks that you should know about before deciding if buying or setting up this kind of service is the best choice for your archive.

Disadvantages

While the advantages that cloud computing holds for an archive are extraordinary, there are some not insignificant issues that can arise if you use a cloud computing service, largely regarding the rights and ownership of your data. Unlike the other methods of storage described in this book, in which your rights to the data you store are undisputed (assuming you have the legal right to store it in the first place), cloud storage gets complicated.

Determining your rights with cloud computing can be a very difficult business. If the company you choose stops offering the service for any reason, you may be unable to retrieve your information, even if you take legal action against the company. There is legal precedent for this situation, as well; if you load your information to an external server as part of a cloud computing process, you may lose some or all of your rights associated with that information. That includes your rights to retrieve it, and possibly your right to store it (Heaven, 2013).

Some cloud computing services use programs that browse through the information stored on their servers and delete anything that may possibly be illegal or pornographic, whether it actually is or not, and, of course, without consulting the user. As an example of these issues, the cloud storage company Megaupload was taken off-line by the FBI in 2012. This was because many of its users were using the service to store pirated films, games, software, or other illegally gained and distributed information. Anyone who was using the service, to store things legally or illegally, lost access to their information, and thus the information was, for all practical purposes, gone (Heaven, 2013). In this case, the redundancies and multiple copies of data that are such a wonderful asset to cloud computing were rendered irrelevant.

To think of this in more tangible terms, imagine that you were using a storage facility to store some old books. However, someone using the space next to yours was using it to store some illegal substance. The police find out about this and confiscate everything in the storage facility, and thus, your books are now property of the police department. They are out of your control and you no longer have access to them. This isn't equivalent to what would happen if you really were storing books in a storage facility, but, given legal precedent, it's very similar to what might happen to your online data.

Your legal rights in regard to who owns and has rights to your data become a sticky issue when you use a company to store your data, as well. For example, the social media website Facebook (which you may not realize uses cloud computing to function) reserves the right to use any image or information uploaded to the site. Their property Instagram can also use the images that the users upload for a variety of purposes. The agreement between you and a cloud storage company can even give the company ownership of the data that you store; in 2011, the online storage company Dropbox changed its terms of service, which gave the company rights of ownership over all of the data its users stored. Though it reversed the policy in response to the public outcry that followed, this is something that could happen with any online storage company (Hallene, 2013). A change in a company's contract could occur at any time, as well. Any company that you choose for such a service must be subjected to severe scrutiny to determine what exactly they reserve the right to do with your data, and if this is a problem for your archive.

Cloud computing is dependent upon the Internet to function. If you lose connection and all of your archive's data and programs exist in the cloud, then your archive's work grinds to a halt for as long as it takes for the Internet connection to be restored. If your Internet connection is reliable, then this may not be a problem, but if it's shaky, then dependence upon this service may not be such a good idea. Similarly, if the cloud computing service becomes unavailable due to problems with the company you chose (technical issues, for instance), then you likewise lose your ability to access your data.

In the section regarding advantages, it was mentioned that putting your data in the cloud could potentially make it easier to share with others. This is true up to a point. If you have a lot of data to share, then transmitting it over the Internet can be slow and cumbersome—it's often easier to simply send a physical object with the information.

Storing information in the cloud is inherently less secure than any other storage method. Cloud computing services, of course, have security and strive to keep your data secure. However, this does not mean that hackers are completely unable to access your data if they want to. Those servers containing your data must be connected to the Internet network, which means that it is possible to get to those servers via the Internet, even for people who are not authorized to access the data. And remember, there isn't a line connecting you directly to a cloud storage server; whenever you store or retrieve your data, it will have to pass through many different lines, as described previously in this chapter.

Strict security may or may not be an issue for you. However, you should always be wary of someone corrupting your data maliciously or causing problems for the company that you purchase your service from. If security is indeed an issue for your archive, there are some steps that you can take to make your data more secure. For instance, if you encrypt files before you upload them to the server and keep the key to decrypt them only on the computers at your archive, this makes it significantly more difficult to access your data illegally (Hallene, 2013).

The legal issues that arise become less of a problem if you have a private or community cloud, but you may lose some of the advantages that come with a public service. For instance, you don't have to own any equipment or pay anyone to maintain that equipment with a public cloud service, but this is something you may require if you decide to, essentially, create your own cloud. You may also lose the variety of locations for servers if you collaborate with local organizations; cloud storage services can potentially have servers located all over the world.

Unless you decide to create a private or a community cloud—in which case, you control the servers and how they are used—you will need to deal with a company that offers cloud storage as a service. This will involve navigating a contract.

Cloud computing is going to be much different from your other storage options. While it's a good idea to read over warranties and other information from the manufacturer, once you buy a CD or a tape cartridge or a flash drive, it's yours and you can essentially do whatever you want with it. When you purchase cloud storage space, you don't own anything. Think of it more like a rental of someone else's space. What all this means to you, in practical terms, is that there will be a contractual agreement involved, and you must read this very carefully and understand all the implications that it holds for your archive. If possible, use legal counsel or someone similarly qualified to assist you in determining your archive's rights when purchasing such a service.

THINGS TO LOOK FOR IN A CONTRACT

Does the company offer backups of your data?

How long does the subscription last, and how is it terminated or renewed?

How do you pay, and what happens if you miss a payment?

What warranties or guarantees are offered by the company?

Can the company be held liable for service or security issues?

Does the company collect data, and what is done with the data?

Where are the servers located?

What kind of security is offered? Are there multiple layers of security available?

What are your rights in regard to data ownership with the service? Does the service claim any ownership rights?

Subscriptions

It should be clear in a contract how long your subscription lasts and how, when, and under what circumstances it terminates or can be terminated by you. It should also be clear how the contract is renewed should you decide to continue using the service. Some contracts can make it difficult to stop your service. Be wary of early termination fees, as well. If you decide to terminate your contract, you also need to find out what happens to your data—whether the service gives you time to migrate or whether it is deleted immediately (McKendrick, 2013).

Backups

Not all providers offer data backups. Find out whether a service does or not, and what kinds of guarantees they offer as well as whether or not the company accepts any sort of responsibility for data loss (McKendrick, 2013).

Payment

This will be an ongoing expense for your archive. Learn what it costs and what you get for what you pay, how your institution pays for this service (for instance, what methods of payment are acceptable), how often you need to pay, and what exactly happens if you don't pay, miss a payment, or are late with a payment. As an example, it would be inconvenient to you if a company deleted all your data because you were a day late with payment.

Terms of Service

Learn exactly what the terms of service are. For instance, what kinds of warranties or guarantees are offered by the company? All companies, even ones that are impeccably run, are subject to technical issues, and so you must expect that the service will be unavailable sometimes, but you should find out how often the company guarantees that the service will be up and operational. A company that doesn't run smoothly is of no use to you.

Additionally, look for any stated issues or breaches of the terms of use that might cause a company to terminate a contract early. As stated earlier, the misuse of a service by one user might impact others; be sure that you understand if and how this could happen (McKendrick, 2013).

Data Collection

Cloud storage companies may monitor the usage of their service, and so you should find out exactly what the company monitors and collects as far as personal data goes. You need to find out if data that the company collects about the usage of your data stays solely within the company and how it is protected or, perhaps more importantly, if they share or sell any data about their users.

Server Location

Though you're buying storage space "in the cloud," remember, the servers that store the data have a physical location somewhere. It's in your best interest to know where. The actual, physical location of the servers determines the laws, rights, and regulations that apply to the data stored on those servers (a phenomenon of particular interest to lawyers who store data in the cloud). Your contract may specify where the servers are located or even guarantee that they are located only within a specific area, or it may have other information regarding locations and your rights in regard to the server location (Hallene, 2013).

Security

Find out what kind of security the company offers. Different companies offer different types and levels of security. Authentication, or verifying the identity of the user, is one of

the most common methods of adding security. This is typically a username and password; some companies can offer multiple layers of security, though. For instance, Dropbox requires a six-digit code whenever a user logs in using an unrecognized device, or one that they haven't used to connect to their account before; this helps to ensure that the user is not a stranger who is not supposed to be accessing the account (Hallene, 2013).

In some services, you are able to allow different levels of authorization. For example, you can allow some users to access the data but not change it, and others to both access and change data. This is perfect from the perspective of an archive, as you can permit your staff and your patrons to have different levels of access. Along the same lines, your cloud computing service should be responsible for notifying you of any security breaches, what they mean to you and your data, and any potential security breaches (that they should be working on repairing, ideally).

Data Ownership

Finally, one of the most important parts that you should look for in a contract is how the ownership of the data you store is affected by the terms of the contract. You need to know whether agreeing to the contract gives the company you use any legal ownership over what you store and what happens to your data should you terminate your contract. The company may retain copies of your data, even if you delete the files that you have stored or if you stop using the service.

The Future of Cloud Computing

The storage methods discussed in this book all have their own pros and cons, and all face the possibility of obsolescence in the future as technology improves and new ideas are created. Some are already obsolete or becoming obsolete.

It's unlikely that cloud computing will become completely obsolete any time soon, but already new models are coming into use to compensate for some of the problems of cloud computing. As mentioned previously, modern cloud computing generally requires the Internet (unless you have your own cloud setup), and transmitting data over the Internet takes time. It also uses a lot of electricity to run the equipment and the data centers required for large-scale cloud computing services, and so alternate models are seeking to reduce time and improve efficiency.

Cooling computers in data centers is actually a problem that is leading to some interesting innovations. For example, there are plans to build large data centers in Norway due to its naturally cool climate and water available for cooling (Kelion, 2017). Microsoft has plans to build data centers that are sealed inside cylinders under the ocean off the coast of northern Scotland because of the water's naturally cooling effects (Rutherford, 2018).

Edge computing is a concept in which data is processed as close to the source as possible. This means that, on a phone, for instance, the data gets processed on the phone itself, then transmitted for storage in the cloud. This results in less data being transmitted and an improvement in speed and efficiency. This is not really a novel concept in that computers have been functioning this way for decades, but it's a variation on the cloud computing model with some important advantages. This concept may also have an impact on the

concept of the “Internet of things,” or devices that are not perceived as computers, but require the Internet to function (such as “smart” thermostats) (Markman, 2018).

Fog computing, also known as *fogging* or *fog networking*, is yet another variation on the model. Again, it processes data closer to where it was created, then transmits the data to a data center or elsewhere on a network. However, rather than being processed on the device itself, it may be processed by a *node*, which is a device configured for the purpose of processing data in a fog computing model (a router might be designed to do this, for instance). Again, this can improve speed and efficiency (DeMuro, 2018).

The simple cloud computing model is likely to be sufficient for most archives; these models solve several problems, but speed is the major one (for example, a smart car needs to process data extremely rapidly in order to operate safely, and so the edge or fog models can be helpful). However, it may be of use to you to know about the alternatives and changes in technology, as they may impact services in the future or may have an influence on how you decide to operate your archive.

Key Points

- Cloud computing offers you an alternative to other storage methods and allows you to store data on large server computers outside of your archive rather than on computers in the archive.
- Cloud computing services come in several models: a public cloud service, a private service, a community service, or some combination of the three. Each has different benefits and drawbacks.
- Cloud computing companies can offer infrastructure, platforms for software development, software applications, or data storage as part of their services.
- For an archive, cloud computing’s off-site storage and data-backup features are very appealing as they protect your data from local disasters or hardware issues, and can provide extra processing power and other conveniences, saving your archive time and money.
- Cloud computing services do have some serious drawbacks in that they are dependent upon an Internet connection, and if you use a company to provide the service, then you may encounter difficulties in regard to the rights and ownership of your archive’s data.

While the items or services that you use to store your data are very important, you don’t just need methods of data storage to create a working digital archive. You will, at the very least, require additional computers and software, and in the case of digitizing information, items like scanners and cameras. In the following chapter, you’ll learn about some of the useful equipment that your archive may need for your project.

References

- Arms, William. 2015. “The Early Years of Academic Computing: Commercial Timesharing Systems.” Cornell CIS: Computer Science. <http://www.cs.cornell.edu/wya/AcademicComputing/text/earlytimesharing.html>.
- Corrado, Edward M., and Heather Lea Moulaison, eds. 2011. *Getting Started with Cloud Computing: A LITA Guide*. New York: Neal-Schuman Publishers.

- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- DeMuro, Jonas. 2018. "What Is Fog Computing?" TechRadar. <https://www.techradar.com/news/what-is-fog-computing>.
- Edwards, Benj. 2016. "The Forgotten World of Dumb Terminals." *PCMag*. <https://www.pcmag.com/feature/348634/the-forgotten-world-of-dumb-terminals>.
- Hallene, Ashley. 2013. "Clearing Up the Cloud." *GPSolo* 30, no. 1: 34–38. EBSCOhost.
- Heaven, Douglas. 2013. "Lost in the Clouds." *New Scientist* 216, no. 2910: 35–37. EBSCOhost.
- Kelion, Leo. 2017. "Record-Sized Data Centre Planned inside Arctic Circle." BBC. <https://www.bbc.com/news/technology-40922048>.
- McKendrick, Joe. 2013. "9 Questions to Ask before Signing a Cloud Computing Contract." *Forbes*. <https://www.forbes.com/sites/joemckendrick/2013/01/14/9-questions-to-ask-before-signing-a-cloud-computing-contract/#3626a8721e3b>.
- NIST. 2018. *Evaluation of Cloud Computing Services Based on NIST SP 800-145*. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-322.pdf>.
- Ranger, Steve. 2018. "What Is Cloud Computing? Everything You Need to Know about the Cloud, Explained." ZDNet. <https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-from-public-and-private-cloud-to-software-as-a/>.
- Rutherford, Sam. 2018. "Microsoft's Newest Data Center Is a Giant Metal Can at the Bottom of the Sea." Gizmodo. <https://gizmodo.com/microsofts-newest-data-center-is-a-giant-metal-can-at-t-1826606291>.



Equipment for Digitizing and Editing Archival Materials

IN THIS CHAPTER

- ▷ What types of monitors are available, and what is best for my project?
- ▷ What are the different kinds of scanners available, and what are they best used for?
- ▷ How can I digitize delicate items or items that can't be scanned?
- ▷ What kind of software do I need to process text and image files?
- ▷ What equipment do I need to digitize audio and video materials?
- ▷ What kind of software do I need to process audio and video materials?

It's easy to forget that a computer is just a device that performs calculations. But how does it know what calculations to perform?

Because humans cannot interface with computers (at least not yet), it's necessary to use items known as *peripherals* to communicate with the computer. A keyboard, one of the earliest and most basic ways of putting data into a computer, is a peripheral. A printer, one of the earliest ways of viewing the results of a computer's calculations, is also a peripheral.

Similarly, software (sets of instructions for computers) is required for modern computers to do anything. While the computer itself is, of course, the major hardware component necessary for your archiving project, a number of other items and software programs will be of use to you. This is especially important for a digitization project, in which you create digital copies of physical items, but much of the hardware and software discussed in this chapter can be of use to you even if you are simply storing digital items.

The equipment that you will require will largely depend upon what type of items you need to store. Are you archiving audio recordings, or are you archiving newspapers, for example? These will require radically different types of equipment to create a digital version of the original.

This chapter will give you an idea of what kind of equipment you might be interested in for your archive. It will not cover every possible situation or everything you might possibly need, since it would be beyond the scope of this book to cover everything. These are just some basics that you would need to get started in most situations.

Determining When to Outsource

Equipment is expensive. This chapter will give you an idea of what kind of equipment you might be interested in for your archive. Most of the items discussed have versions that are aimed at the public and those that are aimed at professionals. Items aimed at professionals, of course, usually come with a much higher price tag, but typically offer more options, more control, and higher-quality results. It's up to you to decide which will best suit your needs and your budget. Sometimes the best equipment available isn't necessary, and sometimes it is.

After learning about your options, you may ultimately decide that you don't have the money for what you want or that you don't have enough staff to work on a project. It is possible to outsource a digitization project to a company that does this professionally. This has some pros and cons for you: You won't be able to monitor what's going on or be able to control when and how quickly the digitization process happens, and you'll need to send your items away. It's also quite expensive and you won't have made any kind of investment in equipment that could be used again if you want to digitize more items.

On the other hand, the facility that you send your materials to will likely have the best equipment available. Your materials will be handled by people who have experience with digitizing materials, meaning that you and your staff don't have to learn anything to get quality reproductions and can spend your valuable time on something else or on some other project. While outsourcing digitization is expensive, you may discover that it's worth your while if you can't afford the necessary equipment or if you can't spare the necessary time (digitization is a lengthy, tedious process no matter what you digitize).

While you may be interested in outsourcing to save time or money, there are some items that you must have to run a digital archive, and one of these is a good monitor.

Monitors

A monitor is a type of peripheral; again, this means that it connects to the computer to exchange information between the computer and the outside world. Early computers actually had no monitors at all—the main ways of interpreting the function and output of a computer were lights (basically indicator lights) and printers. The very first monitors were very basic CRT monitors that had no colors (Edwards, 2010).

This is unthinkable today, where most of a computer's functions are interpreted using a monitor, and where simple visual interfaces are essential for the vast majority of users in order to use computers at all. Many computers now are completely integrated into their monitors, making the monitor and the computer into a single item.

If you buy a new monitor, it will almost certainly be a Liquid Crystal Display, or LCD, monitor. In the past, there were two other options for consumers: the Cathode Ray Tube (CRT) and the Plasma Display Panel (PDP).

CRT monitors contain a cathode ray tube, which is essentially a big glass tube behind the screen. The back of the screen is coated with tiny dots of red, blue, and green phosphorous material. When these dots are activated with electrons from the cathode ray tube, they glow, and these three colors, red, blue, and green, can be combined to create a wide variety of colors (Schmidt, 2000).

With PDPs, the screen is composed of two pieces of glass, between which are little cells of gas (xenon and neon) and phosphorescent material, similar to CRTs. When electricity is applied, these little cells glow (Samsung, 2018).

LCD monitors contain a substance known as *liquid crystal*, which is put between two layers of polarized glass (there are several other layers involved, as well). It's a little more difficult to visualize how liquid crystals work, since they function on a chemical level rather than a more mechanical level, as a CRT monitor functions. In chapter 7, you learned about how lasers can change the opacity or transparency of the material in a rewritable CD. For an LCD monitor, electrical currents change the structure of the liquid crystal material in a somewhat similar fashion to the rewritable CD, causing it to temporarily change form and block some light waves, or certain colors, so that they aren't visible to the user (Fuller and Larson, 2008).

There is actually another type of monitor, an LED monitor. This is essentially the same as an LCD monitor except that it has LED, or *light emitting diode*, lights as the backlighting; LCD monitors use CCFL, or *cold cathode fluorescent lamp*, lighting.

Keeping outdated equipment can potentially be useful, as it might be required to retrieve old data or to reproduce old data faithfully. As an example, the Nintendo videogame Duck Hunt involves players hunting a digital duck using a plastic light gun that can detect where the duck is on the screen. The game does not function using modern televisions and requires a CRT TV in order to play, as other types of televisions do not refresh the way that a CRT does and the refresh feature is part of how the game is designed so that the light gun can detect the duck on the television.

CRT and PDP monitors both work perfectly well. In fact, PDP monitors have some advantages over LCDs, but they are no longer manufactured; LCD monitors are cheaper and use less electricity, among other important advantages. CRT monitors are problematic in that they are heavy and have poor resolution in comparison to a newer computer, among other disadvantages. However, depending upon your project, a CRT monitor could be handy to keep. Software, websites, and other digital materials originally intended for display on a CRT monitor may be experienced most faithfully to their original design on a CRT monitor.

When selecting a monitor, there are several easy characteristics that you can look for.

Resolution

A monitor's resolution is one of the important aspects to take into consideration when making a purchase. The term "resolution" means something slightly different for monitors than it does for images. With a monitor, the resolution is the maximum number of pixels that a monitor can display. For instance, a resolution of 1366×768 means that a monitor

can display 1366 pixels horizontally and 768 vertically. The numbers always go in that order, horizontal, then vertical, separated by an \times .

An image may have more information than there are pixels in your monitor. That is, you may have a monitor that is 1366×768 , but the photo that you want to look at is actually 2000×2000 pixels. When this happens, the monitor essentially has to take an average of the pixels and resize the image to fit your screen (Dale and Lewis, 2013). Because images can be quite large, this can help you by allowing you to see an image at a more comfortable size. When working with images, though, a bigger monitor with higher resolution is almost always better.

Video Card

A *video card* (also known as a *graphics card*) is not part of a monitor, but has an impact on speed and quality of images on a computer. It connects to the motherboard (covered in chapter 2). It has its own memory and processor, and performs the processing for graphics information on a computer. A good video card can improve the speed and quality with which images are rendered. If you work with videos or potentially images or software, it might be worth investing in a good video card. Note that the card needs to be compatible with the computer it's installed on and that a graphics card can use a lot of power and generate a lot of heat. Many come with integrated fans for this reason.

Size

Like televisions, a monitor's overall size (apart from the number of pixels it can display) is listed not by the height or width, but by the approximate diagonal measurement of the screen. Again, bigger is usually better, since you'll be able to see images more clearly or look at multiple images at once on a large screen. However, big screens take up more room and cost more money, so you may need to compromise.

Ports

Like a computer, a monitor can have a variety of ports included. HDMI, DVI, Display-Port, VGA, and USB-C are some of the more common types, but there are others. These are basically all ways to connect the monitor to another device. You may also find other ports on a monitor, like USB ports. You'll need to decide ahead of time how you'd prefer to connect your monitor to your computer, if you'd like to be able to connect any other devices, and if there are any other kinds of ports that you'd find useful.

It should be noted that there isn't a great deal of difference between LCD monitors and a flat-screen LCD television. If an LCD television has a compatible port (HDMI ports are particularly common and easy to use), it might be possible to use that. However, it's typically best to use a monitor for computer use.

LCD monitors are popular because they are lighter, less expensive, and can be larger than the other options. However, this could easily change in the future. For example, organic light emitting diode or OLED technology allows for screens that are even thinner and more energy efficient than LCDs. This technology is more expensive than the other options, however, but this could change.

Other Features

Some monitors are adjustable and some are not. That is, some monitors allow you to adjust the height of the monitor on the stand, the angle, or the orientation (vertical or horizontal). These can all make work easier for whoever is using the monitor.

Some monitors come with built-in speakers, and some don't. Having the speakers built into the monitor can save space and reduce the number of cords lying around, which is convenient. However, if you are going to be working with audio or video data, you'll probably want separate speakers that are higher quality than what is typical for a monitor.

Whatever type of monitor you choose, it's important to have the monitor properly calibrated. Monitors and televisions can both have differences in how they display color from one monitor to the next. Calibration helps ensure that a monitor is displaying properly, or that the colors displayed are true to real-life colors. There is software available to assist with this; it's typical for computers to come with some software to calibrate a monitor or for the monitor itself to come with software. There are also online tools that can help, as well as physical tools that can help calibrate a monitor, such as a colorimeter. If exact color is important, you may want to invest in more professional software or tools.

Scanners

You're probably familiar with scanners already. These devices have the capability to capture an image of an item (like a typed paper or a photograph). Most scanners have a couple of things in common.

The scanner has a device inside that can detect light. If you read chapter 3, you know a little bit about how this works, as the devices used to record digital photography and those used for scanning are essentially the same. A bright light inside the scanner shines on the object within the scanner. The photosensitive device within detects the light reflecting off the object and records it digitally.

There are basically four different kinds of devices that a scanner can use for this process: a PMT, or photomultiplier tube; a CCD, or charge-coupled device; a CMOS, or complementary metal oxide semiconductor; or a CIS, or contact image sensor. Of the four, PMTs are the highest quality, but are usually part of a drum scanner, which is a specialty scanner that will be discussed a little later. CCD scanners are generally common and have good quality.

As you learned in earlier chapters, while there are differences between computers and software programs, sometimes different designers and manufacturers agree to use standards to make things easier for everyone. For instance, the hypertext transfer protocol is used to make transmission of information over the Internet easier. Scanners have a standard, too, called TWAIN. TWAIN is a software standard that applies to scanners and cameras, along with many graphics programs and even some word processing programs, and facilitates communication between cameras and scanners and a computer. Any software that is marked as TWAIN compliant can work with any scanner or camera that is also TWAIN compliant.

While TWAIN is the most frequently used standard for this, it's not the only one. Windows Image Acquisition, or WIA, is another standard that belongs to Microsoft and basically has the same function as TWAIN-compliant software. A lot of scanners have drivers that will work with both standards.

Because all scanners are similar in many respects, it can be difficult to choose which one is best. There are a couple of things that you can keep in mind when you make your purchase, though.

