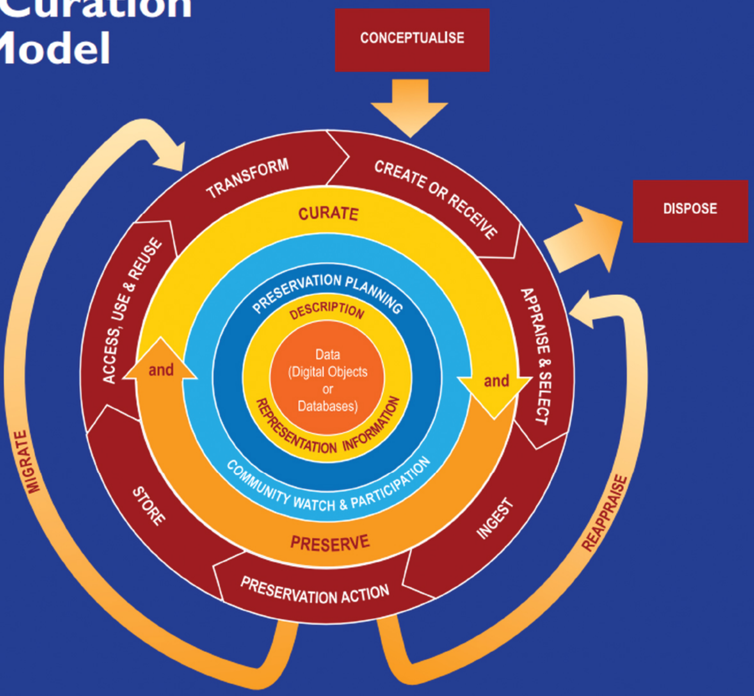
The DCC Curation Lifecycle Model



DIGITAL CURATION FUNDAMENTALS

JODY L. DeRIDDER

Digital Curation Fundamentals

Digital Curation Fundamentals

Jody L. DeRidder

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

Unit A, Whitacre Mews, 26-34 Stannary Street, London SE11 4AB

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Names: DeRidder, Jody L., 1957– author.
Title: Digital curation fundamentals / Jody L. DeRidder.
Description: Lanham : Rowman & Littlefield, [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2018017881 (print) | LCCN 2018033215 (ebook) | ISBN 9781538106792 (electronic) | ISBN 9781538111215 (hardcover : alk. paper) | ISBN 9781538106785 (paperback : alk. paper)
Subjects: LCSH: Data curation in libraries. | Digital preservation.
Classification: LCC ZA4080.4 (ebook) | LCC ZA4080.4 .D47 2018 (print) | DDC 025.8/4—dc23
LC record available at https://lccn.loc.gov/2018017881

∞™ 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.

Printed in the United States of America

Contents

List of Figures and Tables vii

Preface ix

1 What Is Digital Curation? 1

2 What Can I/We Do? 11

3 Models: Which Ones Do I Use When? 21

4 Encapsulation, Emulation, or Migration? 31

5 How Do I Identify and Select Content? 37

6 What Foundational Work Will Prepare Content for Preservation and Access? 49

7 Storage, Protection, and Monitoring: What Do I Do? 75

8 How Do I Provide Access Throughout Time? 87

9 How Can I Leverage the Community? 99

Appendix: Resources 109

Notes 117

Bibliography 141

Index 151

About the Author 155

Figures and Tables

Figure 1.1. Obsolete Floppy Disks 5

Figure 1.2. Cat Photograph Rendered in Figure Viewer (on Left) and Text Editor (on Right) 6

Figure 2.1. Levels of Rights Available in Creative Commons 14

Figure 3.1. DCC Curation Lifecycle Model 22

Figure 3.2. Open Archival Information System Functional Entities 24

Figure 3.3. Digital Preservation Outreach and Education Modules 26

Figure 4.1. Model of the KEEP Emulation Framework 33

Figure 5.1. Useful Tools for Content Identification 41

Figure 5.2. Internet Archive Web Archiving Lifecycle Model 44

Figure 6.1. File System Organization Reflected in File Name 59

Figure 6.2. FITS Test of PDF That Is Not Well-Formed or Valid 62

Figure 6.3. PREMIS Object Categories and Their Relationships 65

Figure 7.1. Example of a Corrupted File 84

Figure 8.1. Disk Image Access for the Web 96

Figure 9.1. Diagram of Peer-to-Peer Replication of Content 102

Table 3.1. Aspects of Curation in Three Models 28

Table 7.1. NDSA Levels of Preservation 80

Table 7.2. Format Reference Table 85

Preface

Websites and digital news stories disappear daily; researchers can't access their own data for reuse; students don't know how to make their work last for the next ten years. Knowledge is built on previously gathered information; but what happens when that information is no longer accessible? And where does the archivist, librarian, or administrator fit into this picture? This book describes the basic steps of data curation, in clear, easy-to-follow language, and clarifies the many potential roles that can help ensure our most valuable information will be viable for generations to come.