- What is the dpi or ppi range on the scanner? Be sure that this falls into the range you need to scan the items that you want to digitize. Typically, the higher this number, the better the quality of the scanner is, but you may or may not require a very high dpi for your scans. Note that using a higher dpi than necessary can potentially be a poor decision, as the scan time will be slower and the file size will be larger.
- What is the bit depth of the scanner? That is, how many bits are used to store information about each color? The higher the bit depth, the more colors are stored.
- Do you need color? Gray scale? What color range is offered by the scanner, and can it render images into two-tones only (bitonal or monochrome)? Which of these best suits your needs and will scan your collection most effectively?
- How large are the items that you need to scan, and how large is the scanner bed? Bigger is better if you need to scan large items, since trying to scan an item in parts and splice it together in a photo-editing program is tedious at best.
- Is the scanner designed for the operating system on the computer that you're using? Sometimes the manufacturer will offer drivers (software that you need for the machine to communicate with your computer) for a variety of operating systems, which are available for download online, but this may present extra challenges. It's typically easier, when possible, to just purchase a device designed to be compatible with your operating system.
- How fast can the scanner operate? How important do you find the speed to be? The manufacturer may list how quickly a scanner works, but the scanning time might vary depending on what's being scanned (text versus a photo, for instance). A manufacturer will probably list the fastest scenario, which is almost certainly a monochromatic scan.
- When the scanner completes a scan, what image formats can it convert the data into—for example, a tiff, a jpeg, a png? Which ones do you need? If you have to convert formats for your project, this will take time and can result in loss of data.
- What kind of port does the scanner attach to? Does your computer have such a port available? USB ports are typically used on modern machines, but this may be a consideration if using older devices.
- How intuitive is the scanner to use? In the best-case scenario, you can hook the scanner up to a port, install the necessary driver software, and press a button to start scanning. Ease of use makes a project go faster and makes training people to do the work much simpler.
- Does the scanner offer features that you *don't* want or need? Unwanted features may cause confusion when you're trying to set up and learn to operate the device. Extra features may also increase the price, and the more complex an item is, the more difficult repair and maintenance is.
- How expensive is the scanner? It is often unnecessary to buy the highest-quality scanner available; again, what you need depends upon what you will be scanning.

It can be a bit difficult to determine exactly how high the quality of a scan is based on the specifications of the manufacturer. If possible, see a sample scan from the scanner.

While the basics of how scanners work are the same from scanner to scanner, there are a few different types of scanners, and each of them offers different benefits to you for your archiving project. Along with deciding what features you need, you'll also need to pick out a general model of scanner.

TYPES OF SCANNERS

Flatbed

Sheet-Fed (Sheet-Feed)

Portable or Handheld

Overhead

Drum

Specialty (Film, Negative, Microfilm Scanners)

Flatbed Scanners

Flatbed scanners are pretty common and many people have them in their homes for scanning photos or documents. These are horizontally oriented scanners with a flat glass bed. When you use such a scanner, you place whatever object you want to scan on the bed, close the lid, and scan it. The fact that these scanners are so common can be an asset to you, particularly if your archive is on a tight budget. Most stores that sell computer equipment will have flatbed scanners available at a reasonable price. Often, they're part of an "all-in-one" machine that can perform multiple operations, such as printing, copying, or faxing. Typically, it's best to have a scanner that is just a scanner, since devices with more than one function tend to be a bit delicate and break down more easily, but you may find that you like the convenience that getting a device with multiple functions offers for your archive. If you need a printer, copier, and fax machine as well as the scanner, it's a nice space saver.

Flatbed scanners are good for scanning just about anything. You could lay a book flat on the scanner and get a scan of the open pages (not a suitable idea for a delicate book, though), scan photographs, pages or papers, or, if it's large enough, items like newspapers or maps. Using a flatbed scanner is tedious and time-consuming, but it's certainly versatile and is useful for most situations.

It should be noted that flatbed scanners inherently have some problems when scanning books. They can cause wear on the books, transfer dust, crush spines, and may have poor image quality in the spine of the book. There are scanners intended for scanning books, however. Such scanners can have a wide variety of features, including software to improve scanning for books specifically, or adjustable lids, or they may be shaped to cradle books during the scan.

Sheet-Fed Scanners

Sheet-fed scanners have a vertical orientation rather than a horizontal one. These scanners are faster to use than flatbed scanners; with a flatbed scanner, you must open the lid, place the item on the scanner, straighten it, close the lid, scan the item, then remove it and replace it with the next item. Though all these tasks are very easy, they take a long time and become tedious for whoever is doing the scanning.

In contrast, with a sheet-fed scanner, you place the item that you want to scan into the scanner. The scanner runs the item through the device, scanning it as it goes, and the item comes out at the bottom of the scanner. The automation involved speeds up the process. It doesn't take away the tedium of the task, but speed and efficiency are valuable aspects of any part of a digitizing process.

Sheet-fed scanners are particularly nice for digitizing large collections of text. However, they are limited in comparison to a flatbed scanner. You can't use them to scan anything really delicate, since the item could get damaged by the scanner when being pulled through the machine. If you want to scan a book of any kind, you'll need to take it apart, essentially destroying the original object. These scanners are best used for quickly digitizing text in particular.

Portable or Handheld Scanners

You might not think that having a portable scanner is of use to an archive, but it could be, depending upon what you need to scan. Portable scanners are exactly what they sound like—scanners that you can easily move around. These contain the photosensitive device within the scanner, but rather than it moving automatically under an object that you place in the scanner, as in a flatbed model, you move the scanner by running it over the desired object. Software for the device corrects (in part) for human error, such as jiggling the scanner or not moving at a perfectly even pace.

This kind of scanner won't get the crisp, perfect representation that a flatbed or sheet-feed scanner will. Even the correction software and a steady hand will leave some distortions in the final image. However, it does have some practical uses, and so you may want to consider purchasing one in addition to other scanners and software.

If, for any reason, you need to scan objects and they can't be brought into the archive, a portable scanner can collect the information that you scan in a memory card (these were covered in chapter 10) and you can bring the data back to the archive for processing later. You don't have to bring a computer of any kind with you, just the scanner. It should be noted that there are also such things as portable scanners with sheet-fed characteristics.

Overhead Scanners

Overhead scanners are just what they sound like—they are positioned over the object that they will scan. This makes them really ideal for scanning books in particular, as there is no lid to crush a book and the pages can be turned quickly and easily by the operator, saving time. It is also possible to scan other types of objects with such a scanner. While this is an excellent choice for a number of situations—book scanning in particular—it should be noted that this type of scanner is more expensive and much less common than a regular flatbed scanner.

Drum Scanners

Drum scanners are scanners typically aimed at photographers, and may be of use to you if you want high-quality scans of photos, transparencies, or negatives. These scanners consist of a drum that spins rapidly during the scanning process, and the item that you want to scan is mounted within the drum. As mentioned a little earlier, these scanners use a device known as a photomultiplier tube, or PMT, which is more sensitive than other types of photosensitive devices and produces very high-quality scans. However, they're far less commonly found than your other choices, and tend to be large and very pricey.

Specialty Scanners

The scanners discussed so far are capable of scanning a decent variety of items, but you may find that you need a scanner that's designed to get a scan of something unusual. If you need to get copies of slides, for instance, it may be in the best interest of your archive to simply purchase a dedicated piece of hardware: a slide scanner. There are specialty scanners for other items such as photographic negatives, microfilm, and microfiche. Purchasing the correct equipment can save time and result in better quality images for your collection.

Cameras

While scanners are very convenient, cameras are quite versatile and can be of great use to you in your project. For instance, suppose that you have an item that is simply too large to be scanned, such as a newspaper. Instead of attempting a scan, you could take a photo instead. While you might have a hard time getting a normal film camera to capture a high-enough resolution to, say, read text, modern digital cameras capture a very high number of pixels in a shot. When you zoom in on the image using a program, if the shot is clear, you'll be able to read the text.

If you have an item that is too delicate to scan, a camera works for digitization, as well, since you can treat the item more gently than you would an item for scanning. Remember, a flatbed scanner can damage a book and a sheet-fed scanner requires taking a book apart. Like an overhead scanner, a camera can capture an image of a book without putting strain on the book.

Cameras can also be used for recording items that don't lend themselves well to scanning, including items that you might not have considered for your collection. For example, suppose that your archive is part of a university, and you want to keep a record of the art collections there, which includes a large collection of pottery. You can use your camera for something like this, as well. Other kinds of artwork, like paintings, are also well suited for digitization via camera.

The quality of your camera will have an impact on the quality of your captured images, and you should take into consideration several aspects of a camera.

Is the camera intuitive? Like a scanner, you want your camera to be simple to use and to train others to use. You probably won't want or need a lot of special features.

How does the camera connect to the computer to download images? USB ports are common; be sure that you get one that matches your needs. If it's convenient, it is also possible to take out a memory card and put it into a computer to get the images.

How many megapixels does the camera capture? The size of the image a digital camera captures isn't measured in ppi or dpi (which would make your life infinitely simpler), it's measured in megapixels. Typically, more megapixels are better, but this is only true up to a point. In general, a larger *sensor* is more important than more megapixels for getting a better image. The sensor is the part of the camera that detects light from the subject, and a bigger sensor can detect more light and produce a more accurate image. As an example, smartphones often have cameras that are capable of the demands of an archive as far as resolution goes (the number of pixels), but you still might get better results with a dedicated camera due to sensor size (Dolcourt, 2013).

What are the ISO, shutter speed, and aperture of the camera?

1. *Shutter speed* refers to how long a camera's sensor is exposed to light. For archiving, a fast shutter speed is generally better (slow speeds can be used for long exposures, getting pictures in dark conditions, and similar situations). Shutter speeds are indicated in fractions of a second (Plicanic, 2014).
2. The *aperture* is a physical part of the camera, an opening in the lens that controls how much light enters the camera. Apertures should be adjustable and are measured in *f-stops*, with a smaller f-stop letting in more light and a larger one letting in less light. A large f-stop can be better for focusing on a subject (Plicanic, 2014).
3. The *ISO* indicates how sensitive the camera is to light, with higher ISOs being more sensitive. While shutter speed and aperture also apply to film cameras, ISO is a digital-only number (Plicanic, 2014). A more sensitive camera with a higher ISO is generally better.

The technology for smartphone cameras has vastly improved, and if you can't have a dedicated camera, many of the more recent smartphones will work just fine. Dedicated cameras do offer more control and higher quality, so having one may be appealing depending upon your archival needs (Abbot, 2018). You can even get tripods and other helpful accessories for a smartphone. Remember that security may be a consideration with smartphones, though, as smartphones are typically connected to the Internet.

If you use a camera to digitize items, there are a few other tools that you may need, particularly if you're going to digitize fragile books. For instance, you may want to invest in a book cradle, which will safely hold the book while you capture images for digitization. You may also want something that will unobtrusively hold down pages while you capture images. Again, there is a problem with digitizing books in that the curvature of the page may cause distortions in the image, and holding it down can help reduce this. Bone handles and vacuum plates are both often used to compensate for this, but you have other options (Lee, 2001). Software may also be designed to compensate for this problem.

When you use a camera for digitizing objects, the ideal situation is to have a permanent setup, a location in your archive where a platform for the objects that you want to digitize and the camera to digitize them are permanently located. The camera can be positioned so that it's at the ideal height and orientation, and any time you want to digitize items, the camera is ready.

The camera needs to be oriented parallel to the surface where you will place the object that you want to digitize. For instance, you could have a table to lay the items on, and orient the camera directly over them. You can also use a book cradle to hold open books

for digitization. It's important that the camera maintain its distance from the item to be digitized and that it doesn't move during the digitization process. There are a number of ways that you can go about doing this; for instance, you may want an arm that holds the camera over the item. You could also use a plain photo tripod, though you'll have to ensure that it doesn't get jostled.

With a permanent setup, you may want to directly connect the camera to a computer so that it downloads images into the computer as you work; you can also potentially operate the camera directly from the computer rather than pushing a button on the camera. If this seems inconvenient, it's also possible to take the SD card out of the camera and transfer the images to a computer if you have a port for the card or if you have a device to read the card. SD cards were discussed in chapter 10.

You'll also need a good source of lighting. Lighting must illuminate evenly and not create any shadows. Heat can be a concern for your materials, particularly brittle or fragile ones. The bulb you choose makes a difference; LED lights, for example, do not generate significant amounts of heat, and so can be a good choice for your setup.

Software

In order to process and store the data that you create, you'll need software, but what kind will depend upon what you need to archive.

Images

For images, you'll need photo-editing software. The program Photoshop is often held up as the paragon of a photo-editing program (and for good reason), but unless you also want to do something like digitally restore or enhance your photos, you don't really have to use a program as large and complex as Photoshop. To make an image suitable for archiving, the program you use needs to be able to do a few things.

- Your program needs to be able to control the level of black and white and the levels of different colors precisely.
- Your program needs to be able to de-skew an image—that is, it should allow you to rotate an image by a few degrees should you accidentally scan a photo crookedly (or possibly if the original object is skewed).
- Your program needs to be able to crop an image to the meaningful data, erasing everything that isn't meaningful (such as the lid of the scanner, should the item be smaller than the scanner bed in the case of a flatbed scanner).
- Your program needs to be able to read, and possibly convert image files to, the image file format that you have decided to use for your archive (such as a tiff, jpeg, etc.). You may want to save image files in several different formats for different uses, and your program's ability to convert formats can be important.

Your computer may come with software that is good enough to cover the basics. Find out what software you already have available to you and if the software has any limitations (both software limitations and limitations placed by the terms of use) that would make it unusable for your archive.

Text

Deciding what kind of software you need for storing text is a little trickier, because there are many things that you can do with text. Most likely, though, you'll be interested in software that reads and creates PDF files or the archive-friendly PDF/A file formats, as PDFs are one of the most convenient ways to store text for archiving.

The Adobe company created the PDF format. They offer a program, conveniently named Adobe Reader, that will read PDFs; this software is free online for download from their website. They also sell software that will edit and create PDFs, called Adobe Acrobat. While Adobe is the company that created this file format, there are many other companies that offer software that will create PDFs. If you're on a tight budget, it may be in your best interest to shop around; there are a number of companies making software that will read or create PDFs and PDF/As. Otherwise, it may be better to use Adobe for the sake of simplicity—using a common software product can also be beneficial in that resources for troubleshooting or effectively using all the features are more easily available.

You may need word processing software. Some software is typically included on computers when you purchase them, but you may need to buy updated software. If you're storing born-digital text, you may also be interested in software that converts one file format to another. For example, with this kind of software, you may be able to “rescue” documents with an obsolete (or unusual) file format, or you may be able to convert a proprietary format to a more archive-friendly one. Sometimes regular word processing software is capable of this kind of reformatting, as well.

If you're scanning text documents and want to convert them to searchable text, you'll need Optical Character Recognition (OCR) software. This is software that converts an image of text to searchable text, allowing the user to search for words within the document. This technology was discussed in chapter 4. Sometimes, OCR technology is offered as part of other software or is bundled in. For instance, Adobe Acrobat conveniently offers OCR technology as part of the software for creating PDFs. Microsoft Office, which is a set of several helpful office programs, offers OCR technology, too. You may want to purchase software specifically for converting text in this manner anyway, and you'll need specialized software if you want to attempt converting handwriting or unusual fonts, as these are beyond the capabilities of typical OCR software. However, before you do purchase OCR software, find out if you already have OCR software or if you will obtain it through another program that you'd like to purchase. You could save some money this way if you don't need to convert unusual text.

Audio/Video

Digitizing an audio or video collection is a little more complicated and a bit less straightforward than digitizing text or images. If you have an audio or video collection, then there are a number of ways to go about digitizing the collection and storing a digital copy, so there's a challenge in deciding how you want to do it.

If your collection is on CD, DVD, or Blu-ray, then, of course, all you need to do is to “rip” the data from the physical object onto a computer hard drive, then store it in your desired method of data storage. If the data is on a tape or record or is on film, then you have a more challenging task.

Audio

Some types of audio recordings are so old that you're better off leaving it to a professional to get a digital copy for two reasons: first, older recordings are going to be more fragile and delicate, and a professional will be able to handle the recording appropriately (E-MELD, 2006). Old wax cylinders and early records would be examples of this kind of a situation. Second, you must have a device that is capable of playing back the recording in order to digitize it. This means specialty equipment, which may be difficult or next to impossible to find.

More recent recording methods, like cassette tapes and records, are probably within your capabilities of capturing digitally. The general method of getting the audio data for both of these types of recordings is basically the same.

Hardware

Digitizing audio materials has become quite prevalent, which is beneficial to you, as you have some options for how you can go about such a project. Rising popularity for vinyl records in particular has led to a variety of options for consumers who wish to connect their record players to different types of audio equipment, including computers.

Digitizing cassettes is a fairly straightforward process, and you can get adequate results using an ordinary cassette player. The player must have an output port of some type (a normal 3.5mm headphone port or an RCA port with the white and red connectors). You can then purchase a device that will convert the audio to digital information on the computer by connecting from one of the player's ports to the computer's USB port (Nield, 2019). Remember, audio on a cassette tape is analog (continuous) in nature, and needs to be broken up onto bits of data for processing on a computer (digital). The device you select will do this for you.

While an ordinary player will work, it should be noted that you will get better results with a high-quality player or with a device specifically designed to convert cassette tapes to digital files. It's important to get a device that will play back the recording as faithfully as possible. In the case of a cassette tape, you'll need to know if the tape was recorded in mono or stereo (monophonic or stereophonic); more information on what exactly this means is in chapter 5. You'll get the best results by using a stereo deck with stereo tapes and a mono deck with mono tapes. You'll also need to know at what recording speed the tape was recorded; most tapes were created with the same recording speed, but some were recorded at a lower speed and will require a player that can compensate.

If you don't have experience with records, you may not know that both the records and their player require cleaning, as dust can accumulate on the record and the player's stylus. There are specialty brushes available for cleaning records and equipment, as well as cleaning solutions designed for records (some common cleaning chemicals will damage a record). Cleaning can improve sound quality (Schiff, 2019). Cassette tapes are somewhat self-cleaning; there is a little piece of felt in the cassette that clears dust. Players can be cleaned with a specialty cleaning cassette, which is shaped like an audio cassette but is intended for cleaning the player.

For records, there are a variety of options for turntables that have ports that can be connected to a computer. The simplest of these is a USB port, as cords intended to connect one device with a USB port to another are very common. Otherwise, you'll need a cable

that can connect from the player's output to the computer's audio input (a microphone port) (Hanson, 2019).

It may also be desirable to get something known as a *preamp* (also known as a *phono stage*) if one is not built into the player itself. This is a device that improves the audio signal from the record player for another electronic device; in this case, a computer. It may be desirable to get one even if a player does have a built-in preamp, as a good-quality preamp can improve the sound quality and fidelity to the original.

Like tapes, record recordings are analog in nature, and thus their data will need to be converted into digital signals. An analog-to-digital converter is a device that can assist with this. For records, one may be built into the player or the preamp, but it can also be purchased separately if needed (or if a higher quality is desired than what the built-in one offers).

Other hardware items that you will need in order to acquire a good sound recording are speakers and/or headphones. You need to be able to listen to your recording to see if the quality is good, and this is especially important if you need to make any adjustments for volume or clarity. Either will work, but headphones are a better choice; you can listen more closely to the sound with headphones and you won't bother others in the archive as you test the recording. The quality of the speakers and headphones is very important to determining the quality of the recording.

Software

While the hardware is an important component, you must have software to capture the sound as well as to process it afterward, just as you had photo-editing software for an image. If your computer has a microphone port, it most likely already comes with software that will capture audio, but you may want something more sophisticated so that you can capture the optimal sampling rate and speed.

Recording from cassette tapes and records will involve playing the entire tape or record from beginning to end; it cannot be done more rapidly than the play speed, as is possible with a CD.

As with photo-editing software, you probably won't need software that does a lot to change the recording unless you are attempting to restore a damaged recording or are making sound enhancement part of your project. You'll need to be able to control the sampling rate, recording speed, and bit depth, and you'll need to be able to make some basic changes, such as adjusting the volume on the recording and dividing your recording into individual tracks; for example, a cassette tape plays continuously from the start of one side to the end, even though there might be several different parts (like different songs). You'll likely want to divide your recording into these logical breaks in order to improve searchability and to make it easy to find exactly what your patrons are looking for. In addition, you'll need to know if the software allows you to add metadata to help your patrons find your audio files; software that allows for metadata is more helpful than software that does not. Metadata allows you to attach extra information to a file, such as creators, dates, and search terms. (Metadata and its importance will be covered in more detail in chapter 13.) You have quite a few options and some of them are free—for instance, Audacity is a free software program that will fulfill the basic requirements. Plenty of software is available that is capable of more sophisticated audio capture and editing, as well.

Video

Like audio capture, digitizing video is essentially going to involve playing what is on the video and capturing it digitally by recording what plays. Again, as with digitizing audio, it's probably better to go to a professional for very old or very delicate recordings. Digitizing video can be more complex than audio, since you are attempting to capture both images and sound at the same time.

Hardware

For a video that requires a reel-to-reel projector, like Regular 8 or Super 8 film, you have two options. The first is to use a device that will convert film to a digital format, essentially scanning each frame and rendering it digitally. This type of equipment is quite expensive, but potentially worthwhile if you have a lot of reel-to-reel film to digitize (Murphy, 2018). Note that there are a few different sizes and types of film, so it's important to be sure that the equipment can handle the size you have.

The possibly cheaper, but more complicated, option is to play the video using a projector, then capture it using the digital video camera. In this case, the quality of the camera is extremely important. You'll also need a screen or a surface to project the image onto, as well as something to set the camera onto and hold it steady, directly parallel to the projection (Herrman, 2012). The concept behind this is rather like using a camera to capture an image when digitizing photos, paintings, or delicate materials.

The audio in such a situation can be captured either directly from the speakers, which is not the ideal situation, or from the speaker's line output, which is more optimal (Herrman, 2012). As with digitizing audio, you'll need to connect the speakers to a microphone port in your video camera. It should be noted that a specialized converter may not capture sound, and so any sound would need to be captured separately—again, via speakers or the output.

Capturing audio and video, especially in older formats, tends to be more complicated than digitizing other materials, and if you have old reels of film to digitize, this is another situation in which you may just want to outsource the project.

If you are digitizing VHS tapes, then things become much less complicated. As with audio cassettes, you can digitize a VHS tape using a normal VHS player and a device that will connect the player to a computer and convert the data from tape output to a file that the computer can process. Again, like cassette tapes, there are likely two different options for ports on the player: an S-Video port or RGA ports (a yellow one for video output, and one red and one white one for audio) (Nield, 2019). Note that other, similar formats, such as the Betamax, can potentially be converted in a similar way (Murphy, 2018).

Software

If you purchase a device to convert video to a digital format, it will most likely come with appropriate software. However, it may or may not have any editing features. As with audio recordings, you probably don't need very many features. You will likely want to be able to adjust colors or volume, divide a capture into shorter tracks, and add metadata (which will be covered in more detail in the following chapter). You may also need something that can help with audio synchronization if you needed to capture the audio and video separately. In addition, you may need the software to be able to create a certain file type, depending upon what files you decide to use for your archive.

Key Points

- A digitizing project requires equipment, and you may wish to outsource part or all of your project to a vendor, who will be able to use the optimal equipment available to digitize your materials.
- Some of the equipment needed for a successful project includes digital cameras, scanners, video cameras, music players, and film projectors and players.
- A digitizing project requires software. Some equipment comes with the appropriate software, and most computers have some software that you can use, but may not be ideal or of a high-enough quality for your project.
- Most archiving projects do not require software that is capable of a lot of editing, since the goal is typically to capture an item as is, not to change it—but quality is still important.

After creating a digital version of an item, the next step is to make it accessible. In the following chapter, you will learn about metadata and how it pertains to digital materials specifically, and you will learn some basic information regarding patron access of your digital collection.

References

- Abbot, James. 2018. "Smartphones vs Cameras: Do You Still Need a DSLR?" TechRadar. <https://www.techradar.com/news/smartphones-vs-cameras-do-you-still-need-a-dslr>.
- Dale, Nell, and John Lewis. 2013. *Computer Science Illuminated*. 5th ed. Burlington, MA: Jones & Bartlett Learning.
- Dolcourt, Jessica. 2013. "Camera Megapixels: Why More Isn't Always Better." CNET. http://www.cnet.com/8301-17918_1-57423240-85/camera-megapixels-why-more-isnt-always-better-smartphones-unlocked/.
- Edwards, Benj. 2010. "A Brief History of Computer Displays." *PCWorld*. <https://www.pcworld.com/article/209224/historic-monitors-slideshow.html#slide1>.
- E-MELD. 2006. "E-MELD School of Best Practice: How to Digitize Analog Audio Recordings." <http://emeld.org/school/classroom/audio/howto.html>.
- Fuller, Floyd, and Brian Larson. 2008. *Computers: Understanding Technology*. 3rd ed. St. Paul, MN: Paradigm Publishing.
- Hanson, Matt. 2019. "How to Convert Vinyl into MP3." TechRadar. <https://www.techradar.com/how-to/how-to-convert-vinyl-into-mp3>.
- Herrman, John. 2012. "Digitize Your Home Movies before They're Gone Forever." *Popular Mechanics*. <http://www.popularmechanics.com/technology/how-to/tips/digitize-your-home-movies-before-theyre-gone-forever>.
- Lee, Stuart D. 2001. *Digital Imaging: A Practical Handbook*. New York: Neal-Schuman Publishers.
- Murphy, Laura. 2018. "How to Convert Film and VHS to Digital." *Consumer Reports*. <https://www.consumerreports.org/audio-video/how-to-convert-film-and-vhs-to-digital/>.
- Nield, David. 2019. "How to Digitize All Your VHS and Cassette Tapes." *Popular Science*. <https://www.popsoci.com/how-to-digitize-tapes/>.
- PCMag*. n.d. "Definition of: Drum Scanner." Accessed September 28, 2019. <http://www.pcmag.com/encyclopedia/term/42020/drum-scanner>.
- Plicanic, Khara. 2014. *Getting Started in Digital Photography*. San Francisco: Peachpit Press.
- Schiff, James. 2019. "For the Record: How to Clean and Care for Your Vinyl Collection." *Rolling Stone*. <https://www.rollingstone.com/music/music-news/how-to-clean-vinyl-re-cords-850080/>.



Metadata and Accessing Information

IN THIS CHAPTER

- ▷ What is metadata? How is metadata for digital items unique?
- ▷ What are the general types of metadata?
- ▷ What are metadata schemas? How are they useful?
- ▷ What are the considerations to make regarding patron access to a digital collection?

If you have a collection of any kind, one of the main obstacles that must be overcome is the problem of finding what exactly it is that you're looking for within the collection. If you need information about agriculture in medieval Scotland, how do you know which book might have it, and where it is?

It's a problem that has been addressed in a variety of ways over time, from nearly the beginning of writing itself. There is a Sumerian clay tablet approximately 4000 years old that is essentially a record of a collection of works of ancient writings. Even the scrolls in the library of Alexandria were labeled with tags indicating the title, author, and subject of their contents (Library of Congress, 2017). One of the more recent solutions is the *card catalog*. Card catalogs were very prevalent before computers came into common use; you may have used one, or possibly still use one or know of a library or archive that still does.

A card catalog can tell you a lot of useful things. If you've never used one, a card catalog is essentially a set of small cards with information written on them. Each item in the library will appear on at least one card. For instance, if you're looking for a book by Mark Twain, you can browse through the cards for the name "Twain" to find books written by Mark Twain. If you knew the name of a book instead, you could search for the title card for that book, which would indicate the author's name on the card.

There were a variety of useful ways to search for items, with author, topic, title, and publication year being common. Again, before the rise of computers in everyday life, this was convenient and is still a good system (although not as good as a digital catalog).

Most libraries now have a digital equivalent of the card catalog. Instead of browsing through cards, you can just type keywords into a search window and the computer will retrieve matches for your search. Just as with the old card catalog, you can typically refine your search by author, topic, title, and so on. This information is *metadata*.

One of the important steps in developing a digital collection is to make it possible for your patrons to find the items that exist in that collection. This is where *metadata* can be important. Metadata is essentially extra information about a file and what it contains. Adding metadata to a file will not only help your patrons, but will also help you when you try to locate items or evaluate your collection.

Metadata

So, what is metadata? It's sometimes a tough concept to get your mind around, because the common way to explain it is that it's data about data. But what does that mean?

As you learned in chapter 2, a file on a computer is simply a mass of electrical impulses that can be interpreted as ones and zeroes. The computer itself doesn't know what these are; it's up to a human to make sense of the information.

That's where metadata comes in. Metadata can be information such as whether a file is a music file or whether it's an image, how many pages it contains or how large it is in the case of a book with a physical format, or who created the item, such as an author or an artist. All of this provides information about the item that is useful to the user.

Your files will contain some metadata upon their creation. Your computer will be able to detect a few things about the file, such as how large it is, how old it is, and the file extension, which indicates which programs are appropriate to open the file.

Metadata in Everyday Use

People who use the World Wide Web often use metadata without even realizing it. For example, suppose that you have an account with a website like Instagram, which allows its users to share and store images. You load an image to your account of an adorable brown and white beagle puppy chasing a tennis ball. The site may then prompt you to use *tags*, or words that accurately describe the image for users who might also want to see the image. You decide to oblige (you don't *have* to add tags in most sites like this) and you type "puppy" (the subject), "beagle" (the breed), "brown, white" (the colors), "tennis ball" (an object in the picture), "chasing" (what's happening), and "adorable" (an opinion, but in this case still a helpful descriptor). Now, if another user comes along and decides to do a search on this site and types "adorable brown puppy" into the search box, your photo will be one of the ones that show up because you used those words as tags. That is, you have told the search engine that you are looking for photos of an adorable brown puppy using your search terms, and those terms match the tags you added for the photo.

This puppy photo, for a computer, is just a series of information regarding the colors of pixels to display. Detecting a subject in an image is a complicated task for a computer, and your tags are a type of metadata that helps human users find the photo by describing it in a meaningful way.

The technology for identifying objects in photos is improving rapidly. “Object detection” refers to the concept of software programs determining what items are in a photo. Some programs can also create suggested tags for a photo. This type of software will likely become more prevalent and accurate in the future, and can potentially be useful for a digital collection, either for more accurately describing subjects or for more quickly adding relevant metadata.

On the social media site Twitter, people can communicate with others using short messages known as *tweets*. People using this site use words known as *hashtags*, which are words or series of words preceded by a # symbol, to describe what their tweets are about or who their tweets are designed to communicate with. This can create “trending topics” when many people use the same hashtag descriptor. This can actually be very relevant, as it might help get information out to many people regarding something like a natural disaster, as the unique hashtag used for the tweets will make it easy to look for up-to-date news from people using the Twitter service. For example, there is a #wildfire hashtag used to help spread information about wildfires, such as the devastating California Camp Fire.

Sometimes people use the hashtag for humorous intent, too—for instance, they might describe a situation in which they made a social blunder and use a hashtag like “#awkward.” Though not exactly used in an efficient or professional manner, these hashtags provide useful metadata and can assist users of the site. These hashtags are metadata that describes the general contents of a message, helping users find desired and relevant information.

Why Is Metadata Necessary?

When you search for information using a search engine, what essentially is happening is that the terms you use are matched to keywords from the search engine program. Your main purpose with adding metadata to your digital information is to make it possible (and hopefully simple) to interpret what is in your collection as well as to make it possible to locate relevant items. Without metadata, you’ll have a lot of files, but you won’t necessarily be able to do anything useful with them.

Computers are becoming more and more sophisticated and can potentially add meaningful metadata to files without human help, too, but this technology is not perfect and humans can make it much easier for computers to describe files in a meaningful way. In addition, search engine technology is improving all the time. For example, online search engines are starting to be able to predict what sites might interest you or what sites might be able to answer a question based on your search terms and even the order in which you type them.

Metadata in Libraries

Cataloging and adding metadata are overlapping concepts. Creating a catalog entry creates a type of metadata, and it’s basically all you need to do with tangible items in a collection, such as books or CDs. The catalog entry will have all of the information that you need about the item. However, digital items can be a little more complex.

Traditional metadata is separate from the item it describes. Again, think of an old-fashioned card catalog in which the information about the books and other items in your collection is written on cards. You wouldn’t keep cards with the books they describe; that wouldn’t make sense, as you are using the cards to find the books in the first place.

Although keeping cards with the item makes no sense for tangible items, it can work just fine for digital ones. It's possible for computers to attach data to an item but have it remain "invisible," just as the tags in the earlier example of the photo of a puppy are. The tags aren't written over the image, labeling it "a picture of a puppy," they're simply attached to or part of the file so that a computer knows that it's relevant information to anyone searching with the keyword "puppy."

As an example, many people who create web pages add metadata content to the web page. It serves no purpose whatsoever to the usability of the web page and does not change the layout in the slightest, but it adds information that search engines can use to decide if the web page is relevant to a person's search. For example, suppose that a person named Robert Smith discovered an amazing way to make pancakes and wanted to share it with the world, so he created a web page to show how to make his pancakes. To make his web page easier for people to find, Robert Smith adds metadata using the following HTML code:

```
<meta name="description" content="Amazing Pancakes">  
<meta name="keywords" content="pancakes, breakfast food">  
<meta name="author" content="Robert Smith">
```

This code helps the search engine determine whether this site is relevant to a person when they conduct a search online. Although it's meant for a computer to read, you don't have to be an expert in HTML code to get the idea of what this means. The "description" here indicates to the search engines that the page is about amazing pancakes. The "keywords," pancakes and breakfast food, are helpful, too, and might match keywords that a person types into a search engine. The "author" code says that the author of the page is Robert Smith. Again, this is all invisible to the user unless they look at the page source; instructions on how you can do this if you're curious are given in chapter 4.

Metadata for Digital Items

Along with metadata that you add yourself, sometimes software adds its own metadata to a file without instruction from the user. Many digital cameras will automatically add metadata, unbeknownst to you, to whatever photos you take. This can include information such as the camera model, when the image was taken, or even where the image was taken (many smartphones have GPS capabilities that can add this type of data to an image). Digital video can record much of the same type of information—again, invisibly. To make things simpler, many computer programs will add at least some relevant metadata to a digital item.

It is possible for metadata to present privacy concerns, especially automatically generated metadata, as users often don't realize it's being created. For instance, a smartphone that takes digital photos might add the time the photo was created and where it was taken. This might be problematic; if a person takes an image inside their house and posts it online, for instance, this could reveal the location of their home and/or personally identify the person who took the picture. On occasion, it is possible and potentially desirable to remove metadata.

But in many instances, you'll need to add the meaningful information yourself. This can be a daunting task, especially if what you have in mind is a large project. How can you

decide what the meaningful subjects of a photo are? How can you determine what the best keywords of a book should be? What if you don't have much information about the item at all?

There are some shortcuts that you can take regarding digital materials that are impossible to use with tangible ones. For example, say you had a PDF created from scans of a book of recipes. If the text is searchable, users can search for specific words and find the location of recipes in the book (enabling computers to search for text on an image of text was discussed in chapter 4). There are software programs that can help you determine what some good keywords are for an item.

It's possible to automatically generate metadata from a wide variety of file types, including audio and video. For instance, a software program can detect words from spoken audio and transcribe it, making it possible to search for an audio clip based specifically on what was said in the clip rather than relying on keywords to describe the clip (NISO, 2017).

Using and creating metadata does not have to be difficult. Remember, while there is a lot of useful information you can attach to a data item, your main goals should be making it possible to search for relevant data and to organize your collection in a meaningful way. If your time and resources were very limited, you could simply add some relevant keywords to the file name or use a highly descriptive title. For example, if you had a photo on your computer of an elm tree that was hit by lightning in the summer of 1937, taken by a person whose last name was Smith, you could name it "photograph elm tree lightning strike summer 1937 Smith.jpg," and it would come up if you did a search on your computer for any of those words in the name, such as photograph, elm tree, lightning, and so on. Computers can pick out individual words from a file name, and while typing in "elm" would bring up every other image of an elm tree on your computer, it's an easy way to get your collection organized and useful to users.

Some photo-editing software allows you to add metadata to photographs. Photoshop, a program discussed in previous chapters, is one of them, but there are others. When you add metadata to a photograph, your image will become more searchable. For example, in the above scenario, in which you use the file name to make the image more meaningful, you're limited by what you can put into the file name. Computers don't like it when file names get too long, and early computer files were very limited on the allowed number of characters, which would make this particular method difficult or confusing. If you can add metadata to a photo, then the name of the file becomes less important. You could add all the same information—that the image is a tree hit by lightning in 1937 and the photographer was Smith—but have a shorter file name, such as "image37.jpg." Searching for the keywords in the associated metadata on your computer will still bring up the image that you wanted, regardless of the file name. This can help you name your files more efficiently, as the actual name of the file will be less important.

When using photo-editing software in this manner, it's better if the metadata is *embedded* into the photo—that is, it's permanently attached to the image file. If software merely *associates* metadata, then the file for the metadata and the item are two separate files, and you may be unable to access the metadata with programs other than the one you used in the first place. The program that you used will merely associate certain metadata with a file name (Ashenfelder, 2013).

Usually, using more descriptors is better. You can have metadata that helps your archive in particular, rather than your patrons, like letting you know if something is part

of a series, if an item has been damaged, if an item has only been partially digitized, and so on. You may want a more in-depth method of describing your data than simply a descriptive file name, in which case, you'll be interested in some of the metadata schemas that are in use today.

Metadata Types

There are many ways to categorize metadata, and some people break down metadata into many different types. Though it's not a formal way of dividing things, an easy way to think about metadata is to break it into three categories: descriptive, structural, and administrative (NISO, 2017).

Descriptive metadata. This is the easiest type to understand. This is anything that simply describes the item that you've archived for the purpose of making it easy to find. Things like titles, authors or artists, and keywords are all types of descriptive metadata.

Structural metadata. This describes the structure of the item or how it's put together. For example, it could describe the order of pages or chapters in a book. It could also note what versions are available (both a JPEG and a PNG, for instance) or things like scan resolutions (Taylor, 2004). If it describes what the file is and how it can be displayed, it's probably structural metadata.

Administrative metadata. This is typically information that's really only useful to your archive. This can include information about when your file was created, who is allowed to access the file, and other information along those lines. For you, it might also include information about how to preserve the item for archival purposes. If it's useful information, but it doesn't describe the item and doesn't help with defining how the item goes together, generally, it's administrative metadata.

It's up to you to decide what information is or is not relevant and how much data you want to include. For instance, you might have a collection of photos that go together, but only want to create one catalog entry or describe them as a set, not as each individual photograph. You could include the dimensions or even the weight of a book that you want to digitize, or omit that information. You do require a balance with metadata—the more information you have, the better, but the more information you have, the more time-consuming it will be to enter all that metadata and to describe the item you are attempting to save. Whatever you do, be consistent. If you want to include dimensions for one book, you should really do that for all your books. The importance of consistency in your project will be further discussed later in this book.

Metadata Schemas

There is no reason whatsoever that you can't use your own system for organizing your data. If you have a small archive, are limited on funds and the available software and resources, or have an unusual collection, then you might even benefit from coming up with your own system tailored to your needs.

If you want to be able to communicate with and share your data with other archives or libraries, though, then you're probably better off using an already-established metadata schema (or at least being aware of the different options available today). A metadata schema is basically a method of organizing and describing an item in a consistent man-

ner. Sometimes a schema will be designed to describe something specific (images, for example). Consistency is important for making metadata easy to use and compatible with software, and so things like format and vocabulary may be controlled as part of a metadata schema.

Using an already-established metadata schema can make it easier to search for and process your metadata, too, since you can purchase software from a vendor that is designed to work with the standard you choose to organize and search for your information. For example, there is software that offers templates with fields that the user can simply type relevant information into, and then the software will generate the formatted metadata for you.

If your archive already uses a particular metadata schema but you'd like to try a different one or if you are collaborating with another library or archive that uses a different metadata schema from the one you want to use, it is possible to convert metadata from one format to another. A table that helps identify equivalent fields (e.g., the author field) in two different formats is known as a "schema crosswalk." Converting metadata can have a variety of problems, however, such as fields in the original format with no equivalent in the new format.

There are many metadata schemas that are aimed at a variety of items that you might want to describe. For example, there are metadata schemas designed specifically for government documents. It would require an entire book to explore all of your different options, and so this chapter will only discuss a few common ones.

MARC

MARC stands for MACHine-Readable Cataloging (not a perfect acronym, but easier to say than MRC). This kind of cataloging has been around since the 1960s. It was remarkable at the time of its invention because it didn't require a standardized field length (Taylor, 2004). To understand the problem of using a standardized field length, imagine that you wanted to use a cataloging program to enter the title of a book. The software programmer decided that you'd only need forty spaces, or room for forty characters, at the most. But your book has a very long title with a subtitle, and you need more space than this. You'll have to decide what to do, and to compromise on accuracy. Or, suppose that you have a book with a very short title. All of the extra room for characters could be considered wasted space.

But why would a programmer set things up that way? If a computer "knows" that there are only going to be forty characters for the title, then it can process the data you submit and know that, after 40 characters (even if those spaces are blank), it's reached the end of the title and that the next set of data is new information, such as the author or the publisher. It knows what data is what by *where* it is in a series of binary ones and zeroes.

The things you might want to catalog are not so tidy, though. You might want to catalog a photograph, not a book. A book might be part of a series. You might want to make all kinds of notations or enter a variety of data about an item. If you have a series of standard fields, or areas in a form that can be filled out but have a limited amount of space, then you're limited in what useful information you can add about an item. A more flexible system is better, which is why the MARC system was remarkable at the time of its invention.

With MARC records, you can add data that signals to a computer what the information is. You use sets of numbers and characters that can signal to the computer “this is the author” and everything that comes after that set of symbols is then interpreted as the author of the book. In this instance, a MARC record could look like this: “100 Mark Twain.” The “100” part signals to the computer that this is the author, and the author is Mark Twain. The specialized series of numbers or characters signals to the computer that a new field has started; for instance, after “100 Mark Twain,” it might read “245 The Adventures of Huckleberry Finn,” and the “245” part signals that this is no longer the author. This is the title of the book. You can do this for the publisher, the call number, even the retail price of the item (Library of Congress, 2009). These number indicators are known as *tags*, just like the “tags” for social media discussed earlier in this chapter. There are many options with MARC records, and this is good because, as discussed earlier, more information is better toward helping your patrons find what they need.

There are several variations on the MARC system. Librarians in the United States typically use the MARC 21 standard, which is essentially an agreement between librarians in the United States, Canada, and Great Britain to use the same method of cataloging. Previously, each country used its own version of MARC records.

MARC has some advantages: It’s well established, which can be a good thing. Many libraries use the format, and so there are many items already cataloged with this format. This can help you if you want to communicate easily with other facilities, and can potentially save you time and money if someone else has already created a MARC record for an item that you want to archive and digitize.

There are some drawbacks, though. It’s not intuitive. The MARC standard was designed at a point in time when computers were huge and slow. A modern cellphone can do more than one of the computers in use at the time of design of MARC records. MARC records are very easy for computers to handle and process. Computers like things such as using a short set of numbers or unique characters to label information. People, not so much. Humans like words. For a human, staring at a MARC record doesn’t immediately make sense. MARC records also don’t lend themselves well to some things that a modern archivist might want as part of the metadata for an item, such as a book’s table of contents, thumbnails of images, or pictures of book jackets (Tennant, 2002).

Since computers today are much more capable of accommodating human preferences, such as words and pictures, some librarians look to other methods of encoding metadata. MARC records were generally designed to describe things like books. Extensible Markup Language is more flexible in many ways and can potentially include more information, which is good for a digital collection that may include some unusual items that need to be described in a way tailored specifically to the item.

PREMIS

PREMIS, which stands for PREservation Metadata: Implementation Strategies, is one of the major standards specifically for digital archiving (it is not the only standard available for this). This is an XML-based metadata schema. XML, or Extensible Markup Language, was described in chapter 4, but to review a little, XML is a system of tags that are used to describe the content and sometimes structure of a digital file. As a simple example, the tags `<author>Mark Twain</author>` would indicate that the author of the file associated with the XML document is Mark Twain.

You may notice that this doesn't seem a lot different from the MARC method of indicating an author, but MARC uses a number instead. In both instances, a computer program could search through the metadata and look for the indication of "author" (for MARC, the number 100; and for XML, the tags <author> and </author>). There are a variety of ways in which XML differs, however.

You know that MARC fields have standardized lengths. XML fields do not. The <item></item> tag system is how a computer "knows" when a field has started and ended. The first tag indicates the start of a field, and the same tag, but with the forward slash symbol at the beginning, is the indication that a field has ended. This is a very flexible system in comparison to the MARC system.

For the current standard, MARC 21, the field numbers always indicate the same values. Number 100 is always the author, number 245 is always the title. XML-based schemas are much more flexible. You could potentially come up with any kind of tag to indicate any type of information, so long as how you use the tags is consistent and the software you use is programmed to look for the tags you used in a way that is useful and meaningful. That is, if you did use the tags <author>Author Name</author> to indicate that the name between the tags is the author's name, then you could use software that will retrieve the author's name correctly when you search for items in your collection.

XML schemas can include a very wide variety of information, however. It's possible to include general information about the item, such as titles, authors, and copyright dates; information about how the item should be formatted and displayed to the user; and information about how the item relates to other items (the item could be part of a set or a series, for example).

So, PREMIS is an XML-based system of metadata in which the creator of the schema, the Library of Congress, for the specific purpose of creating metadata for digital archiving, defined what tags would be used, how they would be used, and what they would mean. PREMIS allows for some metadata that is unique and very relevant to digital files. For example, there are tags that are used to indicate rights (copyright will be explained in more detail in the following chapter). There are also tags that are used to provide information about how the file can be used—that is, how a file can be opened or a program can be run, and so forth (PREMIS Editorial Committee, 2015).

It should be emphasized that there are many XML-based metadata schemas that are all different, but once you understand one, understanding the others will become easier, as they all work in the same basic way.

Dublin Core

The Dublin Core Metadata Element Set started at a workshop held in Dublin, Ohio. The metadata schema that arose from this workshop is simply known as *Dublin Core*. This particular method of adding metadata has only 15 different categories of information for an item, but is designed to be flexible and to be able to describe a variety of different items. The original object of this particular schema was to address the problem of adding metadata to and cataloging web pages.

The people who created Dublin Core recognized that the number of web pages is growing at an enormous rate, and if anyone wants to store information about these pages or study them in a meaningful way, they require metadata that will help with this process. As discussed earlier in the chapter, there are already ways for web authors to add data

to their web pages with the purpose of making their pages easier to locate with search engines. Dublin Core may therefore seem like a redundant concept, but it's really not.

Because the number of web pages greatly exceeds the number of catalogers who can handle all of them, Dublin Core was created for the purpose of attempting to get the creators of web pages to catalog their own web pages, removing all the work from librarians. Dublin Core is therefore very simple and easy to follow, because the people for whom its use was intended aren't catalogers and don't have the time to create a comprehensive catalog entry, since this isn't what they are paid to do. There are 15 elements in the Dublin Core schema, most of which are self-explanatory, such as "Title," "Creator," "Subject," and "Publisher."

Dublin Core is simple, and that can be a good thing. It's also capable of describing more than just web pages, so if you decide that Dublin Core is a good choice for your archive, you can use it to add metadata or catalog a variety of items. It can also be made more complex; qualifiers are used with the main elements to refine them or make them more specific. The basic version is unqualified Dublin Core, and the more complex version is qualified Dublin Core (referring to the qualifiers, not who is able to use this particular schema).

If you're working with a small staff or staff that doesn't include many (or any) catalogers, or you're relying on student workers, Dublin Core's simplicity and the fact that it aims for general users could be an additional bonus to you. Even if you aren't interested in this schema, knowing what it is can help you if you want to store web pages as part of your archive.

Other Considerations with Digital Metadata

As you know, digital materials are vulnerable to obsolescence, with either the file format or the software or equipment required to read the file becoming impossible to locate. Metadata can help you protect your digital items.

This is another advantage to keeping the metadata with the item—no matter what happens to your card catalog, physical or digital, the data you need is safe with the item, as part of the same file as the item. You can keep adding metadata over time if needed, as well.

But why would you need to do this? Suppose that you are working with two other archives to digitize books, and you have an entry for a 100-year-old book on botany with many illustrations. Suppose that it's many years from now and your equipment has improved, so you'd like to recapture some of the illustrations. But you're working with two other archives. Where is the original, physical item? You're collaborating, so who owned the book you digitized? You can, of course, go through the records of your collections, but this will take time. For something like a book, what do you do if there are multiple copies of the item among the three archives? Logically, you would have digitized the best copy, but who held it? What if the clearest copy had missing pages, so you digitized two different books to get a whole copy? How will you remember?

If you added metadata about when and how the digital item was created and who made it, then this isn't a problem. The metadata for the item can indicate who made the scans and, therefore, who holds the book in their collection and possibly even where it is. If you needed two or more copies to get a good digital copy, you can leave some notes for yourself *with the digitized item*. If, for instance, someone working on the project moves to

a different archive, whoever takes over will have invaluable information with the collection and won't struggle to continue the project.

You can add all kinds of useful information to born-digital collections, too. For example, suppose that you want to archive a kind of software. Software becomes obsolete over time for many reasons, one of which is that the software is no longer compatible with a new computer operating system. You can indicate to future users which operating system it was most compatible with—which will help them determine how they can make the software run, if desired.

As you know, file formats can become obsolete, as well. Imagine that someone created a better format for images—say, that it had all the benefits of a TIFF but took up even less space digitally. All the new software companies want to use this new file format instead of the old clunky TIFF. The TIFFs you have are in danger of becoming obsolete, so you want to save them in this new file format. If you used metadata to indicate that your TIFFs are, in fact, TIFFs and perhaps which program or programs you used to create them, you can start locating those files and either converting them or making notes in their metadata that they require conversion. Knowing which programs you used can help you determine how you can view the TIFFs again or help you start to determine the most efficient way to convert your files.

You could also use metadata to indicate that a file is not optimal. For example, suppose that you had an image of a mayor that you think is important for local history. This photo, though, has a big crease down the middle. You'd really like a better copy, but it's the only one your archive has. Someone from the community might have a better copy, if you ever locate one. You can also indicate that your copy is not optimal in the metadata, as a note to yourself or others in the future.

Metadata helps your patrons locate relevant items. As part of the process of creating a digital archive, you need to decide how your patrons will access your items and how much they will be allowed to do with your digital materials once they have located their desired items. You have many options and many possible levels of access for your patrons.