WHO THIS BOOK IS FOR

Digital Curation Fundamentals is for the average librarian or archivist who wants to help save knowledge for future use but knows little to nothing about digital curation or how it fits with their jobs. This book is also for administrators who need to stay on top of things but don't yet have a good grasp on the purpose and scope of digital curation and how central it is to the future of librarianship. Additionally, this book is a reference handbook for those who are involved in digital curation in some form but need the context to know how their work fits into the big picture and what comes next. Most of all, this book is for people who want to make a difference but have limited resources with which to work.

For example, staff members involved in capturing digital content will use this book to determine what file formats to use and what metadata to capture in what standards. Staff members who prepare digital content for long-term storage will refer to this book to determine what technical and administrative metadata to capture and how to store it. Administrators developing digital preservation solutions will refer to this volume for storage options and priorities, for example, the need to protect the provenance and usability of the digital resources. Employees involved in digital curation from the point of content development by users to usability studies for provision of access will find recommendations here for methods, approaches, standards and best practices, and resources for more information.

This book takes a straightforward, commonsense approach to a complex problem and portrays the challenges and opportunities in an approachable, conversational style that lowers the bar to include those with little to no technical expertise and few resources.

WHAT THIS BOOK COVERS

Within this book you will find not only the basic theoretical underpinnings of digital curation and the current options for creating access to older content, but also guidance and methods for every aspect of digital curation, from selection through provision of access. For outreach archivists and instructional librarians, there is guidance on how to educate content creators. For hands-on practitioners, there are sections on file formats, metadata schemas, technical metadata extraction and testing, and crucial aspects to consider for selecting content for curation. For those developing infrastructure, there are guidelines on storage and protection, usability studies, and options on web delivery. For those involved in digital curation, there are resources to help leverage community efforts and expand expertise even further.

Digital Curation Fundamentals is a result of the research, testing, implementations, and expertise gained during the fourteen years the author spent developing digital library and digital preservation solutions at both the University of Tennessee and the University of Alabama. As policy maker, collaborator, infrastructure developer, teacher, programmer, and administrator, the author speaks clearly and knowledgeably about every level of digital curation. Keenly aware of the funding and technical expertise limitations within current libraries and archives, she points out practical options and crucial concerns to be weighed when choosing the most appropriate solutions for your organization. As a result of reading this book you will have a comprehensive understanding of the various aspects of digital curation; know how the proponent theories fit together, and how and why they vary; and be armed with sufficient information to make intelligent choices and decisions, based on your own resources and capabilities.

HOW THIS BOOK IS STRUCTURED

This book explores various roles we can play in digital curation, models for developing our policies and processes, and how to best leverage our resources. It lays out the basics necessary for digital curation in an easy-to-understand manner.

Chapter 1 explains the meaning of digital curation, the reason it is important, and the environmental challenges that make it both difficult and rewarding.

Chapter 2 explores roles in digital curation for librarians, archivists, administrators, organizations, and the community.

Chapter 3 explains three top models for digital curation, including how they overlap, why they differ, and when to apply each one effectively.

Chapter 4 speaks to the pros and cons of the three main methods of providing access to older content: encapsulation, emulation, and migration.

Chapter 5 addresses the difficult tasks of sorting through and selecting content for curation, from locally created or obtained materials to those of donors or found on the web.

Chapter 6 covers selection of archival formats in which to store that content and discusses rights issues and the various forms of metadata needed, including technical metadata and useful tools.

Chapter 7 explains how content must then be protected, monitored, and “replicated” (multiple copies in dispersed locations) to ensure longevity.

Chapter 8 covers options for organizing and providing access to older materials throughout time.

Chapter 9 discusses how we can contribute to and leverage community efforts to accomplish much that we cannot do separately.

The appendix contains highly recommended options for next steps.