Accessing Digital Materials

Once you've created and catalogued your digital materials, you need to think about how you want your data to be accessed and who you want to be able to access your data. You'll need to consider this even further if you happen to be archiving anything that shouldn't be accessible to the general public, information that is sensitive in any way, or data that you are simply archiving for the sake of archiving and are not legally capable of making accessible to the public (situations in which this might happen will be discussed in the following chapter).

Information that is available via the Internet is always less secure than information that is not available online. No matter how much security you use or how many promises a cloud computing company makes, data that is accessible through a network is always less secure than information that is not. If you have sensitive information that is particularly desirable and valuable, hackers will attempt to access it. Even if you don't have sensitive information, hackers occasionally will hack systems solely to see if it can be done or for their own amusement. While an archive isn't a particularly brag-worthy target in this respect, it's always best to be prepared. If your archive is part of a university,

for instance, your local bored computer science majors may see even your databases as worthwhile targets.

While Hollywood may portray hackers as reclusive geniuses with programming powers bordering on the supernatural, this is not reality. While there are certainly computer geniuses, it is much, much more common for hackers to access data simply by tricking someone into giving them usernames, passwords, and so on. This tactic is commonly referred to as “phishing,” pronounced the same way as “fishing.” Tricking a user into downloading malware or spyware is another common tactic. Inadequate security software or inappropriately configured software or databases are also ways for someone to access data that they should not have. If you have data that needs to be protected, be sure that anyone working at the archive is up-to-date on the latest phishing tricks and knows how to identify manipulative e-mails and malware.

If you do have information that is sensitive, the best thing to do is to store it off-line. You would archive the item storing your data as an object, keeping it away from a computer, or a reader in the case of magnetic tape. If there is no online access, then there is no way for someone to access your data remotely.

If you'd like patrons to be able to access your data but only want them to be able to do it locally—that is, only when they are within your building—then you could load your data onto a computer that is not connected to a network. This would involve saving it on a computer's hard drive. While it's possible for the data to be copied via items like CDs or flash drives, you can take steps against this, too, such as physically blocking or removing ports and CD drives if you don't want patrons to be able to take the data with them, or limiting access to relevant folders or programs on the computer. You could also have a local network, or a series of computers that are only connected to one another and not the Internet, which is similar but allows for patrons to access your data from several computers.

You may not have any security issues and want the general public to be able to freely access your collection, as well. In this case, you'll likely want to load the data to a server computer that can be connected to the Internet, or load it to a web-hosting service. This will involve creating a website for your archive, which you likely already have.

You may also want only authorized users to access your data—for example, your archive is part of a larger library system, and you only want registered patrons of the library to access the data. This is also possible, but is a little more complex to set up. It is, of course, definitely possible to do (otherwise, it would be impossible to have security on the Internet for things like making purchases from online shops, and it would be possible for anyone to access your e-mail), and may be a feature of a cloud computing service should you decide to use one. Cloud computing services were described in detail in chapter 11.

Another thing you should keep in mind is what you want to do if you want patrons to have access to your data, but only as a physical object. For example, suppose that you have a collection of audio interviews and decided that it was best to put it on an external solid-state drive. A patron could request to use the drive, treating the data item more like you would a tangible item in your library. However, solid-state drives are easily altered. The patron could erase your data or write in their own, whether accidentally or maliciously.

It's possible to make files “read-only,” which means that the file cannot be altered. The file can be deleted, however, or moved to different folders and copied, so this isn't a perfect solution to the problem. You can change the permissions on files, which limits

access and alterations by unauthorized users (your patrons, in this instance). This is part of the security options for a file. You need to be sure that patrons who know something about computers aren't able to simply go and switch the options back. This means that you require more than one account on a computer—one for you and your staff and one for patrons. This will allow you to use your account to deny some things to anyone on the other account. The account for your archive and staff needs to be password protected in order to limit access like this.

Key Points

- Metadata is information about a digital item and is a necessary part of creating and maintaining a digital collection.
- Adding metadata and cataloging are similar processes with similar goals, and someone trained in cataloging has the skills needed to create good metadata. However, there are ways for amateurs to create useful metadata and even for computers to automatically generate some metadata.
- Metadata can describe what an item is, who created it, its structure, and more.
- In modern times, a number of metadata schemas are used that are similar to one another in many ways but are sometimes tailored to achieve a specific goal or are designed to be used for a certain type of collection.
- Determining how to access a digital collection is an important concern. Security of the collection should be considered carefully.

Digital collections work differently from paper collections in many ways, and the legal aspects of creating and maintaining a digital collection are not always readily evident or intuitive. In the following chapter, you will be given an overview of copyright law, both in general and as it pertains to digital materials specifically.

References

- Ashenfelder, Michael. 2013. "Personal Digital Archiving: Adding Descriptions to Digital Photos." Public Libraries Online. <http://publiclibrariesonline.org/2013/09/personal-digital-archiving-adding-descriptions-to-digital-photos/>.
- Library of Congress. 2009. "What Is a MARC Record, and Why Is It Important?" <http://www.loc.gov/marc/umb/um01to06.html>.
- Library of Congress. 2017. *The Card Catalog*. San Francisco: Chronicle Books.
- National Information Standards Organization (NISO). 2017. *Understanding Metadata*. https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.
- PREMIS Editorial Committee. 2015. *PREMIS Data Dictionary for Preservation Metadata*. Version 3.0. Library of Congress. <https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>.
- Taylor, Arlene G. 2004. *The Organization of Information*. 2nd ed. Westport, CT: Libraries Unlimited.
- Tennant, Roy. 2002. "MARC Must Die." *Library Journal*. <http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/>.



Copyright Law

IN THIS CHAPTER

- ▷ What types of materials are under copyright?
- ▷ Which materials are no longer under copyright, or can't be copyrighted?
- ▷ How is copyright law for digital materials different from laws for tangible materials?
- ▷ What is "fair use," and how does it affect an archive?
- ▷ What are the alternatives to copyright law?

In the past, written material of any kind needed to be copied by hand. This was a long, difficult, and laborious process; the idea that there might be only one copy of a book or document in the world was not at all unheard of. Books were exceedingly valuable because of this, and a person's most valuable possessions (monetarily, if nothing else) could be books.

The same was true for art; it was quite possible for only one copy of a work to exist. Photography had not been invented to make it possible to reproduce a work. The general attitude toward copying a work was different, too. During the Renaissance, copying a masterwork was considered a way of learning, and owning *copies* of a masterpiece was something to be proud of (Sotheby's, 2018).

This norm rapidly changed with the ability to mass-produce works of art and writing; for example, prints of woodcuts or etchings were a way to mass-produce art, as opposed to having a single work of art, such as a painting. To compensate for this drastic change in society, new laws were created. These laws were the beginnings of what is now copyright law.

In modern times, we tend to think of copyright law as a way to protect creators of works from having those works copied for profit, but this was not really the original intent. Though copyright laws did protect the creators somewhat, the original purpose of copyright law was to prevent commoners from gaining materials that might incite either

political or religious rebellion, thus upsetting the church or monarchy's power. It wasn't until the early eighteenth century that copyright began taking on the sentiments of modern copyright law (Hoffmann, 2001).

For an archive, many things that you would want to preserve are or were under copyright protection. Copyright issues are a common problem for all libraries, but digital archiving presents some unique challenges. Part of your archiving process will most likely involve determining whether or not you actually have the right to own or create a digital copy of information and, if you do, whether or not you can offer it to your patrons or make it available online.

It is the aim of this chapter to refresh your memory on the current copyright laws regarding tangible materials, then to discuss how these laws are similar to those for digital materials. Digital materials are easily copied and shared, which is a good thing in many ways, but also makes it very easy to violate copyright law. Therefore, more laws were created that are specifically designed to help copyright holders protect their digital materials. This will also be discussed. You'll need to understand your archive's rights as well as the rights of copyright holders when it comes to dealing with both types of items, digital as well as tangible.

This chapter can't cover every aspect of copyright law that you would ever need to know. It's always a good idea to learn more if you have doubts about an item that you are archiving. The laws also keep changing, so you'll need to keep updating your knowledge to ensure that your archive is working within the law. If you have any doubts about the legality of your project and what you want to digitize or store, consult an attorney or other legal counsel.

Basics of Copyright

So, what exactly is copyright law? It's important enough that it's part of the Constitution—Article I, Section 8 gives Congress the power to promote the arts and sciences by giving creators the exclusive right to their works or, in other words, copyright protection. Several branches of government are involved in creating and overseeing copyright laws. Congress can enact new copyright laws, while the Copyright Office of the Library of Congress is in charge of the administrative aspects of copyright laws, and the federal courts interpret these laws and enforce them.

Copyright law is important. In essence, it gives people the right to make a profit from their own creations. America would be a very different place if anyone could use the works of any other person for any purpose. For instance, the Disney company is well known for animated movies for children. If their movies were not protected by copyright, then anyone could copy their movies and sell them, which would make creating animated movies unprofitable. It would also make it very difficult to control the use of their characters, such as Mickey Mouse, or to control the company image. Without copyright, there would be no monetary reason to make anything creative in nature, because you couldn't make a significant profit from it. The protection that is offered by copyright law motivates people to create intellectual properties, like stories or song lyrics—motivation that would be lost in many ways without this protection. America would be without many works of art, writing, music, inventions, and other works that were created in the hope of making a profit.

Copyright law is also designed to protect the general public by *not* giving the creators of copyrighted works absolute control over those works. There are exceptions to copyright law—instances in which people who do not hold the copyright to a work are permitted to use the work. The concepts of *fair use* and the *public domain* are part of this. These concepts will be explained in more detail later in the chapter.

An archive's aim is to preserve the past and sometimes the present for the sake of the future. Trying to adhere to copyright law can therefore seem to be a burden, or even irritating, since archival preservation is done with good intentions. However, it's essential to keep within the rules of copyright law for several reasons. The most practical of these, of course, is to avoid a lawsuit from an unhappy copyright holder. Violating copyright law can carry some steep fines on top of legal fees, and no one wants to have to deal with this when you could be using your time and resources to do more archiving. The penalties for violation can range quite a bit, based on whether the infringement was done unintentionally (a minimum of a \$200 fine) or willfully (up to \$150,000), whether a profit was made from the infringement, and so on. For instance, if someone created prints of another person's artwork with the intention of selling them without the artist's permission, this would be an intentional violation of copyright. A person winning such a suit is also entitled to recovering attorneys' fees. There can be other problems or penalties arising from a violation of copyright law, whether it was intentional or unintentional or whether it was for a benign or malicious purpose (Library of Congress, n.d.).

But the other important reason for copyright law is to keep protecting the people who make things worth preserving for the future. With the Internet, the transmission of information is extremely easy, and you may wish to put your digitized collection online to help people all over the world who want information. If you put something online that you do not have the rights to or that someone doesn't want put online, then this harms the person who created the item—most likely by making it difficult or impossible to profit from their work, or making it possible for less scrupulous people to steal and profit from the work. This also discourages people from creating and sharing things. It's essential that, in the rush to preserve things for the future, people are not harmed in the present. Even a project that may seem noble, like sharing information with the world, is not noble if someone gets hurt by it in the process.

Copyright law is simultaneously simple and complex. You're probably already familiar with copyright law, but to understand the additions to copyright law that address digital materials in particular, it's helpful to remember the basics of copyright law in general, without the complications that digital materials add to it.

Works Protected by Copyright

While the types of things that can be copyrighted vary greatly, most copyrightable works can be sorted into a few basic categories.

- Music is copyrighted, but it must be recorded in a tangible format, like a tape or a CD, in order to be copyrighted.
- Movies, television shows, and other audiovisual works are copyrighted. A DVD of a movie, for instance, is a copyrighted work. As with music, these performances must be recorded.
- Performed works—such as dances, public performances, and pantomimes—are under copyright protection.

- Written works—such as books, news articles, poems, sheet music, and plays—are copyrighted.
- Artwork—such as paintings, sculptures, and photographs—are copyrighted.
- Computer programs are copyrighted. This may be of interest to you if you want to preserve software: the software that you preserve may be under copyright law.
- Architectural works—such as building plans—are under copyright protection.
- Derivative works or compilations of works are under copyright protection.

In general, works must be “fixed” in some manner to be under copyright, and thinking of works in this manner is a good way to help you figure out if something is potentially copyrightable. You cannot copyright a mere idea. For instance, suppose that you came up with an idea for a fictional story. Unless you write the story down, it’s not copyrighted, and if you were to tell your idea to someone else and they actually did write the story, you would not be able to claim a copyright violation. If you were to perform an original piece of music for an audience, but it was never recorded and you never wrote down how to play it, and someone else did, you would also have difficulty claiming a copyright violation.

Works are under copyright upon creation. There is no need to register a work for it to be copyrighted; this is simply helpful for identifying the owner of a work and is useful should there be any issues regarding rights and ownership. When a work is registered, the creator can easily prove that they hold the copyright should there ever be a legal dispute or a violation. A work doesn’t have to have any kind of notice for it to be under copyright, either; this is essentially assumed. While it was needed in the past, a copyrighted work doesn’t have to have the copyright symbol, which is the symbol ©; this has not been needed since March 1, 1989. Anyone owning a copyright on a work has six exclusive rights in regard to the work:

- The owner can reproduce the work. As an example, suppose an artist creates a painting. The artist has the exclusive right to create prints of the painting and sell them.
- If the owner of an audio recording holds the copyright to the recording, they can publicly play it.
- The owner can publicly display a work, such as a sculpture or a photograph. This particular right can get tricky when it comes to digital copyright law and will be discussed further later in the chapter.
- The owner can create derivative works, or works that are based, in some way, on the original work.
- The owner can publicly perform a work. This applies to someone who writes and performs music, for example. It would also apply to a choreographer, and others.
- The owner has the right to distribute the work. For instance, if someone wrote a book, they have the right to distribute and sell copies of that book.

But, it is possible for someone *other* than the original creator to hold the copyright for a work. The creator of a work can sell their copyright to someone else, giving that person all the rights regarding the work. Creative works can also be commissioned with the intent that the copyright will belong to the person who requested the commission. For example, suppose that a company wanted a new logo. They hire a graphic designer to create it for them. While the graphic designer created the logo, the company that hired the graphic

designer owns the copyright on the logo. This is considered a “work-for-hire” situation in copyright law (Butler, 2011).

In general, the rights that a copyright holder has are designed to help them profit from the work. Determining what works are subject to copyright and figuring out what rights a copyright holder has are therefore not too difficult: works that are original and creative in nature and have been recorded in some manner are usually subject to copyright law, and anything that interferes with a copyright holder’s ability to profit from the work is probably a violation of their rights. However, the protections offered by copyright are not indefinite. Because the laws regarding how long copyright lasts keep changing, figuring out whether something is or is not still under copyright is challenging.

Works No Longer under Copyright Protection

Copyright doesn’t last forever. Works eventually fall into what is known as the *public domain*. What this means is that, after a certain period of time, a work that has been created becomes free for anyone to use for any purpose whatsoever. For instance, there is a parody book titled *Pride and Prejudice and Zombies* by Seth Grahame-Smith, which is simply the book *Pride and Prejudice* by Jane Austen, but Grahame-Smith has altered the text to include zombies, largely for the purpose of comedy. While there are special rules regarding parody and copyright, using large amounts of text straight from the original is possible and lawful owing to the fact that the book *Pride and Prejudice* is no longer under copyright. It should be noted, however, that *Pride and Prejudice* was written in England and laws in other countries are often different from those in the United States.

Essentially, anyone can do whatever they like with works that have fallen into the public domain, including profit from publication. However, *Pride and Prejudice and Zombies* itself *is* under copyright. Remember, derivative works are protected by copyright law.

Knowing how long copyright law is in effect is not simple. The laws keep changing to extend the length of copyright. Earlier, Mickey Mouse was used in an example of what might be a consequence of not having copyright laws. In order to prevent early Mickey Mouse movies from falling into the public domain, the Disney company is a major force behind extending the time that copyright is in effect for a work.

Because the laws keep changing to extend the length of copyright ownership, whether or not a work has fallen into the public domain is becoming increasingly difficult to determine. In addition, laws vary depending upon what type of work was copyrighted and whether or not a work was published and when. These are some *general* guidelines regarding U.S. copyright law:

- Any work published in the United States before 1925 is in the public domain.
- If a work was published between 1925 and 1963, and there is a copyright notice attached, the copyright on such a work at the time could be renewed for another 67 years. If the copyright *wasn’t renewed*, it’s in the public domain.
- If a work was published between 1925 and 1963 and it has no copyright notice, it’s in the public domain.
- If a work was published between 1964 and 1977 and has a copyright notice attached, then the copyright was automatically renewed on the work for a total of 95 years of copyright protection.
- If a work was published on or after January 1, 1978, then it’s under copyright for the life of the owner plus 50 years. If more than one person was involved in cre-

ation or holds ownership, then the lifetime of the longest-living holder is used to determine how long copyright is in effect. January 1, 1978, is the date on which the 1976 Copyright Act came into effect, which is why the rules change after this date.

- If a work was published on or after January 1, 1978, and has a corporate author or is a work for hire (as in the example with the company hiring a graphic designer for a logo), then the work is under copyright for either 120 years after creation or 95 years after publication, whichever one is shortest.

The minimum date of 1925 is going to change over time and is the date before which copyright is no longer in effect as of the publication of this book.

Again, the length of the effect of copyright law keeps changing. Current copyright laws apply to works for the life of the creator plus 75 years after their death. This is an extension from the previous life plus 50 years. In 1998, the Sonny Bono Copyright Term Extension Act was passed in an effort to preserve the rights of musicians during the rock and roll era who died young, so that the copyrights would not run out while their music was still popular (Wherry, 2002). While these musicians can no longer benefit from the work, their families benefit from this law and the resulting income generated by copyrighted works.

Owing to the Sonny Bono Copyright Term Extension Act, the year 2019 was the first year since 1998 in which copyrighted works fell into the public domain.

If a work is anonymous, then copyright law becomes a little more difficult to follow. The work *is* still under copyright. It must be emphasized that because a creator cannot be found or is not associated with a work, that does *not* mean that it isn't under copyright. If a work is anonymous, is a corporate work, or is a work created by a person for a company (work for hire, as mentioned earlier), then the copyright lasts 95 years from the date the work was published or 120 years after the date it was created, whichever span is shorter. Remember, a work does *not* need to be registered to be copyrighted. Again, this is an increase in time, also due to the Sonny Bono Copyright Term Extension Act (Hoffman, 2001).

Determining whether something has fallen into the public domain is not easy, and neither is determining whether a copyright holder renewed their copyright. If you need to find out whether a copyright was renewed, there are basically two ways to do it: The U.S. Copyright Office will find out for you, but for a rather steep fee. You can also browse their online records and try to find out for yourself if an item is still under copyright. There are a couple of other online databases that can help you search for copyrights, as well.

You may wonder what happens if you can't find the copyright holder to determine if you are free to archive something. If finding out who holds the copyright or if copyright is still in effect is difficult or even impossible, then the item in question is known as an *orphan work*. Orphan works are problematic, since you may want to use an orphan work in your archive, but it may still be under copyright.

You could potentially search for an owner and then determine that you've given it a sufficient amount of effort and use the work anyway, believing that it is no longer under copyright protection or that the owner has abandoned rights to the work, but this has problems, too. For example, a group of research libraries and universities known as HaathiTrust digitizes materials and has more than 17 million works on its servers. Some of these are orphan works. To determine if a work has been orphaned, this group puts lists online of works that they want to digitize that they believe are orphaned, and then they

wait for a certain amount of time for a copyright holder to claim ownership and prove that it is not orphaned. This is an easy method of addressing the issue, but the group was sued for this practice in 2011 by three authors' groups and eight individual authors, who were concerned about copyright infringement issues and the security of the digitized files (Bosman, 2011). In this case, the court eventually ruled in favor of HathiTrust for a few reasons—in part because the judge decided that the project constituted a “transformative use” of the works (users can search for terms or phrases, but actual access to the works is restricted). Another major reason for ruling in favor of HathiTrust was the fact that the authors were not able to show that they were harmed financially (Ax, 2014). This precedent may be relevant to you as you plan your archiving projects.

While what you do in the event of an orphaned work is up to you and your organization, in order to avoid problems like this, it's best to be safe and ensure that you do, in fact, have the right to digitally archive all the items in your collection.

Works Not Copyrighted

So far, this chapter has discussed what types of tangible works can be copyrighted, what rights a copyright holder has, and how to find out if something has fallen into the public domain, making it free for you to use. Copyright does not apply to all works, and so there are many items that cannot be copyrighted at all. Any item that falls into this category is free for you to digitize, regardless of when it was made and by whom. The following are some of the kinds of works or creations that can't be copyrighted:

- Slogans, titles, names, and other short phrases cannot be copyrighted. It should be noted, though, that things like company slogans can be protected by other laws, such as trademark laws (Wherry, 2002).
- Lists of ingredients, processes, and methods can't be copyrighted. For instance, you can't copyright the list of ingredients in a recipe for chocolate chip cookies.
- Phone books aren't copyrighted. This is an item that might interest genealogists in particular.
- Ideas, as discussed earlier, can't be copyrighted. They must have some tangible format, such as a book or a music CD.
- Any facts or news can't be copyrighted. This includes the kind of data you'd find in an almanac or an encyclopedia, though the text itself could be copyrighted.
- Familiar symbols, like a flag, can't be copyrighted.
- Common-property works can't be copyrighted; this would cover things like height or weight charts. Calendars can't be copyrighted, either, though artwork on the calendar might be.
- Any work that falls into the public domain upon creation is not under copyright. For example, almost anything created by United States government employees automatically falls into the public domain if it was created as part of their job.

It's possible for a person to voluntarily give up the rights to their copyright, in which case, a work is free for anyone to use. There are also alternative licenses to copyright that a person may use to make their work more freely available for others to use; this will be discussed later in the chapter.

Once again, works that aren't fixed in a tangible medium are generally considered not under copyright. A photo or a book would be an example of a tangible medium. If

someone gave a speech and it wasn't recorded in any manner, though, the speech wouldn't be under copyright.

In the past, this distinction—whether or not something was recorded in a tangible way—made it pretty clear whether or not something was under copyright in most instances. If someone paints a picture, they hold the copyright. Their idea—the image—is fixed in a tangible medium: the painted image. But in modern times, this distinction is not so clear, and the general rule of thumb leads to some obvious complications. For example, is a website tangible? It's ephemeral, in some ways, and can change radically and without notice. It is composed of electronic signals. Are websites tangible, or aren't they?

Problems with Digital Copyright

It should be obvious that yes, a person who creates a website controls the copyright of the site and all the data on it, assuming that they are not violating copyright themselves. The website is a creative work, and there would be chaos if website designers and people who post their works online had no protection; it is also possible that there would instead be very little to see online without copyright protections, since creative people would not want to share their works.

Each time something new is invented that makes it easier to share information or to profit from someone else's work, copyright laws need to change in order to protect people who create intellectual works. The problems that are arising from digital information today are not new. For example, in 1909, copyright law needed to adapt to protect composers and publishers, since they feared that the new invention at the time, the piano roll, would represent a serious cut into their profits. The phonograph was a similar cause for worry, and so the laws were again changed to protect people's rights (Wherry, 2002). Photocopiers, VCRs, and many other inventions that make the distribution of information faster, easier, and cheaper all required the laws to adapt to new technology.

The Internet and the rise of digital information is really no different, except for the fact that some of the things that have been used to make judgment calls about whether copyright law has been violated with tangible materials don't transfer well to digital materials. For example, you learned in chapter 11 that when you "visit" a web page, what actually happens is a server sends you a copy of the necessary data to re-create the web page via a browser program, and all of that data is temporarily stored on your computer. Technically, you've just made a copy of copyrighted material. Is that a violation, or isn't it? If you were to put a link on your web page to an article on someone else's page, does it count as distribution? Since translations are normally considered derivative works, are you violating a creator's exclusive right to derivative works if you use an online language translator (Hoffman, 2001)? What happens on wiki pages, in which the authors of the content may be numerous and anonymous?

Things like links and making copies of digital items are common online and are necessary for web function. However, linking doesn't have a clear tangible equivalent, and making copies is clearly a copyright violation for tangible materials. Current laws are designed to help address this kind of confusion and both to protect people whose rights can be violated by the easy transmission of information and to protect from an erroneous lawsuit people who have no ill intentions.

If you are digitizing tangible materials, the laws regarding use of tangible materials will be somewhat more important to you. However, you may be interested in archiving born-digital materials. In this case, you'll need to know about the laws pertaining specifically to born-digital media.