The bibliography provides a wealth of further reading.

This book is a valuable resource that can help turn the tide against the loss of our digital culture, including many valuable scientific, legal, and historic resources. Used effectively, the guidance and recommendations provided here can help ensure that future generations have the information they need to leverage and build new knowledge, instead of recreating the wheel. This can make the difference in our ability to solve complex political and environmental problems, overcome terrible diseases we have researched for years, and avoid repetition of the mistakes of our past. The nature of our future depends on our collaborative effort. The author offers her depth and breadth of knowledge in the hopes that you will take it, build on it, and put it to work, to help ensure a hopeful and healthy future for our children and grandchildren. Thank you in advance for all you do to provide users with the resources needed, in the form needed, at the point of need. *Digital Curation Fundamentals* will help you move us forward.

1

What Is Digital Curation?

For centuries, librarians and archivists have collected valuable information, organized it, sought to keep it in usable condition, and provided access to users. Throughout the years, the forms of content have changed tremendously, from tablets and scrolls to handwritten tomes to published books and now to digital content. The purpose of librarianship and archiving has always been to collect and provide effective access to (especially curated) information. Currently, providing access extends to teaching users how to perform research and use computers, databases, software, and the web.

Why is it so important to provide effective access to information? The primary reason is because all knowledge builds upon information and knowledge that has been generated before. Year after year, decade after decade, century after century, we build upon what others have learned and done before us. This is particularly evident in the scientific arena. Examples include Isaac Newton's studies of gravity in 1664, Michael Faraday's development of the first electric motor in 1821, Louis Pasteur's realization that disease comes from microorganisms that can be killed by heat and disinfectant, and the discovery of DNA by James Watson and Francis Crick in 1953.[1] Our ability to build structurally sound high-rises and launch space explorations is dependent upon the understanding of gravity and the creation of exceptionally strong materials.

Throughout the years, we have developed almost all our vehicles and machinery as variations of the original electric motor and combustion designs. Current medicine has evolved dramatically from Pasteur's discovery of microorganisms to develop many remedies to countless diseases, and doctors and researchers are still unfolding what knowledge of DNA can mean for us and our progeny. The National Cancer Institute demonstrates the importance of building on prior scientific work in their 250-year time line of milestones in cancer research, from the first X-ray to the development of the first cancer vaccine.[2]

Imagine: What would happen if those initial discoveries were lost? What kind of world would we live in, if we could not fight disease, build tall buildings, and drive cars? If we could not access previous research to build upon it, that research would have to be repeated, over and over, or simply abandoned; we could not compete in the international arena in terms of scientific discoveries, inventions, medical cures, technological advances, economic production, cultural evolution, or any conceivable realm. The loss of our ability to build a better future would be effectively destroyed.

In 1982, Meijer concluded that the main characteristic of librarianship was the "stimulation of the optimum use of mankind's cultural heritage insofar as it consists of coded thoughts recorded in documents that are and must be held in readiness for use with the ultimate objective of making possible cultural progress."[3] The work of librarians and archivists is crucial to cultural progress, although this fact is rarely acknowledged.

There are also more immediate reasons why we must provide effective access to valuable information. Without access to crucial records, businesses, institutions, and individuals alike may be subject to litigation and legal penalties against which there is no proof of innocence. Patents cannot be protected; ownership cannot be proved; actions cannot be justified. Institutions that are mandated to retain certain information permanently, for example, student records, financial transactions, and more, are beginning to find they cannot effectively access and use the information they have stored, especially content in proprietary formats. Copious amounts of content are usually stored in databases, and most databases are proprietary, which means the encoding and organization of their files is secret, so that others cannot build systems to access their files. If a software company goes out of business, it is unlikely that the files that software created can ever be made usable again. And as software changes, it may or may not be able to access older stored content created by that same type of software.

For example, in 2015, the author was asked to try to repair corrupted databases from 2008 that contained legal information that had been created with a proprietary database system. A review of the content at a low level showed that shifting of the content had occurred, so that information assigned to some fields was now in other sections, and some information was totally unreadable. Although the original software company was hired to restore the files, they were only partially successful. Not even software companies retain all the previous versions of their code in forms that can be reused. From tax records to presidential e-mails to documented policy decisions, our digital history is at risk, and that risk grows with every software and hardware update and change.