Digital Materials under Copyright

So far, this chapter has discussed what kinds of tangible materials are under copyright. But what kinds of *digital* materials are under copyright? Some digital materials have a “tangible” equivalent, so sometimes determining what digital materials are under copyright is somewhat intuitive. The following are some digital items that are under copyright protection:

- Blogs, vlogs or video blogs, and podcasts (just as books, video recordings, and audio recordings are)
- Web pages and wikis, or sites that allow for user collaboration, such as Wikipedia
- Images in a digital format (just as regular photos, paintings, or other images are): This is important, considering that many artists are able to create works of art using computer software; art created in this way will have no original, tangible medium. Digital cameras also create images with no original, tangible format.
- Digital movies or other videos (just as film videos or movies are): again, these may not have an original format that is tangible in the way film has a tangible format.
- Digital advertisements (just as printed ads are)
- Software programs

Though most of these items are intuitive, something that is copyrighted that you may not know is protected under copyright law is e-mail. If you use e-mail to send messages, every time you create a new e-mail, it's under copyright. If you reply to an e-mail and it contains the contents of the original e-mail, you're technically in violation of copyright law by using the original sender's material without permission. This is not typically something that people go to court over, however.

Something that could be problematic, though, is an attachment to an e-mail, or additional files sent with the e-mail (Butler, 2011). For example, suppose a friend of yours sends you an image of a kitten. For the sake of simplicity, say that it's a picture that they took of their kitten. You now have a copy of that image, but you don't necessarily have the right to use that image. Forwarding it to someone else who appreciates cute kittens could technically be a violation of copyright law. If the picture that your friend sent was taken from somewhere online and doesn't belong to them at all, then things start becoming quite murky.

It's often easy for people who create born-digital materials to specify certain conditions under which their materials can be used, as well, which makes following the law even more difficult. This is a situation that is common with software in particular. When you install software, typically, you must agree to a license before the software will install. Oftentimes, people skip reading the license because it's very long and technical. However, if you want to archive materials like software, then reading the license is something you need to do. Many times, software companies limit your rights, but companies also often

allow for “backups” to be made of their software, which would allow your archive to make extra copies for the purpose of archiving.

It’s also possible for the creators of born-digital works to specifically give up some or all of their rights to a work, just as they can with tangible materials, which allows you to freely archive an item, just as with tangible materials. Laws regarding digital works in the public domain are similar to those for tangible materials.

Digital Materials in the Public Domain

It may have occurred to you that most things that are born-digital won’t have lost their copyright and fallen into the public domain yet, since the time needed for the copyright to expire won’t have passed. There *are* digital items that are in the public domain, though, since, as mentioned earlier, copyright holders can give up their rights. However, figuring out which is which is just as difficult, if not more so, as determining what is and is not under copyright with tangible materials. There are a lot of misconceptions when it comes to materials found online in particular, which should be cleared up.

Misconceptions about Online Copyrights

The following are myths about online materials:

- All works that can be viewed online are in the public domain.
- All works that can be viewed online are free to use.
- Digital materials without a copyright notice or attributable creator are not under copyright.
- If a copyright holder doesn’t respond to a request to use an item, then the item is free to use.

A lot of people treat everything online as being in the public domain, and a lot of people feel that materials offered via the Internet should be free to use—all the while, ignoring copyright law. That is more of a philosophy than a reality.

Some people assume that because an item is freely available to use or view online, this means that it can be used for any purpose. For example, suppose that a blogger, or someone who writes for a blog, wants a picture of a sunset to illustrate a post. This blogger does a search online to look for a picture that fits the bill and finds one. The blogger then makes a copy and posts it on their blog.

This image *may* be freely available to use, since there are many sites online that offer images that anyone can use for any purpose (further muddying the distinction between what is and is not under copyright with digital materials), but it’s more likely that the image is under copyright and the blogger does not have the right to use it, even if they *really* want to. Just as with tangible materials, digital materials don’t have to have a notice of copyright to be under copyright protection. Creators do not have to register with the copyright office, either. Digital materials are under copyright upon creation, just as tangible ones are, including things like computer software or websites.

Oftentimes, on social media sites like Twitter, users will post photos that they have taken that are relevant to current events, and reporters will ask the poster for permission to use the photos for their news reports.

Orphan works are extremely common online, and all the same rules apply to digital orphan works as tangible ones. Even if the copyright holder cannot be located, this does not mean that their work is available to use. Sometimes, it's possible to contact the copyright holder, for instance, via e-mail. If the copyright holder doesn't respond, however, this also does not mean that the item can be freely used.

The ease with which people can create and post online material creates some problems for you. While you will most likely find the copyright information for a professionally published book with the copyright office, you're less likely to find legal information regarding the copyright of the material on a blog or other online resource (though you may find some general information at the bottom of a web page or on a website's "About" page).

It's important to keep within the law for many reasons. However, your archive will have some special rules that are designed to make things more fair; again, copyright law is not only for protecting copyright holders, but helping the general public, as well.

Laws for Archives

Seventy years after the death of a creator is a very long time for an item to be under copyright, and since the creator can't benefit from the work any longer, many libraries feel that such a long extension of copyright law has no benefit to anyone—not the creator, and not society in general (though certainly corporations benefit greatly from the extension). It can be argued that, because the creator can't benefit, extensions like this get away from the original purpose of copyright laws (Hoffman, 2001).

There's a provision in copyright law that is aimed at archives in particular to address this issue. This provision gives your archive, as well as libraries and nonprofit educational institutions, the right "to reproduce, distribute, display, or perform in facsimile or digital form a copy . . . for the purposes of preservation, scholarship, or research" (Hoffman, 2001) so long as one of the following provisions is met:

- The work is no longer "subject to normal commercial exploitation." This condition is not defined by the law (Hoffman, 2001).
- The work can't be obtained at what the archive considers a reasonable price.
- The copyright holder gives permission.

These conditions, while helpful, are unfortunately not defined very well. You may decide that a creator can't benefit from a work any longer, but they may have a different opinion on the matter (or their family). You still need to make a decision as to whether or not something falls into one of these categories.

In addition, often times archives have the right to make a digital copy of an item for the purpose of preservation only. However, there is a difference between preservation and dissemination. That is, if you make a copy and store it, but don't make it available to the public until the copyright has run out, this may be within your rights, according to the Digital Millennium Copyright Act.

The Digital Millennium Copyright Act

In 1998, Congress passed two bills that are essentially additions and amendments to the 1976 Copyright Act. One of these was the Sonny Bono Copyright Term Extension Act mentioned previously in this chapter. The other was the Digital Millennium Copyright Act, or DMCA, which is designed to address copyright issues that are specific to digital works and the Internet.

Many of the rules of the DMCA simply make it illegal to circumvent or ignore security that is already in place on digital items. For instance, there are many sites online that exist solely to distribute files like music, PDFs, games, or movies, most likely in the belief that online information should be freely available. Oftentimes, these are items that are actually under copyright. As an example, you might be able to go onto one of these sites and find a copy of a movie that was recently released on DVD and download it. This, of course, is illegal, and is known as *online piracy*.

To prevent their products from winding up on such a site, many companies embed software code into their products that protects them from piracy, or they require a password of some kind for their product to function. The DMCA includes a law that makes it illegal to do anything to circumvent this, or to remove the anti-piracy software or to copy something that is protected in this manner (Wherry, 2002).

While it has good intentions, as with many aspects of copyright law, there are many issues with the DMCA. The rules preventing circumventing software protection are intended to protect companies from competitors copying their product but are problematic for a variety of reasons. One of the most straightforward of these is that it's easy for a successful company to develop a monopoly by restricting, for instance, third party-created parts for a device. It's also problematic in that many essential devices function using proprietary software that might have issues, but third parties are not allowed to investigate; this includes many modern medical devices. In 2016, the Electronic Frontier Foundation, a nonprofit digital rights group, filed a suit against the U.S. government in regard to this part of the DMCA; the suit is still ongoing (Doctorow, 2016).

It's also illegal to do this for online websites. While much of the World Wide Web is open for anyone to browse, there's also a significant portion of it that is kept hidden from access for commercial purposes or for security reasons. It's illegal to try to get around any software or security that limits access to something via the Internet (Hoffman, 2001).

Similarly, it's illegal to attempt to make items available for the public to access that should otherwise be under password protection. For example, if an item was part of a digital course at a university and was limited to student access only via password protection, and you knew the password and made it freely available, this is in violation of the DMCA (Wherry, 2002).

Along with these restrictions, the DMCA also provides some extra rules that pertain to archives and other nonprofit institutions, giving these organizations extra rights, which is helpful to you. Under some circumstances, it's legal to circumvent software and web page protections. You may be able to copy DVDs or CDs, for instance, for the purpose of archiving. As mentioned earlier, in the terms of service of many kinds of software, there is a clause that states that it's legal to make a backup copy, and so it would be legal for your archive to store a backup copy of the software so long as you (1) owned an original copy and (2) did not distribute your backup copies.

It would likely not cause any problem if you were to, say, make backup copies of a copyrighted audio collection and put the backups into archival storage. If you were to make your copies available as part of your online collection, though, then this might be illegal. When trying to determine whether or not something is legal or illegal under copyright law, a good place to start is to consider whether or not what you want to do could interfere with a creator profiting from their work. If it could, then it's probably a violation of copyright law.

If you make your collection available online, it's essential that you are sure that you have the right to display *everything* in the collection, not just to protect copyright holders, but to protect your organization. There is a rule in the DMCA that is designed to protect the rights of copyright holders by removing material from the web that is being used without permission. Many people who use the web will, as an example, look for nice photos to illustrate profile pages, blog posts, or personal websites, without being aware of who actually owns the right to the photo. If the owner of the photo catches the person who is using it, they can make a complaint to whoever is hosting the website. The host is required by law to take down the offending web page and to notify the owner of the violation. The owner of the website must then either notify the host that the photo was, in fact, legal to use, or remove it from the site if it was not legal to use. The response must be sent to the owner of the photograph, who must then respond as to whether court action will be taken. If the owner doesn't send a notice, then the host needs to put the web page back online in between 10 and 14 days (Wherry, 2002).

This particular part of the DMCA is helpful to copyright holders and gives them quite a bit of power, which is highly beneficial to them, since it's very easy to "steal" material from others online and claim it as one's own. The problem with this law is that anyone could falsely accuse any site of using copyrighted material without permission (Wherry, 2002). What this means to you is that it's important to check and double check to ensure that you have the right to post any material that you might want to display online, since having a web page down for 10 to 14 days could wreak havoc with your archive's website. You can protect yourself by taking the proper steps to ensure that you have the right to use everything in your collection and that you can prove it, too.

As an additional consideration, you might suppose that if it turns out that you don't have permission to display something after all and someone complains, you can simply remove the offending item and this will resolve the problem. However, because it's so easy to copy and distribute materials, if you put something online that should not be there, other people can create copies of your files and put them elsewhere online, out of your control, which means that the damage is already done from the perspective of a copyright holder. Remember, there are sites devoted exclusively to distributing copyrighted material illegally. It's important to be careful, not just for your archive, but for anyone holding a copyright on an item.

Fair Use

Sometimes, you can use an item that has a copyright without asking permission or paying a fee. This is known as *fair use*. The government decided that, while it's important for people who create works to be able to profit from their work, it's also important for people to be able to learn and exchange ideas without worrying about violating copyright law, and so, they created the fair use rules. There are a few kinds of activities that fall under

fair use. Most of this will not pertain to your archive, but it's good to be aware of the rules of fair use anyway, if for no other reason than to know what you can't claim as being covered under fair use.

In copyright law (section 107 specifically), there are four factors that are used to determine if using a work falls under fair use.

1. What is the item being used for, and is it for nonprofit or educational purposes?
2. What kind of work is being used?
3. How much of the work is being used (for instance, a certain percentage of the total work)?
4. What is the effect of using this item on the item's market value, and so on? In other words, will using this item impact the profit that the copyright holder can make?

There are many situations in which using part of a copyrighted work can count as fair use and is, therefore, legal. For example, things like quoting a passage from a book or a few lines from a song for the purpose of criticism or critique counts as fair use (though the creator may think otherwise). Parodies also fall under fair use. Summaries of works count as fair use, and teachers and students are able to reproduce portions of works for the purpose of education (Library of Congress, 2012).

Something that may be of use to you is the fact that reproducing part of a damaged work that is owned by an organization such as a library can fall under fair use practices. If you're preserving a collection that is deteriorating and you can't get new copies of your items, then these laws may enable you to legally save your collection without the difficulty and expense of replacing it (Library of Congress, 2012).

The legal precedent for fair use and infringements of copyright law has been to determine whether the offending party infringed upon copyright deliberately or without determining whether it was legal, or if they attempted to use material under the rules of fair use.

As with all other aspects of copyright law upon the rise of computers in everyday life and the use of digital materials, new rules regarding fair use with digital copyright needed to be created, and in 1994, a group called the Conference on Fair Use attempted to create a fairly extensive set of guidelines that would address fair use with digital materials. Unfortunately, these guidelines are an agreement between the organizations that created the guidelines, and are not actual law. No major library organizations endorse these guidelines (remember, the Library of Congress is heavily involved in copyright law). However, if you're interested in whether using something digital or transmitting information online counts as fair use, then these guidelines can give you an idea of what you may do and what may not be legal (Wherry, 2002).

Other Types of Licenses

There are people who feel that copyright law is too restrictive or doesn't benefit society as a whole. For instance, there are people who feel that all software should be freely available to use. With this mindset, some people have put their works under alternative licenses that allow for others to freely use their works (often software) so long as certain conditions are met. Items under these licenses are typically available for archiving or even

making them available online, regardless of other laws or when they were created, because the creators have forfeited part or all of their rights under copyright law either on moral grounds or for the purpose of helping others.

Free Software Foundation

As mentioned previously, some people feel that all software should be free to use. The Free Software Foundation has therefore created its own license, the General Public License, or GPL. This license states that the user of software under this license has the ability to use it for any purpose, make changes, share the software with anyone, and share any changes that the user makes (Butler, 2011). This enables other programmers to use and change software without worrying that they are in violation of copyright law and sometimes encourages collaborative efforts on complex software.

Open-Source Software

Open-source software is software that is not only free, but may be modified by its users, something that is usually forbidden in commercial software. The term *open-source* refers to the fact that the code for the software is available to view and change. Coding for software programs is often kept hidden from the user to discourage others from modifying it or discovering how it works—in which case, other programmers could make similar programs and profit from them, cutting into the profits of the original programmer.

Some examples of open-source software include the art program GIMP, the web browser Firefox, and the operating system Linux.

If the coding source is open, then programmers can easily see how the code works and make their own changes. Sometimes people who create open-source software will request that others share their modifications in the interest of adding new and improved features to a program or fixing errors, or “bugs.”

Open-source software licenses require that the distribution of the software be free and freely available via the Internet. While the license can restrict some forms of modification, it must allow for derivatives and modifications. Open-source licenses must also not discriminate against a particular person or group and must also be technology neutral—that is, it doesn't work only on a single proprietary device (Butler, 2011). In most cases, open-source software can be archived.

Creative Commons

Creative Commons (<http://creativecommons.org/>) is a nonprofit organization that essentially provides alternatives to copyright. It has a similar sentiment and function to the GPL, but can apply to a wide range of works, not just software. Under a Creative Commons license, a person could create a work and allow for others to use and modify that work to suit their needs (Butler, 2011). There are several variations on this type of license, and so you need to read exactly what your rights are with a certain work, but, typically, these works are free to use so long as it is not for profit, the works are not modified, and the work is attributed to the original creator.

The Future of Copyright Law

Copyright law changes slowly, but it does change, and those changes can have major impacts on a variety of aspects of society. Remember, as explained at the beginning of this chapter, the rules of copyright law have a huge impact on many companies. As mentioned before, the Disney company is of particular note with respect to copyright law, as this company is heavily involved in lobbying for an extension of the duration of copyright to keep its movies under copyright protection.

A recent change in the law is the Orrin G. Hatch–Bob Goodlatte Music Modernization Act or simply the Music Modernization Act. Signed into law in 2018, this act has three parts that impact the music industry specifically. The first allows for companies like streaming services to be able to pay a set fee to copyright holders for use of their work rather than negotiating with each copyright holder individually. The second part addresses rights for older music: sound recordings created before 1972 were protected by state rights only. Now, they are protected by the same digital rights as sound recordings from after 1995, so digital music providers must now compensate those copyright holders. The third part involves changes that make it easier for people who create sound recordings to get their royalties (Music Modernization Act, 2018).

This is a major change in copyright law, and addresses some of the unique issues that creators of music face online. These changes facilitate the process of paying creators appropriately for their works, a topic that is complex and controversial with music streaming services.

However, minor changes happen pretty frequently, and these can be important for your archive, too. It's important to stay up-to-date on recent decisions and especially on any major changes like the Music Modernization Act. While this particular act is not very likely to have importance to your archive, always consider how changes might impact your archive, whether you need to change your policy to protect it or whether a change makes it possible to improve services for your patrons.

Key Points

- Copyright law has adapted to address the issues that arise from progress in technology, with the rise of the Internet radically changing how easily copyright law can be violated as well as what works can be considered under copyright.
- Changes in copyright law make it difficult to determine what works are under copyright and which are not, though there are guidelines that can be followed and online databases that can make the process easier.
- Copyright laws for digital materials are very similar to those for tangible materials, with the addition of specific rules designed to help the copyright holders of digital materials.
- The rights of an archive can vary depending upon what items are to be archived and how these items are to be used. Making copies of digital materials and archiving items for the sake of preservation is often treated differently from archiving for the purpose of distributing information.
- The rise of the Internet and of computers as tools for everyday life has also led to the invention of some alternatives to copyright as a license for use, since there

is a philosophy among some people that information distributed via the Internet should be free for everyone to use.

- Monitoring changes in copyright law and modifying an archiving program accordingly is an essential and ongoing task.

Digital materials present many obstacles to archiving, both legal obstacles and moral ones. In the following chapter, you will learn more about the issues presented by digital items—not problems arising from the costs of archiving, as discussed in previous chapters, or problems arising from user access, organizing, legal problems, or moral issues. The issues discussed in the next chapter will concern the problems arising from archiving digital media due to the fact that it *is* digital in nature.

References

- Ax, Joseph. 2014. “U.S. Appeals Court Rules against Authors in Book-Scanning Lawsuit.” Reuters. <https://www.reuters.com/article/us-books-scanning/u-s-appeals-court-rules-against-authors-in-book-scanning-lawsuit-idUSKBN0EL22020140610>.
- Bosman, Julie. 2011. “Lawsuit Seeks the Removal of a Digital Book Collection.” *New York Times*. http://www.nytimes.com/2011/09/13/business/media/authors-sue-to-remove-books-from-digital-archive.html?_r=0.
- Butler, Rebecca P. 2011. *Copyright for Teachers and Librarians in the 21st Century*. New York: Neal-Schuman Publishers.
- Doctorow, Cory. 2016. “America’s Broken Digital Copyright Law Is about to Be Challenged in Court.” *The Guardian*. <https://www.theguardian.com/technology/2016/jul/21/digital-millennium-copyright-act-eff-supreme-court>.
- Hoffman, Gretchen McCord. 2001. *Copyright in Cyberspace: Questions and Answers for Librarians*. New York: Neal-Schuman Publishers.
- Library of Congress, U.S. Copyright Office. n.d. “Chapter 5: Copyright Notice, Deposit, and Registration.” Accessed November 4, 2019. <https://www.copyright.gov/title17/92chap5.html>.
- Library of Congress, U.S. Copyright Office. 2016. “U.S. Copyright Office Fair Use Index.” <http://www.copyright.gov/fls/fl102.html>.
- Orrin G. Hatch–Bob Goodlatte Music Modernization Act, H.R. 1551, 115th Cong. (2017–2018) (Music Modernization Act). 2018. <https://www.congress.gov/bill/115th-congress/house-bill/1551>.
- Sotheby’s. 2018. “A Brief History of Old Master Copies.” <https://www.sothebys.com/en/articles/a-brief-history-of-old-master-copies>.
- Wherry, Timothy Lee. 2002. *The Librarian’s Guide to Intellectual Property in the Digital Age*. Chicago: American Library Association.



Problems with Digital Archiving

IN THIS CHAPTER

- ▷ Why is digital information less reliable than tangible information?
- ▷ What are the problems inherent with using the binary system of data encoding?
- ▷ Why is outdated technology so troublesome for data storage?
- ▷ Why are digital materials less reliable for archiving than tangible ones?
- ▷ Why are tangible materials not yet obsolete?

In the 1990s and early 2000s, using the World Wide Web was still a pretty new concept. After all, the technology had only been available at all since 1991, and so the web's possibilities were still being explored.

This led to interesting entrepreneurial ideas like the Million Dollar Homepage. Set up by a man named Alex Tew in 2005, the site sold a million pixels of advertising space for one dollar per pixel, a novel idea at the time. Over time, all one million pixels were sold to a wide variety of people and companies, who placed their names, logos, and so forth on the site as well as a links to their own websites. This website still exists, continuously advertising for all the people and organizations that paid for their ads back in 2005 (Dowling, 2019).

What does not exist are quite a few of the websites that were advertised on the site (Dowling, 2019). How websites work has been described in several chapters in this book. When you set up a website, you need to pay for two things: a domain name (your URL) and a host for your web pages and their content (unless you host your own site, which has its own expenses).

Domain names are controlled by an organization, the Internet Corporation for Assigned Names and Numbers (ICANN), and when you register a domain name, the company you use works with ICANN to ensure that the name is valid and not already taken.

What this means is that, if you have a website and do not renew your domain name and your contract with your hosting service, your website essentially vanishes. If there were links to that site elsewhere, they no longer function. If no one else took screenshots of your site or copied it in any way, it has disappeared. There is no way to know what it was like or what was on it. Much of the early web has disappeared in this way, and much of it continues to disappear in this way now.

If you use a link that no longer works, an error displays, “HTTP 404 Not Found.” Having a protocol in place for the event of a link no longer functioning is an essential aspect of the web. There are a variety of status codes for using the web, and the code “404” is not arbitrary. The number 4 is for errors made by users, and 04 is for requesting a URL that is nonexistent. “HTTP 404” is a default message, but many sites have custom messages, often humorous ones, for this particular error (Dunietz, 2019). As some examples, movie studio 20th Century Fox shows brief film clips with a funny message, and Emailcenter UK shows pictures of their developers, allowing the user to select which one to “fire” for the mistake.

The Internet and the web, though they’ve been accessible to the general public for a very short period of time (relatively speaking), have become integral to daily life. Rather than letters, it’s more common to send e-mails. Diaries can be read by the public in the form of blogs. News are shared less on paper and more in the form of pixels on a screen.

Considering that a lot of current history is being made online, the fact that it’s so easy for information to disappear when shared online is a serious problem. This issue is not limited to online information, either. The advantages of using digital information are numerous, and this book has really only scratched the surface of what is possible. For all the advantages, however, there are a variety of problems.

Many of the problems that you will face with digital preservation have been discussed in previous chapters. This chapter will discuss them more in-depth. This chapter is a bit different from the others in this book so far in that it doesn’t tell you specifically how to do anything for your archive. Instead, it is the aim of this chapter to discuss the limitations of computers and digital files so that you can be prepared to deal with these issues in a logical way that is suitable to your archive.

Issues with Digital Information

In general, there are four issues that you will encounter with a collection of digital information:

- Mutability, or the changeable nature of digital information
- Binary data encoding, or the nature of how computers encode their data
- Obsolescence, or the problems that arise when technology becomes outdated
- Data decay, or the tendency for digital data to degrade

Issue 1: Mutability

The word “mutable” is an adjective. It means that something is easily changed or is subject to change, like the weather, and so it’s a very good word to describe how digital information works. Data is mutable in nature, in many different ways.

The goal of an archive is to preserve information and to keep it safe. This can mean different things, but for the most part, an archive aims to make sure that all of its information is accessible and that it stays exactly as it was when it was first created. Trying to do this with something that is mutable in nature, then, presents some obvious difficulties.

People have found a lot of different ways of effectively storing information over millennia—by making symbols on things like clay or stone tablets, leaves, papyrus, or parchment. The method of sharing information that people are probably most familiar with, though, is the book, which is one of the more efficient ways of storing and displaying information. Books are also not particularly mutable in nature, which is good from a preservation standpoint.

Historians have a pretty good idea about the development of the book and of writing over time, and have information about this from countries all over the world. The concept of making symbols that represent words, numbers, or ideas is very old and has been reproduced in many different civilizations. The ability to record information, even something as mundane as how many cows someone has, has a drastic impact on all aspects of a society.

Since writing is so valuable to society, people are constantly working on ways to make writing easier to access and distribute. For example, as discussed in chapter 4, movable type revolutionized the way books were produced. Before movable type and the printing press were invented, all books had to be painstakingly handwritten, an enormously time-consuming task. This meant that books were available only to a few (in European societies, solely to the church and the very wealthy) and that there might only be a handful of copies of a work. An individual book might be the only one in existence, which is a terrible thing from the perspective of an archivist.

Mass-printed books using movable type made books available for the lower classes, and also meant that there might be many copies of a work. Information and the exchange of ideas became open to more and more people as the years went by, making the transmission of information and ideas far easier than before. Further innovations only made printing books easier, cheaper, and faster, making books more and more easily accessible.

Historians can mark the point in time at which movable type was created, since information about this historic event is still available, even hundreds of years later. The Gutenberg Bible was the first book printed this way, sometime in the 1450s. Though the first copies were printed centuries ago, copies of it still survive to today, and these copies are still readable.

The printing press produced a physical object that has historical value. A lot of events in history do, and that makes it easier to study the past, since historians have actual items that they can look at for their studies. A problem with tracing the history of events with computers, then, is that oftentimes, the revolutionary object is not physical in nature, and can be changed, as well.

The year 1989 marked another historic occasion in the transmission of information—work on creating the World Wide Web began, and in 1990, the first web page was designed, an item that may be considered by future generations to be as revolutionary to human society as the Gutenberg Bible. However, unlike the Gutenberg Bible, there are no copies of this pioneer web page. Nobody saved one. The original was overwritten by other files (Ward, 2013).

The World Wide Web (discussed in detail in chapter 11) was developed at CERN, the European Organization for Nuclear Research. When CERN began searching for early files and web pages decades after the web's creation, it was soon discovered that, while the first web page was created in 1990, they did not have any copies of early web

pages from before 1992 (Ward, 2013). If you would like to see what the earliest web page available is like, there is a copy at CERN's website: <http://info.cern.ch/hypertext/WWW/TheProject.html>.

When the World Wide Web was being developed, its creators wanted to promote their idea and created demonstration materials in order to show the benefits of their idea to others. This early web page is available because someone decided to keep a copy of those demonstration materials (Ward, 2013).

No one thought to save the first web page because no one really knew what a revolutionary thing the World Wide Web would be. Probably no one thought this about the Gutenberg Bible, either, beyond thinking about how much profit could be made from printing books this way. Unlike a computer file, though, printing a book leaves behind a physical, tangible object. You can't erase it out of existence with the ease that you can a collection of electronic impulses. You can burn it or do other things to destroy a book, of course, but you can't accidentally forget to save your book or suddenly not be able to access your book due to a hardware failure. Destruction of books is typically deliberately done. The fact that digital information is so easily prone to being lost in this way is one of the many ways in which it is mutable, but not the only one.

You learned a little bit in chapter 13 about how problematic it would be if patrons were allowed total access to your archive. What if they started saving their own files on your storage materials, or what if they started altering the data that you do have? The integrity of your digital materials would be compromised, and one of the major goals of an archive is to keep things exactly the same. While there are things that you can do and ways to save things in a more permanent, unchanging way, all of the information using the storage methods discussed in this book can be changed and is designed to be changed (with a handful of exceptions, such as CD-Rs). Most people appreciate this, since they can do things like get rid of old photos to make room for new ones, or delete an old report that they don't need anymore and make a new one. This is all a *good* thing—unless you don't *want* anything to change.

This makes your digital materials vulnerable in a way that tangible ones are not. If someone goes into an art museum and attempts to vandalize a work of art, for example, there are people trained to restore damaged artwork. The changes will be noticed, and people can attempt to restore the work to its original condition. If someone attempts to change the words in a book, the alterations will be completely obvious by the scratched out or omitted words.

Although the type on the printing press was movable in order to make it possible to reuse letters for many different books, once printed, the words are always in the same place on the page. They might fade, but they'll never be completely erased, change positions, or be replaced with other words. For digital information, this might not be true.

If someone gets into a word processing file and deletes a paragraph, no one might ever notice that something is wrong, since the program will simply reformat itself around the changes. Someone could use a photo-editing program and add all kinds of things to a digital image. If the person changing the photo is good at it, someone looking at the image might never know that something has been changed. After all, the point of sophisticated photo-editing programs is exactly this: so that you do not know that the altered image is not exactly like the original.

As another example, suppose that you don't want to save items that have a tangible form originally. You want to save web pages, for instance. But web pages change, too. Say that you're saving a web page that helps students find open-source texts for study on

the web. The website updates monthly, adding new texts. Recently, it underwent a layout redesign—the navigation buttons are now along the top, and the background is blue. Do you keep updating your copy? Do you save a complete copy of the website every time it updates? And if you're saving a great number of websites, then how do you know that something has changed in one of them? If you're trying to keep track of a forum or other media in which many people are able to contribute, what do you do if someone changes the text of a comment that they made? How can you preserve the integrity of history like this?

Digital information is easily changed, erased, and written over. It can even be done by accident, and not maliciously. This presents a problem for you as an archivist. If you put a photograph into a box, you expect that you'll have the same item when you retrieve it again (assuming that it isn't attacked by mildew or insects or something, of course). Digital materials are different. They are capable of changing. They can change even if no one interacts with the item, altering spontaneously. For example, in chapter 9, you learned that hard drives function best if they are used often, so that the data gets refreshed. In essence, digital materials and tangible ones tend to be opposites. Many digital materials are safer the more they are used, and tangible ones are more prone to damage the more they are used. For your archive, this means that you need to consider a few things with your archiving plan:

- How can you make sure that your digital materials can't be erased or written over?
- How can you ensure that your digital materials are true to the original and are unaltered?
- How will you deal with preserving materials that are inherently prone to change?

There are a variety of approaches that you can take to address the problems of the mutability of data, some of which will require more advanced computer skills. As an example of a useful feature, many Windows operating systems can compare files to determine if they are the same. Simply type FC followed by the file path to both files into DOS or Command Line. You can also run comparisons at a binary level, run a case-sensitive comparison, and more.

Finally, you need to think about digital materials in a different way from tangible ones. Archiving traditionally involves repairing or maintaining an object and storing it someplace safe, where it won't be harmed or touched unnecessarily. Though it depends upon how you store your data, you should probably think instead about how your data can be interacted with in order to keep it safe and to monitor it for degradation and change.

Issue 2: Binary Data Encoding

So far in this book, the concept of binary has been presented in a positive manner. Computers work efficiently with the binary system of numbers. It's a numerical system that works very well for computers and will probably continue to work well for some time. Though there are only two numbers to work with, you can store all kinds of information using the system. Binary is also a problem, though.

The biggest issue with digital media is that you require a tool to use it. And not just any tool, either—computers are capable of detecting fluctuations in a magnetic field, electrical impulses, or, in the case of optical media, changes in light that are far smaller and more minute than anything a human can sense—even with tools like magnets or

magnifying glasses to assist. This is also what makes computers so helpful—the fact that they are capable of storing so much information in a tiny amount of space.

Computers are, as mentioned at the beginning of this book, glorified calculators. Everything is numbers to them. As you learned in chapter 6, although programmers use coding, or sets of words that act as instructions to the computer, what actually happens is more complicated. Those words are all for the benefit of human programmers, working with human language. When a programmer has completed a program, that program needs to be “translated” into numbers so that the computer can understand the instructions. Programmers are working several levels away from what a computer can understand. While it is possible for a human to use machine code (the term for languages that don’t need translation and are directly accessible by the computer), machine code is extremely difficult and tedious to write. Very few people work this way when it comes to programming, or are even capable of it.

Humans don’t work well in binary. Not only is it removed from the comfortable decimal method of mathematics, but also, humans generally like words, not numbers. Even if it were possible for someone to, say, detect the dark and light spots on a burned CD without the aid of a computer, the vast majority of people would not be able to translate what was on the CD because it’s simply too difficult.

If every computer suddenly went down today, all of the digital data saved everywhere would be completely lost. It would simply not be worth the time and effort to get the information without a computer to do it. Everything that was written down in books or other materials, on the other hand, would be perfectly accessible. Books are tangible items that are large enough for people to see without any sort of aid. The characters or letters in a book represent concepts or sounds that are easily translated for most people, since humans are language-based creatures and easily connect a symbol with a word.

The scenario of every computer everywhere suddenly breaking all at once is pretty unlikely, but it serves a point. Your data is only useful if a computer can read it. So, apart from your major archival goal of digitizing tangible materials or saving born-digital materials, your primary goal must be to ensure that a computer somewhere is able to read your data. Without the calculator—that is, a computer—your sets of electronic numbers are completely useless, and all that hard work creating your archive would have been for nothing.

Although every computer in the world breaking simultaneously is quite unlikely, the scenario of having data that no computer in the world can read is not implausible at all.

Issue 3: Obsolescence

Though people tend to think of computers as being fairly modern, it’s really only the personal computer that’s new and innovative. There were plenty of computers similar to the modern concept of a computer available starting in the 1940s. They just weren’t as common as computers are today, and people did not have computers in their homes.

First, as explained in other chapters, people didn’t have computers in their homes for a couple of reasons. Early computers were incredibly slow compared to the capabilities of computers today, so they didn’t do as many things as modern computers can do; basically, they were used as calculators only. There was no World Wide Web, so you couldn’t do banking or shopping or search for information, and computers lacked the capability for things like games, movies, or music and didn’t have monitors capable of showing you the

relevant information for games or movies, anyway. All these activities simply wouldn't be practical. So, only large companies and universities would have wanted or needed one (or would have been able to afford it).

Second, computers required quite a bit of maintenance. You probably do very little to maintain your computers. Maybe you run some antivirus software or you might use some compressed air to blow dust out of the casing. But computers used between the 1940s and 1960s needed constant maintenance. Fuses and transistors had to be replaced on a regular basis—regular meaning daily or more often. Finally, and the one reason that may be most relevant to the discussion, the original computers were huge. Anyone owning a computer would need an entire room to hold it, not just a desk, since their weight was measured in tons.

Early computers had several ways of storing data, but one of the most prominent of these was through the use of punch cards. Punch cards were heavy pieces of paper with holes punched into them to encode the necessary numbers for the computer. A program could consist of hundreds and hundreds of these punch cards, which needed to be inserted into the computer in a specific order for the program to function.

Punch cards work nicely as a data storage method. They were even used before computers were created. Some punch cards were designed for mechanized looms, with the holes being used to designate fabric patterns. It's easy to tell when a punch card has gone bad, too, unlike modern methods of data storage. A bad punch card will be torn, folded, or otherwise mangled.

People keep innovating, however, and though modern storage methods are troublesome in some ways, they store a lot more data and are much more convenient than those punch cards, if for no other reason than they take up less space. The humble punch card was replaced by other storage methods, such as magnetic tape, that offered easier access and stored more data.

There are punch cards around today, and their data is still good. Unfortunately, though, all those cards are completely useless. The computers that used them aren't around anymore. It's simply not practical to keep a massive, energy-devouring machine for the sole purpose of reading programs that are decades out of date. And as mentioned earlier, data without a computer to interpret it is useless. Punch cards are so worthless today that people use them for art or craft projects, and there are suggestions available on the web for interesting things that can be done with these otherwise useless items.

If you really, really wanted to read a punch card, it's not impossible. It's possible to use a service, and there are a variety of sets of instructions online for building your own punch card reader. Interpreting what it says, on the other hand, is a different matter.

It's somewhat of a tragedy for archiving, though. There won't be any equivalent Gutenberg Bible for computer programs, really. Though there's information available about the development of programming over the years, it's not the same as a tangible object, and it's impossible to tell what people in the future will wish that the people of today had preserved. Will it be those early punch card programs?

In some ways, it's simply not practical to keep everything. You probably don't want your archive filled with boxes and boxes of cards that no one will be able to use and that take up precious space in your facility. But the phenomenon illustrates a point—technology becomes obsolete. This has been mentioned over and over throughout this book, but

it really bears repeating. Though punch cards were in use for decades, the speed at which technology becomes useless today is quite alarming.

Floppy disks, for instance, had been around for 20 or so years of common use by everyday computer users. Once recordable CDs and other recording methods that could store more data became available, the use of floppies rapidly declined, and they are rarely used today.

Twenty or so years probably sounds like a pretty long time, but it's really not. Twenty years is a fraction of the amount of time that you could expect a book to be useful. Remember, the Gutenberg Bible is more than 500 years old. There's writing available that is much, much older, as well. Consider the Dead Sea Scrolls, for instance, or the Rosetta Stone—both thousands of years old. The technology needed to use these is still available, since humans have not evolved anything better than eyeballs in the past few thousand years.

Even the coding languages used to create software programs can become obsolete or change over time. For example, HTML, which is used to create web pages, keeps updating. This book discusses HTML for the sake of ease, but the current standard is actually HTML5, the fifth major version of the language since its beginnings in the early 1990s. For this version of the language, some instructions are added to make it easier to create web pages, and some are removed, being categorized as obsolete. So, any web pages that you might want to archive might no longer work with future browsers. They won't display the way they should.

It's difficult to determine if a technology is becoming obsolete. Technically, punch cards have been in use since the 1700s, but they went largely out of use in the 1970s and are now totally useless. Magnetic tape, which was covered in chapter 8, has been around in varying forms for close to a hundred years and is still in use today (though those early forms are probably no longer readable by any device). Paper tape, a sort of combination between the principles of magnetic tape and punch cards, is even older than magnetic tape, but stopped being used in the 1990s and is now exactly as useful as a punch card.

As another consideration, oftentimes, someone will create a new technology that is an improvement in some way or is novel, but it won't catch on with general users. It may have a fatal flaw, or be too expensive or inconvenient to use. It's difficult to predict if something is merely a fad or it's a technology that will be around for decades to come.

For example, the Blu-ray disk works exactly like a CD or DVD, but has an enormous storage capacity in comparison to these other two technologies. Movies on Blu-ray are available for purchase and do pretty well, but they haven't completely replaced DVDs by a long shot. Why not? Probably for a variety of reasons—Blu-ray disks are more expensive and Blu-ray players less convenient, and when it comes to viewing quality with movies, the difference between a DVD and a Blu-ray disk may not be enough to entice people to buy (as opposed to the difference in convenience and quality between a VHS and a DVD). It may be that the ability to stream content online is a factor, as well. There isn't really a clear, tidy answer for it, so all you can really do is to monitor developing technologies and be sure that you don't get caught with something useless—and that you move your data before it's too late.

Part of the goal of digitizing materials is to save space. It would be pretty inconvenient to be storing stone tablets in your archive just because they have an excellent longevity and are generally fire- and water-resistant. But the fact that the technology changes so rapidly should influence how you think about your archiving project. Here are a few questions to consider:

- How will you keep an eye on developing technologies so that you can plan for the future of your collection?
- How will you determine whether a data storage method is becoming obsolete, and that your data needs to be moved to something new?
- How will you determine if a new storage technology is simply a fad or if it's a revolutionary innovation that will replace the current technologies?
- How will you decide what equipment is important to keep for the future?
- How will you migrate your collection from one technology to another in an efficient and logical manner?

Obsolescence is a definite problem and one that you will unquestionably need to address when you create a plan for your archive and determine which technologies you want to start with. However, there is an even bigger problem that digital materials face, which has also been touched on somewhat in the chapters so far.

Issue 4: Data Decay

On January 10, 1921, an infamous event occurred, which causes great frustration for genealogists and historians in the United States today. A fire broke out in the Commerce Department building in Washington, DC. The ultimate result of this is that most of the 1890 census was completely destroyed (National Archives, 2005). There is absolutely no way of getting that data back. It became irretrievable ash, and anyone who needs that data is simply out of luck.

It may be alarming to you to realize that vast amounts of data are just as completely destroyed on a daily basis. Possibly you know someone who has lost years and years' worth of photos (or maybe it's happened to you) due to a hard drive crash, or because the rewritable CDs that were supposed to be so revolutionary didn't hold up as well as their manufacturers said. Maybe you or someone you know was writing an important letter or a paper and suffered computer failure. Because they didn't save the file on their hard drive and it was being stored temporarily in the RAM, the material is gone. Or maybe you have tax records on floppy disks and you have no computer with a floppy drive, so now they sit in a drawer somewhere, slowly deteriorating.

Humans make mistakes and forget things. While some programs automatically save information periodically to help with this, it doesn't completely solve the problem, and computers are rather unforgiving. Even computer professionals make mistakes (remember, that first web page was overwritten with other files).

A book can easily be burned in a fire. Fire is also bad for digital materials, melting things like magnetic tape or CDs (some types of flash drives are designed to be fire-resistant, though). Other disasters are also bad for your archived materials, both tangible and digital, such as floods or earthquakes, although restoring a waterlogged book is a much easier prospect than dealing with a soaked hard drive, for instance.

Many cloud computing services offer redundancies and backups, but remember, a cloud computing service is not infallible. Employees can make mistakes and equipment can have problems, or a variety of other things can happen that can delete data.

Digital materials have an extra vulnerability, though. When you put a book away, you'll expect that when you take it out again, you'll have exactly the same book. None of the

words will have randomly disappeared. The illustrations will all be there in the same place. While the pages may become more brittle over time or the words more faded, they never completely vanish. You'll never open up a book and suddenly get a notice between the pages that you can't read it due to an error.

This is not the case with digital materials. You've already learned a little about bit rot and data decay. These phenomena refer to the fact that digital materials can change over time, with the data on them vanishing. Magnetic tapes can lose their magnetization, and so can hard drives. Flash drives can lose their charge, and writable CDs fade.

How fast this happens is open to debate. There have been many tests done on digital materials, but none of these storage methods have been around long enough to understand their possibilities and limitations the way we do when archiving something like a book. Some methods of data storage become obsolete before a test such as this can even happen. There are many tests and experiments that have been conducted on digital materials in an attempt to simulate an accelerated aging process. These experiments are quite useful and can help with predicting how to best preserve digital materials, but still cannot truly predict the future qualities and the lifespan of these items. Therefore, there are considerations that you must make in regard to the longevity of your project:

- How will you ensure that your digital materials are not deteriorating?
- How will you ensure that your digital materials are safe in the event of a disaster?
- How will you ensure that your archive is not ruined by a storage device going bad?

Digital materials are inherently difficult to archive. That makes your work as an archivist even more important. In modern times, people are embracing the advantages of digital materials. Digital images take up less physical space than paper ones. A music collection on a smartphone or computer is less cumbersome than a collection of CDs or tapes or records. The same goes for movies. Many people's thoughts and ideas are in the form of blogs—not newsletters or pamphlets or books. It's highly possible that the creations of today will be lost due to the mutable nature of digital materials, just as surely (and just as frustratingly) as that 1890 census.

What this means is that it's important to keep track of the creations of today. As an example, the Library of Congress has been saving "tweets" from the popular social media website Twitter (Gross, 2013). Twitter was briefly discussed in chapter 13; tweets are short messages to other users. They can be "tagged" with words to indicate the content of the tweet, and can be directed toward a particular person's account online. Many tweets are lacking in content, and simply link to another location online, or are responses to the tweets of others. Some are venomous in nature, or exist solely to promote something outside of Twitter. Some are completely indecipherable without context. However, tweets are a way that people in modern times communicate. Will they have any value or relevance hundreds of years from now? Maybe not, but they will exist for study if needed, since the Library of Congress has decided to preserve them.

This project originally started by archiving all tweets, but it has since been scaled back to only archive selected tweets. The current collection is estimated to have more than 170 billion Tweets. It is not available to the public for a variety of reasons; because it is such a huge collection, determining how to sort it and make it possible for the public to access is a problem (Wamsley, 2017).

Tweets are digital in nature and have no physical form. This presents an extra obstacle to archiving. However, if you are digitizing tangible objects, you'll have the advantages offered by both a tangible object and a digital one.

Tangible Materials and Longevity

If you plan on digitizing a collection, you should not eliminate the original physical copies of your items unless it's really necessary or you're attempting to preserve something that is deteriorating to a state beyond practical use. As mentioned before, tangible items have a much greater lifespan than digital ones. Not only is their *predicted* lifespan longer, but it's been *proven* to be longer through trial and error over thousands of years.

Many archiving projects that are designed to preserve digital materials turn to tangible items in order to truly preserve their information. For example, the Internet Archive is a nonprofit group that has many goals, including keeping copies of online web pages. However, they have another project that is more physical in nature: collecting a copy of every book ever published—ever. These books are not being digitized. They're being stored in a cool, dry warehouse inside large shipping containers. Along with books, the Internet Archive is also collecting music and movies to store (Stokes, 2011).

One of the issues faced by the Internet Archive is that of copyright. Remember, websites and materials on the web are copyrightable and archiving them can potentially be a violation of copyright law. If archiving online material, this is something that must be considered with your own projects.

E-books are cheaper and more convenient than paper ones. But the object of this vault is not to distribute information. Should digital versions of these items vanish, like so many digital items have, there's still a copy available because of this project. If someone thinks that a digital version has been tampered with, which is so easy to do, then there's a copy available that is not so easily changed and can be used for comparison (Stokes, 2011).

Miniature items, or tangible items that are extremely small, may be a technology to look into in the future. For example, the Rosetta Project is designed to archive the languages of the world. The creators of this project fear that a huge number of the languages in use today will simply vanish over the next century, and so they are attempting to document them now. The method by which these languages are being preserved is via an item that they call the Rosetta Disk. The disk is three inches in diameter and made from nickel, and is stored in a spherical casing made from stainless steel and glass. On the face of the disk are approximately 14,000 pages of information about languages. Rather than using binary data encoding, the pages are actually microscopic images that can be read by a person (using magnification, of course). This disk, should it be kept safely in the casing, could last for thousands of years (Rosetta Project, 2019). Even if every computer in the world were destroyed tomorrow and humans were never able to create computers as people think of them today again, the Rosetta Disk would still be serviceable and useful.

Oddly enough, it seems to be a trend that older technologies do better as far as archiving goes than new technologies, as though things are working backward. Consider the huge longevity of stone or clay tablets, for instance. The company Hitachi is working on a method of storing data that will be similar in principle, etching binary code onto pieces of quartz glass. It's a little like working with pressed optical disks, but the glass is

much more resistant to extreme temperatures and won't degrade over time (Clark, 2012). Scientists are even working on encoding data using the very oldest material on the planet: DNA (Richards, 2013).

The convenience offered by digital information and the rapid speed at which information can be created, shared, and stored are changing human society, and people can benefit greatly from digital technology. You will be benefiting others by making the information your archive holds part of this easily transmitted data. However, the point that must be made here is that, while technology is wonderful and helpful and a great way to make your collection available to patrons, it has limitations and problems that have not yet been fully addressed. You should never expect technology to solve all your problems, and you must always be on the watch for better, more stable methods of storing data.

Key Points

In this chapter, you learned a little about what kinds of questions you should be asking about technology and your collection in order to develop good policies.

- Digital media is not as stable as traditional, tangible materials. As such, the normal methods of archiving don't address the problems facing digital media very well.
- Your archive will need to develop policies that are specific to the difficulties of digital media, and that continue to monitor developments in technology to avoid having your archive become obsolete. The instability of digital media makes traditional media still relevant today.

In the next chapter, you will review the information covered in the book so far so that you can start developing a coherent plan for creating a digital archive.

References

- Clark, Liat. 2012. "Laser-Etched Quartz Will Store Data for Hundreds of Millions of Years." *Wired*. <http://www.wired.co.uk/news/archive/2012-09/25/hitachi-quartz-data-storage>.
- Dowling, Stephen. 2019. "Why There's So Little Left of the Early Internet." BBC. <https://www.bbc.com/future/article/20190401-why-theres-so-little-left-of-the-early-internet>.
- Dunietz, Jesse. 2019. "How the 404 Error Created the World Wide Web." *Popular Mechanics*. <https://www.popularmechanics.com/technology/a24091/404-error-world-wide-web/>.
- Gross, Doug. 2013. "Library of Congress Digs into 170 Billion Tweets." CNN. <http://www.cnn.com/2013/01/07/tech/social-media/library-congress-twitter/>.
- National Archives. 2005. "The 1890 Census." <http://www.archives.gov/research/census/1890/1890.html>.
- Richards, Sabrina. 2013. "DNA-Based Data Storage Here to Stay." *The Scientist*. <http://www.the-scientist.com/?articles.view/articleNo/34109/title/DNA-based-Data-Storage-Here-to-Stay/>.
- The Rosetta Project. n.d. "Concept." Accessed December 8, 2019. <http://rosetta-project.org/disk/concept/>.
- Stokes, Jon. 2011. "Internet Archive Starts Backing Up Digital Books . . . on Paper." *Wired*. <http://www.wired.com/business/2011/06/digital-books-on-paper/>.

- Wamsley, Laurel. 2017. "Library of Congress Will No Longer Archive Every Tweet." NPR. <https://www.npr.org/sections/thetwo-way/2017/12/26/573609499/library-of-congress-will-no-longer-archive-every-tweet>.
- Ward, Mark. 2013. "Online Appeal Unearths Historic Web Page." BBC News. <http://www.bbc.co.uk/news/technology-22652675>.



Drawing Up Policies

IN THIS CHAPTER

- ▷ Which parts of my collection should I store first?
- ▷ How should I avoid and address legal issues?
- ▷ What settings should I use for digitization and file storage?
- ▷ Which methods of data storage are optimal for my archive?
- ▷ What equipment will I need?
- ▷ What type of software will I need?
- ▷ How will I protect my collection?

As discussed in several chapters of this book, the idea of a library and an archive is very old. One of history's most famous libraries, the Library of Alexandria, existed more than 2,000 years ago, and some of the techniques used to store and locate information there are similar to those used today. Librarians, collectors, and scholars have had a lot of time to perfect the process of collecting and storing a wide variety of library materials.

In contrast, the concept of *digital preservation* has only been around since about 1990. Even taking into account that the preservation of digital materials has been a concern since at least the 1960s, creating a program for creating and archiving digital materials is a relatively new undertaking and there is still plenty to learn (Hirtle, 2008).

So far, if you've read every chapter in this book, you've learned quite a few things. You know a little about how computers work, what sorts of things computers can and cannot do, and what the past and the future of computing look like. You know how software files and programs work, how you might make information available to others, and the kinds of pitfalls you might encounter as you work on your project.