The risk of loss is not only in institutional holdings. A tremendous amount of current history is captured on the web, but that information is mutable and may change at any time. As websites constantly change, the record of our history disappears before our eyes. For example, with the change of presidents in 2017, the WhiteHouse.gov site began losing web pages, starting with those concerning policies on climate change, health care, and civil rights.[4] This was certainly not the first time important web content has vanished. In 2013, three major web databases of nuclear history disappeared.[5] From 2005 to 2008, the official history of the war in Iraq was continually revised on the WhiteHouse.gov website, without public notification.[6] With the blink of an eye, policies can change, and the websites previously considered the most reliable sources of information can leave the public feeling blindsided and betrayed. This mutability extends to the entire web and all types of online content. Estimates of how long the average web page lasts vary from forty-four to 100 days.[7] "Link rot" (the continued existence of links to content no longer online) is so widespread that in 2014, more than 70 percent of the links in the *Harvard Law Review* were no longer accessible.[8]

What does this loss mean in terms of our cultural history? What documented evidence of our current history will survive for the use of future historians? Vint Cerf is famous for having coined the term "Digital Dark Age," for where we are now; he fears that little or no record of the twenty-first century will survive for future generations.[9] Published works are not preserved by publishers after they go out of print. News agencies are not retaining old news, nor have they the mandate or resources to preserve it. Scholars are creating content in digital form that may be proprietary and may not be properly managed on a long-term basis. People are capturing history on their cell phones and creating it via social media, which is not only proprietary, but also extremely ephemeral; it disappears quickly if not captured. The Shakespeares of today are writing in Microsoft Word (a proprietary format) and know almost nothing about how to make their efforts last. Research data may be scattered across multiple computers and media, forgotten after the results are published. Later researchers seeking to repeat experiments for verification or expansion, or original researchers seeking to build on their previous work, may be unable to do so. More and more, universities are offering digital forensic services to assist faculty and students in recapturing access to older digital files; however, this may have no relevance for more personal content, which in the past may have made its way into special collections and archives. Personal correspondence, records, and photographs from the past have always become the fertile field for research of the future, particularly in the humanities and social sciences. What happens when this content no longer exists, is not channeled to those who can manage it, or is in forms that cannot be effectively curated? The digital content beginning to pour into special collections poses a series of crucial challenges, for which most institutions are simply not equipped. As databases, websites, cultural and historic information, and scholarly works disappear, we face an incredible amount of loss.

WHAT DOES IT MEAN TO CURATE DATA?

Digital curation has been defined as the "active management and enhancement of digital information assets for current and future use."[10] In other words, digital curation is a multifaceted *effort* to ensure both current and long-term access to and use of digital content. "Access" means to be able to open the file; "use" means that the content must be provided in a form that makes sense to the user. It is insufficient, for example, to be able to render accurately the contents of a database, without clarification of the definitions of each field. If you see a number but have no context for what that number means or refers to, it is useless to you. Similarly, if we retain a file and cannot even open it with current hardware and software, or haven't enough context to even understand what is stored, the preservation has been simply a waste of resources.

Vint Cerf, known as the "father of the internet," has been quoted as saying, "I'm very concerned that digital content will be less and less accessible, not because we can't find the bits, but because we don't know what the bits mean."[11] (By "bits," he is referring to binary digits, the smallest unit of data in a computer.) No matter how well the content is stored throughout time and monitored, if it cannot be rendered meaningfully and made usable, there really is no point in having made the effort. Thus, digital curation goes far beyond storage. For instance, it includes periodic reviews to determine whether, in fact, the content should continue to be monitored, managed, and preserved, and how best to do so. Digital curation may begin with the act of creation and extends through the life cycle management of materials. This involves appraisal and selection; active involvement throughout time to ensure usability, authenticity, and accessibility; provision of access; migration to new formats and media; and even destruction of content when it is no longer of value.[12]

Notice the word "effort" in this statement: "Digital curation is a multifaceted *effort* to ensure both current and long-term access to and use of digital content." No one can accurately foresee the hardware, software, systems, and file formats of the future. Consequently, successful digital curation is heavily dependent upon staying abreast of discoveries, developments, the status of file formats, and agreements among leading institutions within the digital preservation community. It is not possible for an isolated manager to effectively manage digital content for future access. Selection of file formats, emulation methods, storage systems, workflows, and tools will continue to evolve, and new efforts, formats, and other choices are most likely to survive and continue to be supported if embraced by the wider digital preservation and curation community.