Now you're probably wondering how you can practically apply what you've learned toward your collection. This chapter will outline general steps to take when creating a digital archive, review important information from the previous chapters, and ask you questions to consider when creating plans for your own digital archiving project.

There is not a single correct way to go about making your archive, and so you may decide to take a different approach from the one outlined in this chapter. You will also certainly have questions that are not covered in this chapter. As you read, think about the questions provided by this chapter and any others that you may have.

Setting Goals

Though it may seem unimportant or obvious, consider what your goals are before you start. Why are you creating a digital archive? There are a lot of good reasons to make one.

- You want to make it easier for your patrons to obtain information, or make it easier for your archive to share your information with others.
- You want to be part of the effort to archive a digital culture.
- You want to work with other institutions that are digitizing their collections.
- You have a special collection that needs to be protected. This could include preserving something unique that may not be addressed by a national project, such as local history.
- You want to make information available to your patrons while protecting the original object—that is, patrons can use the digital version without ever handling the tangible object, which keeps the original object safer.

You might also have another motivation that doesn't fit into one of these categories. Write down your basic reasons for making an archive and what your goals are. Some of the benefits of creating a digital archive were discussed in chapter 1; considering how your project might benefit others can also be helpful for deciding what your goals are. Your goals might influence how you go about creating your collection and what you need to consider for your project to be a success. Writing these reasons down might also help if you need to talk about your project to groups or communities that might be willing to provide funding or other kinds of assistance to help your project reach its goals.

When you create a plan for your project, be as consistent as possible. Consistency is very important for digital archiving. Having a comprehensive, consistent plan will make it possible for other workers to carry on without you should you leave the archive and make it possible for your other workers to make decisions on their own without needing to consult you. A consistent plan will also make it easier for you to share your work with other archives if desired, or to collaborate with another organization.

After determining what your goals are, your next step should be to decide what you want to archive.

Your Collection

Regardless of what you want to store, the first thing that you must consider is the size and scope of your collection. If you have a small collection, if you only want to archive a few things, or if you want to archive a very specific portion of the collection, then you may be able to store everything and can therefore move on with your plans.

Otherwise, you're going to need to make some difficult decisions. The process of storing born-digital materials and the process of digitizing materials, then storing them, are both tedious and time-consuming for varying reasons. You need to prioritize which

parts of your collection need to be stored first. Chances are good that you won't be able to archive everything you would like, and so you may need to determine what is not worth storing. There are a couple of ways to go about deciding what parts of your collection are the most important to digitally archive. Here are some good questions to consider:

- Is the item particularly useful to your patrons, or would it probably see a lot of use if it were part of your digital archive? Does it already get a lot of use? If so, then this is a high-priority item.
- Is the item unique in any way? Do many other people have a copy, or is it unique to your archive? The former would decrease priority, and the latter would increase it. If the information is unique, then it's important to archive because there are few or no backup copies in other archives.
- Is the item highly delicate? Is the item in danger of degrading beyond reasonable use? If it's a born-digital item, is it in danger of becoming obsolete? Protecting items from degradation or obsolescence should be a major goal for a digital archive.
- Has anyone else digitized the item? This might be difficult to determine, and it's not necessarily a bad thing for there to be two different digitized copies of an item. If you definitely know that there is a digital copy of something already in existence, though, then the item probably drops on your priority list.
- Is the item too fragile for you to digitize? You may find that you have items that need to be given special treatment. For example, scanners use beams of light, which may harm some items. As another example, some early audio recordings were made using brown wax cylinders, which can become extremely delicate and require special handling to digitize.
- Is the original (not digital) item difficult to store? For instance, is it large, or does it require a special shelf or storage unit? Would digitizing the item remove difficulties in this respect or open up more physical space in the archive?
- Is the item difficult to retrieve? Would having a digital copy reduce this difficulty for your employees and your patrons?
- Is there something about the item that would require special equipment to digitize? You don't necessarily want to spend your funds on equipment that can only digitize a small part of your collection when you could spend the funds on something that could be used for a larger part of your collection.
- Is the item already catalogued, making cataloguing the digital version simpler?

Some questions to consider for born-digital items in particular are the following:

- Will this item potentially have historical or cultural significance? This is very difficult to determine, but you can attempt to decide which seem most useful by having protocols in place to make choices.
- Will the item be difficult to store? Will it be difficult to use and retrieve? This is a potential complication if you wish to store software that is designed to work with a specific operating system or piece of equipment.
- Will it be difficult to convert a file to a more archive-friendly format?

While you're pondering these questions—you'll probably think of more that are relevant to your particular situation—write down ones that seem significant to you. You can use these to officially determine which items in your collection are most important to digitize, or to help with discussions if you are making these decisions in a group.

Another important consideration to make is to think of what your patrons might like to see. If you have anything that is particularly fun or interesting, you might want to make it part of your collection to draw patrons to your archive. Even if an item would be otherwise a low-priority item according to these questions, an item that patrons will like is probably an important one to digitize. Items that are simply fun can put your archive into a positive light, draw more patrons in, and even help if you want to demonstrate your project to anyone who might be interested in funding your archive's efforts.

Dividing Your Collection

When deciding what to digitize first, you might want to divide your collection into parts. For example, suppose that your collection has local maps, books on Civil War history, and books by local authors. Your maps don't get a lot of use and you know that there are already digital copies of the books by local authors. You also know that there's an active historical society in town and that your Civil War collection includes some unique primary resources. You therefore decide that your project will digitize all the Civil War books first. By saying that anything in the category of "Civil War history books" should be digitized, you can make things easier on anyone working on this project by making the important choices ahead of time.

You can also make the process of storage and digitization easier this way. For instance, you may decide that *all* the Civil War books are going to be photographed and that the images will be converted to gray scale, regardless of whether or not they are in good condition, while another part of the collection will be scanned with a flatbed scanner. Having a consistent procedure in place also makes the work easier.

Something you may need to consider in your plans is security for your collection. What this entails will vary depending upon what your collection contains. For example, the equipment and storage devices you use can be appealing items for theft. Along with potential human threats, protecting your materials from disasters, such as fire or flood, should be part of your plans if it isn't already. For online collections, be sure that your security is adequate (or that the security of the cloud storage service or online host you use is adequate). This is particularly important if you have items that should have restricted access, but is also important in general, as it is possible (though unlikely) for someone to get into your files for malicious purposes—just to see if it can be done.

If you approach your project in this manner, don't make categories mentally. Write down what parts of the collection you think are most important, as well as instructions for how to determine whether or not an item falls into that category. You might have blurred categories—for instance, a book on Civil War history by a local author. You might also have items that don't fit tidily into your categories—for instance, a book on history that has only one chapter about the Civil War. If everything is clearly written down, then anyone working on the project will have no trouble determining what to do, and the project can proceed without the need for you to constantly supervise or make decisions.

Timeline

It may be difficult to determine how much time everything will take until you actually start a project and see how long it takes to process a digital item, and things may speed

up and slow down during a project. For instance, things will speed up as you and other workers develop skill at digitizing material or processing files, or might slow down if you need to train someone new or someone goes on vacation, and so on.

However, forming goals and having a specific desired timeline for your project is very useful. Again, you can use such a plan to explain your project and your goals to others, and it's very helpful for motivating you and others working on the project. For example, you can use a timeline to see how much progress you've made, to try to keep on a consistent schedule, and to celebrate milestones in your project.

Along with deciding what materials you will archive and what should be archived first, you will probably need to deal with an important aspect of digital archiving before you proceed to creating your collection: legal issues.

Legal Issues

In chapter 14, you learned a bit about what kinds of legal problems can arise from archiving. The last thing you want to do while you work to preserve information for the future is waste time and money getting tangled up in a legal dispute. You should therefore make determining whether or not you have the right to store or digitize something or obtaining permission to store or digitize items an important part of your archiving process.

ITEMS THAT CANNOT BE COPYRIGHTED

- Any work that falls into the public domain upon creation
- Any work whose creator has given up his or her rights
- Slogans, titles, names, and other short phrases
- Lists of ingredients, processes, and methods
- Phone books and similar data
- News and facts
- Common or familiar symbols
- Common property works, such as height or weight charts, calendars, or similar items

Essentially, you need to determine if the work you want to store is under copyright. Items that are not under copyright are free for you to digitize. If the item is in the public domain (has no copyright protection), then you're free to use it in your digital collection. Determining if items still have a copyright is a bit tricky, especially if you want to use something published in another country, since laws may be different there. As stated in chapter 14, in general, if a work meets one of the following conditions, it's in the public domain and is free to use:

- The work was published in the United States before 1925.
- The work was published in the United States between 1925 and 1963 and has no copyright notice.

- The work was published in the United States between 1925 and 1963 and has a copyright notice, but that copyright wasn't renewed. This takes some investigating to determine.

Works published after 1963 had an automatic copyright renewal or had the copyright extended, depending upon the year of publication. This means that, if you want to use something published after that year, you'll probably need to obtain permission to use the work for your digital collection, especially if you want to display the item online or share it with your patrons, and not simply store the item.

Some people use alternatives to copyright law that might be of interest to you; these are typically used for born-digital materials. These alternatives limit the rights of creators and give more rights to the public, and so items that use these alternatives are probably safe for you to archive, though you should always consult legal counsel to be absolutely sure. Some of these types of licenses are listed in chapter 14.

Remember, archives and libraries do get some special treatment when it comes to copyright law. You should know what these are and what situations they apply to. For example, if you simply want to archive the data as a backup and not make it available to the public, then this is usually acceptable. Archives can also sometimes make copies of materials that are degrading and a new or better copy can't be obtained for a reasonable price.

This book can only give you general guidelines for addressing and avoiding legal problems with your archive. When in doubt, always consult an attorney. While your archive may have good intentions, copyright holders and their lawyers may not see it that way.

Once you have determined what needs storing first and what you have the right to archive, a good next step to take is to decide how you are going to store your data.

Storage

Once you know how much of your collection you can or want to store, you can make a reasonable estimate regarding how much storage space you'll need and what will or will not be a practical method of storing your data. Deciding how you'll store your data can be difficult, and you should think of it as an ongoing process because whatever storage method you use will almost assuredly become outdated.

For example, suppose that you have a collection of photos of local historic homes. You decide that you only want to store this part of your collection, so you scan these particular photos. You then back up your scans on CDs and then make the photos available online. Your community really enjoys the photos and you get some funding from a local historical society, so you decide to expand your project. Unfortunately, you determine that the next phase of your digitization project, which involves digitizing books about local history, is not going to be practical to store on CDs. You're going to need to change your policies and plans to accommodate this, and decide what will be practical given the scope of your project at this time.

Replacing outdated or broken technology is something else that must be part of your plans. For instance, you might have three 500 GB external hard drives, giving you a total of a terabyte and a half of data, and you find that this works very well for your collection for many years. But ten years from now, 500 GB might be an extremely small amount of

storage, and this storage method might be considered highly outdated. You might be able to buy something that runs faster or is more reliable.

If you use cloud storage, you will need to research your options regularly to ensure that you're getting a good deal with your service and that you still agree to the terms of service for your cloud storage service, since terms of service can change.

There are many, many ways that you can store your data. This book has covered the most commonly used ones today. Your options will very likely change in the future, as well, as people develop new or better ways to store data.

Your data storage device is an object that will, itself, require a method of storage—that is, you'll need to protect your storage devices to keep your data safe. An important general consideration to make is whether or not you have optimal storage conditions available for a particular data storage method. Magnetic tape, for instance, is rather delicate and needs a room that is ideally situated and has controlled temperature and humidity. Flash memory devices lend themselves better to less-than-ideal conditions.

Each method of data storage has some pros and cons, and you should think carefully about what would be optimal for your archive in particular. You do not have to pick only one, and using more than one is a good idea if it's possible. That is, you might want to use a hard drive, but have the same data backed up on tapes. Using multiple devices makes your data safer. If, in the above example, you take your tapes off-site and leave the hard drive in the archive, if something happens to your main archive and the hard drive is destroyed, then your data is still safe. If hard drives were to, say, become obsolete, your data is still safe on the tapes. Again, having multiple copies of your data on multiple devices is a good idea and will keep your data as safe as possible.

After deciding where you'll store your data, the next step you should take is to determine what kinds of files and specifications you want to use.

File Settings

Deciding what specifications you'll use for your files or determining what equipment you need are good steps to take after deciding how you'll store your data. However, determining what kinds of files you want to store and create may make it easier for you to decide what your equipment needs to be capable of, and so you might want to pick your equipment later.

For digitizing tangible materials, you'll need to decide what formats will most accurately represent the original, or, in the case of items like text, what settings you need in order to further process the data (i.e., you may want to make scans of text searchable by a computer). For storing born-digital materials, you'll need to decide whether or not you want to convert to a more archive-friendly format.

Write down what formats you think will be best for your project; you may want to note why, as well, especially if you need to discuss the matter with others. You may need multiple formats, either to protect the files against obsolescence or to make your collection optimal in varying situations. For example, you might want to have an image stored as a TIFF for archiving, but have another copy saved as a PNG for sharing online.

You may decide that you'll need to outsource your project or that outsourcing is going to be more efficient. Even if this is the case, you should still decide what kinds of files you want and what settings you require, as well as find out what files the company

you choose will create for your materials and whether or not you find this satisfactory for your archive's needs.

Images

When it comes to images, you need to make a few decisions about the most basic parameters of your images: whether or not to use color and what level of color to use, the resolution, and the file formats, as well as how you will capture the images in the case of digitization, and what software is necessary. How computers store image data was discussed in chapter 3. The following are some factors to consider in your project.

Color

Images may be in color, gray scale, or black and white. Digital cameras typically capture images in color, and scanners often have several options. You can also convert color images to gray scale using photo-editing software. Remember, color requires the most data to store, gray scale requires less, and monochromatic 1-bit color requires the least (this might be suitable for scans of text if the text is clean and easy to read).

You may want to have different policies regarding color level depending upon what kind of image is being stored and whether or not it is in color. This is fine so long as you are consistent and can clearly define which parameters for image capture are to be used for which situations to avoid confusion.

QUESTIONS TO CONSIDER FOR IMAGE DATA STORAGE

- Are your images in black and white or are they in color?
- Will you need to convert color images to gray scale?
- Do you have a variety of images to digitize? Would it be simpler for you to simply use the same settings regardless of the type of image, or would it be more efficient to customize your settings to the particular image?
- Do you benefit from having a ppi higher than 300–400?
- Will having a higher ppi make storage and retrieval of data too cumbersome for your archive?
- Do you want to change the ppi for born-digital images, or always use the original format?
- Are you archiving images that require a specialized scanner?
- What is your budget for equipment?
- Which formats do you feel will make it easy for patrons to access your information?
- Do you want to share your data online?
- For born-digital materials, what is the original file format for the item?

DPI or PPI

After determining how you will handle the color level for your images, you must then determine what dpi or ppi you wish to use for your images. For consistency, you should

scan your images at the same dpi or ppi. As explained in chapter 3, it's better to work in ppi, or pixels per inch, whenever possible, rather than dpi, because ppi is technically more accurate. For archiving, a range of 300–400 ppi is usually adequate, but you may want a higher range for greater detail and resolution. Some archiving projects use 600 ppi, and you can use an even higher range.

If you want a higher range, you should remember that images with a high ppi require more storage and take longer to store and retrieve, and so you should take this into account when making a decision. A lower range may produce an image that is too low in quality to be of historical use. As with color, you can use different ppi ranges for different materials, but make it clear when a lower ppi and when a higher ppi should be used for an image.

With born-digital images, you can *reduce* the ppi, but *increasing* it requires the computer to make “estimates” and guess what color information would be in the extra pixels, which is inaccurate. It's best in most situations to either reduce ppi only, or always leave it in the original format.

Capturing Images

Unless your image is born digital, you'll need to use equipment to get a capture of the item or to digitize it. The two main choices are to use a scanner or a camera. In most instances, the scanner is the simpler method, unless the item is either too large to scan easily or the item is delicate. In these instances, a camera can be the superior choice.

As discussed in chapter 12, you have a variety of choices for scanners. A flatbed scanner is a good choice for images in most situations, but not all. You may need a specialty scanner if you want to digitize film, or microfilm, for instance, and if your archive wants to solely create high-quality archival captures of images and you have the funds, you may even find it worthwhile to invest in a drum scanner.

File Formats

Once you've created an image, you need to decide what format you require to store it. In general, the TIFF format is best for archiving, since this format captures a lot of information and uses lossless compression. However, TIFF files are typically quite large, and they are also a poor choice for transmission over the Internet, so you should consider your options carefully. The JPEG, GIF, and PNG formats are better suited for use on the Internet. You can store an image in multiple file formats, having a format for storage purposes only and one for viewing online.

Even if you don't want to have your archive available via the Internet, having multiple file formats for your image helps to protect them from obsolescence. For instance, if you had the same image in both TIFF and JPEG formats and suddenly no one was making software that read JPEGs anymore (though this is a scenario that is highly unlikely to happen anytime soon), then your files would still be safe because the other format, the TIFF, is still in use. A similar (and more probable) scenario would be keeping digital photo files in their native formats for the sake of posterity, but also saving those same images in a more commonly used file format. In addition, if you're storing born-digital images, you might want to keep an image in its original format, but also keep a copy in another format. Table 3.1 summarizes the various commonly used file formats.

Processing

After capturing your images, you may want to do some alterations. This typically involves rotating or de-skewing an image, cropping out unnecessary information, and improving the contrast or color levels. De-skewing an image can eliminate data around the edges of an image, so you should always do that first.

Write down the steps that you want to take to improve your captures; that way, everyone who works on the project does it the same way every time, improving quality and consistency and reducing confusion regarding how the work should be done.

Text

Like photos, your text items may be in a digital format already, or they may need to be digitized. Each will require a slightly different approach. Storing text was covered in chapter 4.

QUESTIONS TO CONSIDER FOR TEXT DATA STORAGE

- Does your archive benefit from having text items in multiple formats?
- Will patrons be accessing this data?
- Do you lose significant amounts of data by converting to another format? Is the formatting of the text item important, or is the information itself more important?
- Do you want to keep a copy of the file's original format?
- Do you want to use a camera or a scanner, or some mixture of both? What kind of scanner is best for your materials? Do you have delicate books that require special equipment? How many books will you be processing, and how many will you ideally process per day?
- What settings are optimal for capturing your text materials? Do you want to further process the images by making the text searchable?
- Are your text materials clear and easy to read, or are they faded? Does this impact what settings you choose for capture?

Digital Format

If your text is already in a digital format, then you essentially need to make one decision: Is it better to keep the text in its original format, or to convert it to something else? You can also do both and keep both copies, which may be better from a preservation standpoint in that you have a copy of what the original item was like.

The biggest problem with items already in a digital format is the fact that many file formats for text are proprietary and may become obsolete if the software to read them becomes outdated. In this book, you learned about some formats that are relatively safe from becoming obsolete. TXT, RTF, and OTD are some fairly safe options. TXT allows for very little formatting, RTF allows for some more, and OTD has essentially the full range of word processing formatting options available today.

HTML and XML are markup languages that work with plain text. HTML is generally only useful for web pages, but since it allows for formatting with plain text, you may find it useful for other purposes. A web page does not need to be online to be viewed in a browser program if the data for the web page is on the computer that you want to view it on. XML allows for things like metadata, which you might find useful, as metadata can be used to generate information about your collection and to help with patron access. Metadata was covered in chapter 13.

PDF files allow for “pages,” and can feel like the original document to the user. The PDF/A in particular is a variation on the PDF format that is geared toward archiving. The PDF format is very user-friendly as well as potentially searchable, and is a good choice if you want your archived documents to be used by patrons.

Digitized Materials

If your text needs to be digitized, then you’ll be taking a picture of it, either through a scan or a photograph. It will need to be treated in a similar manner to an image, and so you’ll need to consider some of the same questions as you did regarding images. As with images, you will be using either scanners or cameras to capture your data. Cameras are good in most situations with text items, and flatbed scanners are good for text items that are either not bound or have a strong binding that can withstand rough handling. However, if you don’t mind destroying the original object and the text is in good condition, a sheet-fed scanner is very efficient with text, and so it is an extra consideration for your choices. Overhead scanners are expensive, but particularly efficient if you need to digitize a large number of items and preserve the original object.

Precision is important with text. Like plain image files, TIFF files tend to be best for your original image captures, but you can use other formats if needed or desired. The JPEG format tends not to work as well with text, because the text can become fuzzy due to the manner in which JPEGs are encoded and compressed, but you can use this file type if desired. In essence, you need a lossless format that can capture crisp edges. You may want to use a higher resolution with text than with ordinary images. This will allow the user to zoom in on words and may help optical character recognition software function better should you choose to use it.

If you want your text to be searchable, then you need to scan the original items with this in mind. As discussed in chapter 4, optical character recognition software “looks” for matches to an internal library of letters and characters. You need to scan items in such a way as to make this as easy as possible for the software. If you have very clean, clear documents with no fading, damage, or images, then you should use 1-bit monochrome to get a plain black-and-white document with a small file size. If there are spots or faded areas that would show up as black splotches in a scan that only has two colors, then gray scale is better.

Once you have scanned your items, you can leave them as a sequence of images, but it’s better to put them into something like a PDF format for convenience. PDFs can be made from text documents or images with equal ease, and, as mentioned earlier, “feel” like pages in a book to the user. Storing the images like this will keep relevant items together, as well, which is helpful to you as an archivist.

Audio/Video

Like text materials, your audio or video materials may already be in a digital format, or they may need to be digitized. Unlike text, in which scans of tangible pages and born-digital text documents are two entirely different types of files, the same file formats are suitable for both digitized and born-digital audio and video materials. Audio and video formats for archiving are less straightforward than those for image and text items, and finding truly archival-quality formats is a struggle. For example, though the MP3 format is a highly efficient and popular method of storing audio data, it uses lossy compression, which means that some of the audio data will be lost when the file is compressed to save space. Lossy and lossless compression methods were described in chapters 3 and 5, and audio and video data storage was discussed in chapter 5.

It's a good idea to save digital files using more than one file format. For example, the WAV and AIFF are two common proprietary formats that are extremely similar in capability. You could store a single sound file in both formats, making it safe in the event that any of the companies or corporations who own these formats go out of business or their file format goes out of use. If your audio data already has a digital format, then one of the formats you use should probably be the original format.

You'll face similar difficulties with video formats, since most formats are proprietary and use lossy data compression. You can use similar considerations with video formats, by storing more than one file format and choosing the formats with the least amount of data loss that you can. With video formats, you'll need to compromise between lossy and lossless storage. Lossless storage methods are superior from an archiving standpoint, but are uncommon for video owing to the huge amount of data a video might contain if stored in a lossless way. Lossy methods are much more common and there are more choices for storing a video with lossy compression. Again, you do not have to use a compression method at all, but this will create very large files. In addition, while there are formats that are fairly friendly for archiving, the less archiving-friendly formats are more likely to have more software options.

QUESTIONS TO CONSIDER FOR AUDIO AND VIDEO DATA STORAGE

- Do you want your files to be available online?
- Do you want to offer a transcription of your audio and video files?
- Does it matter to you how well suited a file is toward digital archiving, if the information is originally in a digital format?
- Does it matter to you if a format or method of compression is lossy? Remember, lossy formats are more common and more efficient, but are inferior from a pure archiving standpoint.
- Is the software you want to use compatible with the formats that you'd prefer? Will the software determine what formats you will ultimately save your files in?

Software and Other Digital Materials

Storing things like software and websites is a complicated task that requires a lot of different considerations. Again, consistency is important. You may not be able to achieve optimal storage for these types of digital materials, so make a plan for what you want to store, what the optimal situation is, and what is suitable in the event that you can't have the optimal scenario.

For example, with software, you need to have access to an operating system that will run the software. Depending upon what it is, this might be complex. You may not have a computer capable of running the software, and so you may need to use an emulator instead.

Additionally, with software, being able to store the source code is the best-case scenario. Remember, source code is the original code as it was written by a programmer. If you have the source code, it is often possible to read the code and determine how the software works (how easy this is can depend on a variety of factors, including the skill level of the programmer and the language in which the code is written). Compiled code, on the other hand, is code that has been converted into a format that is possible for a computer to execute. This is useful to store, but not as helpful as having the source code.

For databases, you will actually have some of the same considerations required for digital text documents, as many database files are, in essence, a type of text document. That is, you need to consider things such as whether it would be optimal to convert the data to another format, and if yes, what format is most useful for you and most practical for your database.

Online materials, such as websites (especially websites with dynamic content, like social media) may again require a compromise. The ideal scenario is to have all of the files required to make a website run, such as images or videos, HTML files, or style sheets, which control the visual aspect of a website. You can gather some data about a site using a program similar to a web crawler. Again, a web crawler is a program that gathers information about materials available online, generally for use by search engines, which match the information gathered by the crawlers to search terms entered by users. In this case, you need to essentially decide which sites or which types of sites you will archive and how often you will capture a record of a site (remember, web content changes all the time).