CONSTANT CHANGE AND LOSS IN THE DIGITAL REALM

How old is our digital history? Computer punch cards, generated to store digital information, were first used in 1890; floppy disks were introduced in 1976; and the first commercially successful word processor hit the market in 1979.[13] (A fascinating overview of digital technology and preservation is provided in the Digital Preservation Management time line, which covers the basics from 1881 to 2013.[14]) Users often stored early word processing files in proprietary (licensed, patented) formats (like Word Perfect) on local attached media, for instance, tape or floppy disks, and later on diskettes or memory cards (see figure 1.1). With the advent of personal computers, word processing software replaced word processors, but the forms of the files they generated continued to change.

Even operating systems update frequently. For example, from 2000 to 2015, Microsoft Windows released more than ten versions of their operating system for personal use.[15] Almost every time the operating system changes, users must obtain new versions of software as well, from word processing to image display to antivirus packages. To compound the problem, current software packages often stop supporting older versions of files, with no notification to users. And files that were not perfectly formed by the originating software are the first to fail to render as new versions come onto the scene. For instance, users can generate PDF (Portable Document Format) files in many ways, including via browser extensions and online services. Each file format has specifications as to what information should be stored where and in what form, just as MARC21 defines what type of information should go into each cataloging field and RDA (Resource Description and Access)[16] or AACR2 (Anglo-American Cataloging Rules, 2nd Edition, Revised)[17] specify how that information should look. For archivists, the EAD (Encoded Archival Description)[18] specification determines what type of information is allowed in each finding aid field, and DACS (Describing Archives: A Content Standard)[19] describes how that information should look. In the same way, an image format specification describes exactly what information must be stored where in the file and how so that software can render access to the file properly. Files that are not properly encoded according to their format specifications will be the first to fail to render correctly as software evolves, since backward compatibility seeks to conform to known specifications. What does all this mean for the digital curator?

BASIC CONCEPTS UNDERLYING LONG-TERM ACCESS

Digital curation can be a fairly complex challenge, but there are some basic principles underlying the tasks. Some of the basic concepts managers need to understand when undertaking digital curation include authenticity and provenance, accessibility and usability versus obsolescence, the three primary methods of ensuring access to older content, and standardization of file formats and the multiple types of information about content. Standardization is key to managing complexity and leveraging the community to ensure success.

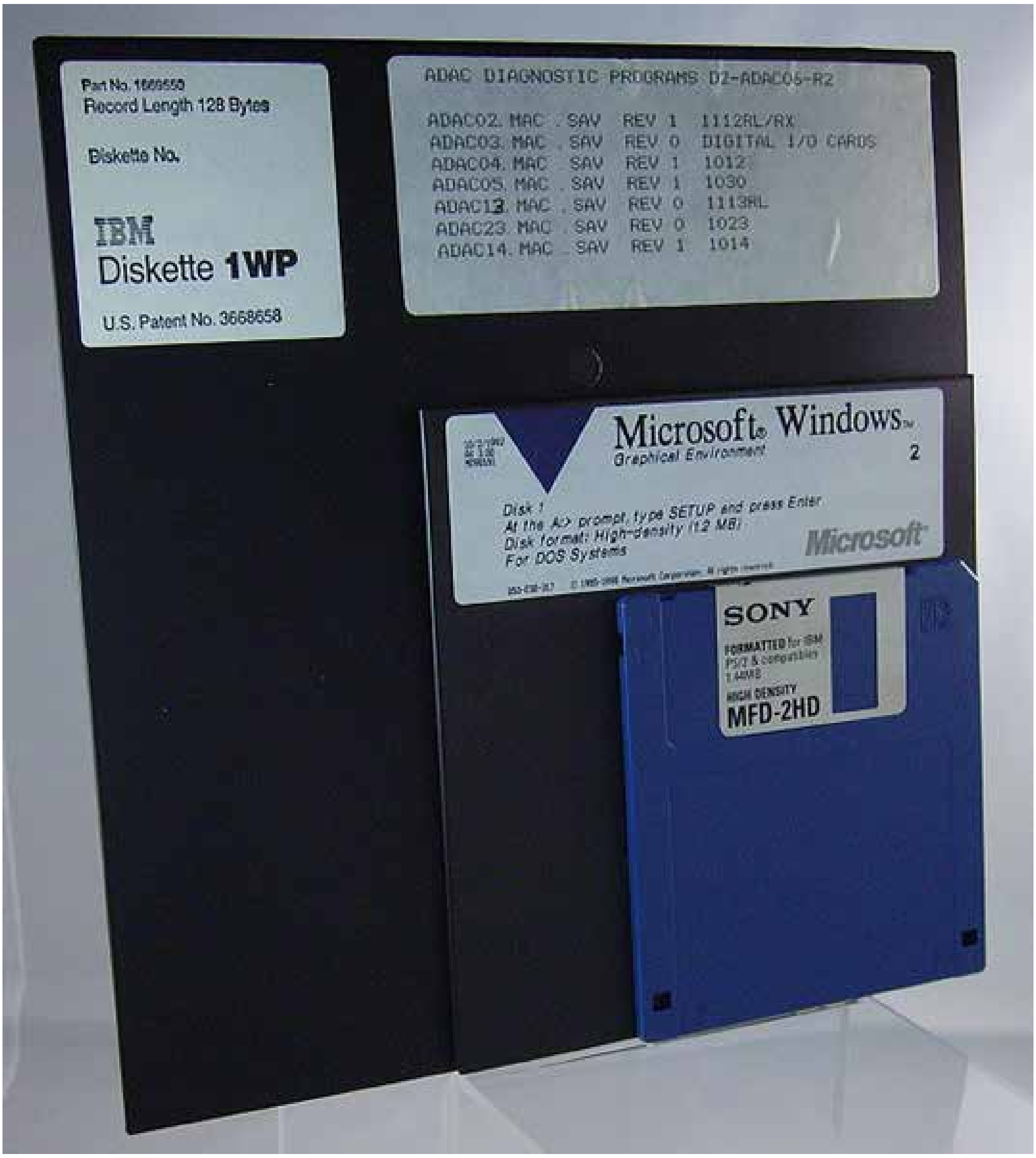
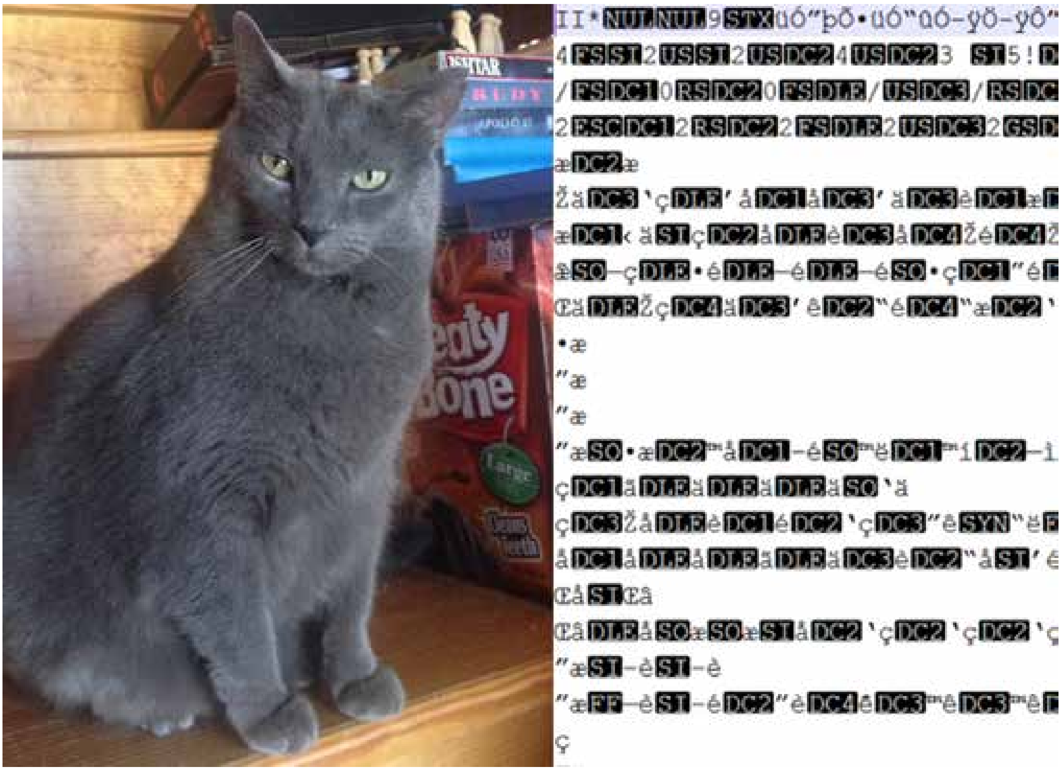