However, some information on a website cannot be gathered in this manner. For instance, many websites have programs running on the server that hosts the website. As a user, you do not have access to the server; remember, you request data from the server, which is then shown to you on your computer (known as the *client side*). Such programs allow for complex functions, like posting comments or filling out order forms. Without the programs running on the server, any pages you save will have their original appearance, but not their original functionality.

There are also such things as "client-side" programs, but these can have security issues and are limited in their use.

Websites may additionally have databases or other files on the server that are required for full functionality of the website. If you don't have access to these files, you won't have fully archived a website. Depending upon what you want to accomplish, this might be unnecessary anyway. It may also be possible to get the cooperation of a website owner to gain all files required to faithfully reproduce a website.

E-mails are like other types of born-digital materials in that you will need to decide whether to keep the e-mail in its original file format or to convert it. There are several suitable file formats for e-mails, and again, using more than one for the same document can be helpful.

Once you know what kinds of files you need and what settings you want to use, you can go about selecting the optimal equipment for your project.

Equipment

The type of equipment that you need will depend upon what you're archiving. If you want to, say, digitize a collection of local photographs, you'll need more equipment than if you were archiving the websites of university professors, although archiving websites may require some specialty software. In chapter 12, this book covered the type of equipment that you will most likely need in detail; this section will review that information. No matter what you archive, though, you have to have two items: a computer and a monitor.

Choosing Computers

Though tablets and smartphones are highly prevalent today, you'll probably still want a dedicated desktop or, at the least, a laptop computer. This is largely owing to the greater computing power offered by these devices. Tablets and smartphones also have fewer and more specialized ports, which makes using equipment more difficult. This may change in the future, but for now, you should really consider a desktop or laptop computer for your project.

When you buy a computer, it can be difficult to compare one computer to another or to understand why one is better than another. The basic parts of a computer were described in chapter 2. Computers may have features that seem complicated or difficult to understand, or you may have trouble determining why two similar computers have different prices, or why one computer might be better than another. Though evaluation options can be a little confusing and some merits of a computer are largely opinion, doing a general comparison does not have to be difficult. Here are some things to look for when you want to make a purchase.

FEATURES TO LOOK FOR IN COMPUTERS

- Operating system, manufacturer
- Size of the hard drive
- Speed of the CPU
- Amount and type of RAM
- Number and type of ports

Operating System

As explained earlier in this book, an operating system is essentially the general software that allows a computer to coordinate its own functions. This affects everything about a

computer and how it runs. Operating systems change over time, with companies updating the systems to meet new customer demands. If you buy a new computer, it will probably have the latest operating system. If you buy a used computer, be sure that you know what the operating system is. Sometimes old computers can be upgraded, and sometimes they can't.

The type of operating system and the company that made it will determine what software you can use with the computer. Software is not universal. That is, software is designed to run on certain computers and operating systems. The software that you want to use can determine what kind of computer you want to buy, though it's more common to select software based on what kind of a computer you have.

Different companies will sell computers with different operating systems. For desktops and laptops, Windows and Apple are the major choices. Remember, software designed for one of these operating systems won't necessarily work for the other. There are also other operating systems currently, such as Linux, but these are less commonly used.

Hard Drive / Solid-State Drive

Another easy thing to look for is how big the computer's main memory is; this will likely be a hard drive (HDD) or a solid-state drive (SSD). Bigger is always better when it comes to memory. Modern HDD or SSD sizes are listed in either gigabytes (GB) or terabytes (TB). Remember, a terabyte is 1,000 gigabytes, so a 1-terabyte hard drive is twice as big as a 500-gigabyte hard drive. However, you should also remember that a bigger HDD or SSD may come with a higher price tag. Computers with HDDs may be less expensive.

CPU

The Central Processing Unit, or CPU, is the part of the computer that handles all calculations and other requests made by the user. It's therefore an essential component. The speed at which the CPU operates is determined by the clock, which is a vibrating crystal inside the computer. This rate is usually designated in Gigahertz (GHz). The bigger this number, the faster the CPU can operate, and the faster the computer can operate as a whole. You may want to see if any software that you want to use has a minimum CPU requirement to help you determine the optimal CPU for your computer.

You may see terms like "dual-core" or "quad-core" or even "eight-core" when it comes to CPUs. A computer can have more than one CPU or CPUs with duplicate components, which also improves the speed.

In addition, the CPU has its own memory, known as *cache memory*. More cache memory also helps improve the processing speed by having plenty of space to store the CPU's calculations. Finally, a term often connected with CPUs is the front-side bus, or FSB, which is the portion of the CPU that allows for communication between the CPU and the rest of the computer. The speed of the FSB directly impacts how fast the CPU can respond to the user's requests.

RAM

There are several types of RAM chips, some of which are faster than others. You can often add more RAM to a computer and it's not too difficult to do. However, you need to be sure that the chips you purchase are compatible with the computer. In addition, the

different types of RAM have some similar acronyms, so it's important to take care when researching and purchasing RAM chips for your computer, as it's possible to get them confused. A computer's manufacturer may have a guide or online store where you can be sure to get the right components (although buying parts this way can be pricey). If you can, it's better to purchase a computer with a sufficient amount of RAM just for the sake of ease. RAM chips were covered in more detail in chapter 2.

Ports

A computer must have ports, or else it is not able to communicate with the outside world. The number and types of ports will be important to you. Most computer manufacturers advertise how many USB ports are available in particular. Many different types of equipment can use USB ports to communicate with the computer, so more USB ports are better. It's possible to buy a hub that plugs into a USB port and gives you more USB ports, though, so it should not be your only consideration. In addition, there are a couple of different types of USB ports. USB A-type ports are extremely common at the moment, but the smaller USB C-type ports are becoming increasingly more common and having such a port could be handy.

Some other common ports are ports for microphones and speakers, and ports for monitors or televisions, such as VGA, DVI, and HDMI ports.

Choosing Monitors

Choosing a monitor is simpler in some ways than choosing a computer. In many stores, samples of the available monitors will be powered on so that you can see the difference between the monitors.

FEATURES TO LOOK FOR WHEN PURCHASING MONITORS

- Physical size of the monitor (expressed as the diagonal measurement)
- Resolution (number of pixels, or how clear and sharp the picture is)
- Ports (how it can connect to the computer)
- Other features (for example, whether it has speakers or whether the monitor is adjustable)

Monitors should be calibrated to match real-world colors and contrast for best results. Bigger monitors are better, especially if you're going to work with images and video files, but this is only true up to a point. Of course, don't get one that's too large to reasonably fit into the space you have to work. A monitor that's too big can be difficult to use, as well, or increase eye strain. Keep in mind that you can also use multiple monitors simultaneously, and thus look at more than one program or image at a time. This can be very handy in some situations.

While a computer and a monitor are the major items of equipment that you'll need, there will probably be others, such as speakers, mice, and keyboards. You'll also need more

equipment, as well as software, if you will be digitizing materials. Some of this equipment can be used for projects other than digitizing materials, as well.

Scanners and Cameras

With images, you need a way to record the image, also referred to as capturing the image. You have two options: a scanner or a camera. In general, use a scanner if the following conditions exist:

- The images are small enough to fit on the scanner bed.
- The items are not delicate, or, in the case of a book, can be taken apart for the digitization process.

In general, use a camera if the following conditions exist:

- The item is large.
- The item is delicate or the item is a book that should not or cannot be taken apart.

There are several types of scanners to choose from:

- Flatbed scanners are common and are useful for most situations, but are somewhat tedious to use.
- Sheet-fed scanners are partially automated and are faster than flatbed scanners. They are not suited for delicate items and can't scan books unless the books are taken apart.
- Overhead scanners are expensive, but are ideal for quickly scanning things like books.
- Portable scanners don't produce archival-quality results, but may still be useful in some instances.
- Drum scanners are useful for high-quality scans of photos. They produce the highest-quality scans, but are far less convenient than your other options and are very large and expensive.
- You may need a scanner designed specifically for some items; negatives and microfilm are some items that really need a specialty scanner.

A good scanner will at minimum do the following tasks:

- Scan items at your archive's target dpi or ppi
- Scan items and convert the data into multiple formats, including the formats that you want to use
- Scan in color, gray scale, and monochrome
- Connect easily to a port on your computer

With cameras, you should be concerned mainly with the size of the sensor and how many megapixels the camera captures. Bigger is better in both instances. You need some extra equipment with a camera setup—such as a book cradle for books, a stand or platform for large images, something to hold the camera steady (such as a tripod or an arm), and something to discreetly hold down pages.

Audio/Video

If your audio data is already on something like an optical disk, then you simply need to “rip” the music from the disk. A typical computer already comes with software that will do this for you.

If you need to digitize your data, then you basically require two things: something that can play the recording (a record player, for instance) and something that can connect the player to your computer so that the computer can capture the audio data. Chapter 12 describes what you need for this process in more detail.

Video data can also be directly copied, just as digital audio data can be. If you want to record a VHS tape in particular, then there are devices that are specifically designed to capture and digitize the information. You will require one of these devices and a VHS player. Film reels are a little more complex, especially if they have an audio component, but there is equipment that can be used for this, as well.

Processing

Once you’ve created your digital files, you’ll then need to start making them practical to use by the public or your staff. This will involve things like cleaning up captures, adding metadata, and creating catalog entries.

Cleaning Up Files

Once you’ve created your digital files, you need to decide what the next step is. These are some questions you might consider:

- Are you going to adjust images for brightness/contrast, rotate or de-skew, or do any other digital cleanup or enhancement?
- Do you want text to be searchable?
- Do you want to divide audio or video captures into segments (like individual tracks), or leave them as complete files?
- Should digital restoration be part of your project? For example, if you have a photograph with a tear in it, you could potentially use a photo-editing program to remove it. Though this isn’t “pure” from an archiving standpoint, your patrons might like it.
- Are the file names important? If they aren’t, then you can save some time by numbering files or using another simple system rather than trying to think of a good descriptive title for each item.

You might want to digitize a few items in order to figure out what process will be best before coming up with an official plan. As an example, suppose that you want to digitize some photographs. After digitizing about a dozen, you decide that the best procedure is to scan the photo, de-skew it in a photo-editing program, enhance the contrast, name the file using the photographer’s last name and a number, save it as both a TIFF and a JPEG, add metadata, and then catalog the entry.

Metadata

As discussed in chapter 13, your files aren't useful if no one can find or access them. Typically, you'll want to add some kind of metadata to the file. Software for images, audio, and video data in particular may give you the option to permanently add metadata to the file, and whether or not software is capable of this might determine whether or not you want to use a certain type of software. There are ways to embed metadata in text files, as well. You may or may not want your patrons to access your digital archive, or you may want to limit the access or who you want to access the data.

METADATA AND PATRON ACCESS

- Do you have programs suitable for adding metadata?
- How much metadata do you want to add, and what kind of information is important to put into the metadata?
- Are you digitizing items that already have a useable catalog entry?
- What type of metadata schema do you want to use?
- How much patron access do you want to allow?
- How will you prevent others from tampering with your files?

Once you've created and stored your files, your work is not finished. Digital archiving is an ongoing process.

Maintaining Your Archive

As mentioned at other points in this book, digital archiving is different from regular archiving. Your collection will need to be monitored and maintained, to protect it against both physical degradation and obsolescence. Chapter 15 discussed some of the major issues with digital storage and digital archiving. You need to have a good plan in place for how you'll deal with changes in technology and hardware or software issues. Some thoughts to consider are:

- How often do you think you should check your collection for corruption, errors, or physical damage or degradation? How will you go about this in an efficient manner? What will you do if you discover damage or errors? There is software available that can help with this.
- If you find that part of your collection is in danger of becoming outdated, what is the best way to move your files without losing any data?
- How will you monitor relevant developments in digital storage?
- What will your archive do in the event that your software is no longer usable for your project, or in the event that a more efficient method of archiving is discovered?

Key Points

- Your collection may be very large or small, or you may only want to digitize part of a collection. In most instances, you will need to prioritize which parts you address first.
- Potential legal problems should be addressed before any digitization or storage takes place.
- You will need some way to store your digital files. Using multiple storage methods is the optimal choice. The ideal method will depend upon how much data you need to store and how much money your archive has.
- You may want to store your data in multiple file formats. You should always choose your formats based on what is optimal for your project's goals.
- You will require equipment; at the minimum, you need a computer and a monitor. Other equipment that you need will depend upon what you want to store. Often times, choosing equipment requires a compromise.
- You will need to add metadata to your items; the best way to accomplish this will depend upon what kinds of files you are storing, your method of cataloging, and what software you have available, as some software can embed metadata.
- One of the primary considerations for your archive must be how you will maintain and upgrade your collection.

Without a doubt, the future of archiving will be digital and require the user of computers. While the process of digital archiving may seem unfamiliar or intimidating, the amount of data that can be stored and shared will be highly beneficial for future generations. Your efforts now can benefit others for many years to come.

Reference

Hirtle, Peter B. 2008. "The History and Current State of Digital Preservation in the United States." *Metadata and Digital Collections: A Festschrift in Honor of Tom Turner*. CIP (Cornell University Library Initiatives in Publishing, Ithaca, NY). https://ecommons.cornell.edu/bitstream/handle/1813/45862/7the_history_and_current_state_p_hirtle.pdf.

Index



- 16-bit color. *See* high color
- 24-bit color. *See* true color
- 32-bit color. *See* true color
- 1976 Copyright Act, 190, 196

- AAC. *See* Advanced Audio Coding
- adaptive multi-rate, 61
- advanced audio coding, 61
- AIFF, 61
- ALAC, 61
- American Memory, 8
- American Standard Code for Information Interchange.
 See ASCII
- AMR. *See* adaptive multi-rate
- analog data, 58–59, 168
- analog-to-digital converter, 168
- Apple File System, 114
- Apple Lossless Codec. *See* ALAC
- ASCII, 43
- aspect ratio, 66
- Audio Interchange File Format. *See* AIFF
- audio port, 24
- autoloader, 105

- back side bus, 19
- basic input/output system. *See* BIOS
- binary, 13–15, 207–8
- binary digit. *See* bit
- BIOS, 22
- bit, 15–16
- bit depth: and scanners, 160; audio, 59–60, 168; image,
 28, 30; video, 66
- bit rot, 120, 212
- Blu-ray, 86–88, 90–91, 210
- Broadcast Wave Format, 61
- browser, 77, 78
- buffering, 113
- bus, 19
- BWF Files. *See* Broadcast Wave Format
- byte, 15–16

- cache memory, 118, 231
- calibration, 159
- camera, 27, 35–37, 163–65, 224–25, 227, 233; and
 metadata, 174; sensor, 35, 164, 233; video, 64, 169
- CARLI, 59–60, 66, 68
- cartridge, 101, 104, 106, 109
- cassette: magnetic, 101–2; music, 60, 98, 102, 167–68;
 player, 167
- cathode ray tube. *See* CRT
- CCD, 159
- CD, 60, 86; construction, 87–88; drive, 87–88; types,
 88–90
- channel: audio, 59–60; color, 38
- character set, 42–43
- charge-coupled device. *See* CCD
- CIS, 159
- client: computer 140, 142, 144, 229; email, 78–79
- clock, 18–19, 231
- cloud computing, 140, 143–53; advantages, 147–48;
 contracts, 150–52; defining, 143–45; disadvantages,
 148–49; models, 145–46
- cluster, 114
- CMOS, 159

codec, 59, 61, 65
 color depth. *See* bit depth
 colorimeter, 159
 comma-separated value file. *See* CSV
 common-property work, 191
 compact disk. *See* CD
 compiled code, 73
 compiler, 74
 complementary metal oxide semiconductor. *See* CMOS
 compression, 28, 31–32, 59, 62; adaptive, 32; asymmetric, 32; lossless, 31, 59, 65, 228; lossy, 28, 31, 59, 65, 228; nonadaptive, 32; ratio, 28; symmetric, 32
 compressor-decompressor. *See* codec
 Conference on Fair Use, 198
 contact image sensor. *See* CIS
 copyright, 185–200; alternate licenses 198–99; anonymous works, 190; databases, 190; laws for archives, 195; orphan works, 190, 194; owner's rights, 188–89; protected works, digital, 193–94; protected works, tangible, 187–88; work for hire, 190; works not under copyright, 191–92
 CPU, 18–20, 147, 231
 Creative Commons, 199
 CRT, 156–57
 CSS3, 77–78
 CSV, 46–47

 database, 76–77
 database management system. *See* DBMS
 data set, 76
 DBMS, 76
 defragmentation, 115
 digital archiving: definition, 2; equipment, 156–69, 230–34; goals, 6–8; organizations, 8–9
 Digital Millennium Copyright Act. *See* DMCA
 digital versatile disk. *See* DVD
 direct access, 102, 115
 disk imaging, 85
 display adapter. *See* graphics card
 DisplayPort, 158
 DMCA, 196–97
 document type definition, 50
 dots per inch. *See* DPI
 DPI, 29, 160, 164, 224–25, 233
 drum scanner, 159, 163, 225, 233
 DVD, 87; construction, 90–91

 DVI, 158, 232
 DTD, 50–51
 dyes, 89, 92–93

 EBCDIC, 43
 electromagnets, 99–100
 e-mail, 78–79
 e-mail client, 79
 EML, 79
 EMLX, 79
 emulator, 75
 enterprise hard drive, 116
 eSATA, 115, 117
 ethernet port, 24
 Extended Binary Coded Decimal Interchange Code. *See* EBCDIC
 extensible markup language. *See* XML
 external hard drive, 115–17, 119, 121

 fair use, 197–98
 FAT, 114
 File Allocation Table. *See* FAT
 file extension, 32, 45, 172
 firmware, 16–17
 FireWire, 23
 FLAC. *See* Free Lossless Audio Codec
 flash drive, 126, 129–30, 133–34, 136
 flash memory, 126; advantages, 131–32; disadvantages, 132–34; storage 135–36
 flash memory cell, 127–28, 135
 flatbed scanner, 161, 163, 225, 227, 233
 floppy disk, 4, 7, 82–86
 floppy drive, 84–85
 fragmentation, 114
 Free Lossless Audio Codec, 61
 Free Software Foundation, 199
 frequency, 13–14; audio, 59
 front side bus. *See* FSB
 FSB, 19
 full-screen, 66

 General Public License. *See* GPL
 GPL, 199
 GPU, 18
 Graphical Processing Unit. *See* GPU
 graphics card, 24, 158

handheld scanner, 162
hard drive, 17, 22, 112–23, 125–27, 129, 131–33, 144, 230–31; advantages, 119; disadvantages, 119–21; storage, 121–23
hard drive enclosure, 116
HDD. *See* hard drive
HDMI, 158, 232
headphones, 60, 168
Hierarchical File System. *See* HFS
High-Definition Multimedia Interface. *See* HDMI
HFS, HFS+, 114
high color, 28, 30
histogram, 38–39
HTML, 45, 47–51, 77, 79, 174, 210, 227
HTML5, 77, 210
Huffyuv, 65
hypertext markup language. *See* HTML

ICR, 54
IEEE 1394 ports. *See* FireWire
image data, 28–38; file types, 32–37
indexed color, 33
infinite loop, 72
Intelligent Character Recognition. *See* ICR
interlaced: image, 33, 35; video, 66
internal hard drive, 115–17, 119
Internet, the, 138–43
Internet Archive, 6, 8, 74, 213
IP address, 142

jewel case, 93
JPEG 2000, 65
jump drive. *See* flash drive

keywords, 77–78, 173–76
kilohertz, 58–59

Lagarith, 65
LAN, 140, 142, 145
lands, 87–89, 91–92
laptop hard drives, 115–16
LaserDisc, 86
laser rot, 92
Latin-1 Extended ASCII set, 43
LCD chapter, 157–58
Library of Congress, the, 8, 53–54, 92, 179, 186, 198, 212
linear pulse-code modulation, 62
liquid crystal display. *See* LCD
local area network. *See* LAN
LPCM. *See* linear pulse-code modulation

machine language, 73
magnetic field, 99–100, 107–8, 207
magnetic tape, 41–42, 98, 100–109, 111–15, 119, 128, 131, 210; advantages, 103–5; construction 101–2; disadvantages, 105–6; history, 97–98; storage, 106–9
magnets, 97–100, 107
mainframe computer, 143, 146–47
MAN, 141
malware, 72, 74
MBOX, 79
megapixels, 164, 233
memory stick. *See* flash drive
metadata 172–81; schemas, 176–80; types, 176
metropolitan area network. *See* MAN
MIDI, 61
monitors, 156–59, 232; resolution, 157–58; size, 158
mono recording, 59–60, 167
motherboard, 16–17, 20
MP3, 61
MPEG-1 Audio Layer 3. *See* MP3
MPEG-4, 61
multi-level cell, 128, 134
Musical Instrument Digital Interface File. *See* MIDI

native formats: images, 36–37; video, 65
New Technology File System. *See* NTFS
noninterlaced image, 34
nonvolatile, 133
NTFS, 114

OCR-A, 53
OCR software, 53–54, 166
ODF, 46–47
OGG, 61
Ogg Vorbis. *See* OGG
OLED, 158
open document format. *See* ODF
open-source code, 74
OpenType fonts, 44
operating system, 73, 74–75
optical character recognition software. *See* OCR software
optical media, 86–94; storage, 93–94
Opus, 61
organic light emitting diode. *See* OLED

outsourcing, 156, 223
 overclocking, 19

PATA, 115, 117
 Parallel ATA. *See* PATA
 parallel port, 24
 PCM. *See* pulse-code modulation
 PDF 35, 45, 51–52, 54, 166, 227
 PDF/A, 51, 166
 PDP, 157
 pen drive. *See* flash drive
 permanent magnets, 99–100
 phono stage. *See* preamp
 photomultiplier tube. *See* PMT
 pits, 87–89, 91–92
 picture element. *See* pixel
 pixel, 28–31, 42, 65–66
 pixels per inch. *See* PPI
 plain text, 45–48, 50, 79
 plasma display panel. *See* PDP
 platform, 75
 PMT, 159
 portable document format. *See* PDF
 portable scanner, 162, 233
 POST, 22
 PostScript fonts, 44
 portable document format. *See* PDF
 power-on self-test. *See* POST
 PPI, 29, 224–25
 preamp, 168
 PREMIS, 178–79
 PREservation Metadata: Implementation Strategies. *See*
 PREMIS
 processor, 19–20, 158
 programming language, 72–73
 progressive encoding, 35
 progressive scanning, 66
 projector, 169
 proprietary formats, 45–46, 61, 166
 PS/2 port, 24
 pseudocode, 72
 public domain, 189–90, 194–95
 pulse-code modulation, 62

RAID array, 119
 RAM chips, 16, 21–22, 112, 126–27, 231–32
 random access memory. *See* RAM
 raster graphic, 35
 rasterizing, 35–36
 raw format. *See* native format
 record player, 167–68
 reel-to-reel film, 169
 reel-to-reel tape, 101
 rendering, 65
 resolution: images, 28–29, 53, 225; monitor (*see* monitor,
 resolution); video, 65–66
 rich text, 45–47
 ripping, 59–60
 ROM chips, 17, 21–22, 88, 125–26
 router, 140–42

sampling rate, 58–59, 168
 SATA, 115, 117
 scanners, 159–63, 224–27, 233
 schema crosswalk, 177
 scripting language, 77
 SCSI, 117
 SD card, 20, 25, 126, 130, 133, 165
 searchable text, 51–54, 166
 secure digital card. *See* SD card
 sensor, camera, 35, 164, 233
 sequential access, 102, 111–12, 115
 Serial ATA. *See* SATA
 serial port, 24
 sector, 114, 120, 123
 single-level cell, 128–29, 134
 small computer system interface. *See* SCSI
 software, 17, 71–75
 solid state drive. *See* SSD
 Sonny Bono Term Extension Act, 190
 source code, 73–74
 SSD, 112, 126–27, 129–30, 231
 standard encoding, 35
 stereo recording, 59–60, 167

tags, 48, 50–51, 76, 172–74, 178–79
 tape library, 105
 television, 29, 65–66, 158, 232
 text data 42–52; file types, 45–52
 three-level cell. *See* triple-level cell
 thumb drive. *See* flash drive
 Thunderbolt port, 23
 timecode, 65
 transistors, 127–28
 triple-level cell, 128, 134
 true color, 28, 30, 38, 42

TrueType fonts, 44
TWAIN, 159
Twitter, 173, 212

underclocking, 19
UNICODE, 43
universal serial bus. *See* USB
USB port, 23, 115, 158, 232; hub, 25

vector graphic, 35
VGA, 156, 232
video card, 18, 158
video codec, 65
video data, 64–68; file types, 66–68
video graphics array. *See* VGA
vinegar syndrome, 101–2
volatile memory, 21

WAV file. *See* WAVE file
WAVE file, 61
wear leveling, 135
web crawler, 77
web font, 44
web-safe font, 44
WIA, 159
widescreen, 66
Windows Media Audio, 61
WMA. *See* Windows Media Audio
WOFF, 44
WOFF2, 44
Wordpad, 47, 50

XML, 76

About the Author



Elizabeth R. Leggett is a technical writer specializing in writing about software and technology. She has worked in libraries and archives at Centre College, the University of Kentucky, and Murray State University. She also began a local digital genealogical collection at the Calloway County Public Library.