Part No. 1669550
Record Length 128 Bytes
Diskette No.
IBM
Diskette 1WP
U.S. Patent No. 3668658
ADAC DIAGNOSTIC PROGRAMS D2-ADAC05-R2
ADAC02. MAC . SAV REV 1 1112RL/RX
ADAC03. MAC . SAV REV 0 DIGITAL I/O CARDS
ADAC04. MAC . SAV REV 1 1012
ADAC05. MAC . SAV REV 1 1030
ADAC13. MAC . SAV REV 0 1113RL
ADAC23. MAC . SAV REV 0 1023
ADAC14. MAC . SAV REV 1 1014
Microsoft® Windows™
Graphical Environment
2
Disk 1
At the A> prompt, type SETUP and press Enter
Disk format: High-density (1.2 MB)
For DOS Systems
Microsoft
SONY
FORMATTED for IBM PS/2 & compatibles 1.44MB
HIGH DENSITY
MFD-2HD

Figure 1.1. Obsolete Floppy Disks. ***"Floppy disk comparison" (http://www.obsoletemedia.org/8-inch-floppy-disk/floppy-disk-comparison/), by Jason Curtis, is licensed under CC BY 3.0 (https://creativecommons.org/licenses/by/3.0/).***

If you were to be presented with a digital file today that was originally created by John Lennon on the day before he died (in 1980), what would be most important to you? You'd want to know whether this file was authentic: Was it really created by John Lennon? Is it the original or a copy? Has this file been changed or modified in any way? These are questions of "authenticity" and "provenance." In the digital world, it is simple to make duplicates and versions of almost any file. Forgery may be undetected, and authenticity is difficult to prove. Once proven, it must be protected, and this can only be managed via an "unbroken chain of custody from the creating agency to the archives."[20] Such custody and protection require regular verification that the file has not changed and documentation of "provenance." Provenance in the digital realm is the chain of events and people that have owned or touched this document throughout time. This includes documentation of any transitions that have impacted the contents, for instance, migration to newer formats. Only by documenting provenance and authenticity is it possible to verify for users that they are indeed encountering the original document or a version of it that has only been modified to the extent necessary to provide access in today's computing environment.

Once you are assured of the validity of the claim that this file was indeed created by John Lennon and that all handling of the file since then has been documented, you'd want to be able to access the file. If you can't open it on current hardware and software, that's a problem: This is the issue of "accessibility." Digital curation includes the work that will enable continued access to the content throughout time, regardless of changes in hardware and software. And when you open the file, the contents need to be rendered in a form that is meaningful to you. For example, while you can open an image file with a text editor, the contents are not rendered in a form that makes sense to humans (see figure 1.2). This is what is meant by "usability." If the file can no longer be opened in the current computing environment in a form that is usable, this is described as "obsolescence." Obsolescence can be created by changes in hardware (media) or changes in software. Cassette tapes, zip disks, VHS videotapes, and floppy disks are now obsolete due to media changes,[21] and most file types used by software packages that are no longer in existence are now obsolete due to changes in software. To prevent obsolescence



```
II*NULNUL9STXuó"þõ•uó"uó-ÿõ-ÿõ"
4FSST2USST2USDC24USDC23 ST5!D
/FSDC10RSDC20FSDLE/USDC3/RSDC
2ESCDC12RSDC22FSDLE2USDC32GSD
æDC2æ
ŽäDC3'çDLE'åDC1åDC3'äDC3èDC1æD
æDC1‹äSTçDC2åDLEèDC3åDC4ŽèDC4Ž
æSO–çDLE•èDLE–èDLE–èSO•çDC1"èD
ŒäDLEŽçDC4äDC3'èDC2"èDC4"æDC2'
•æ
"æ
"æ
"æSO•æDC2™åDC1–èSO™ëDC1™ìDC2–ì
çDC1äDLEäDLEäDLEäSO'ä
çDC3ŽåDLEèDC1èDC2'çDC3"èSYN"ëF
åDC1åDLEåDLEäDLEäDC3èDC2"åST'è
ŒåSTŒä
ŒäDLEåSOæSOæSTåDC2'çDC2'çDC2'ç
"æST–èST–è
"æFF–èST–èDC2"èDC4èDC3™èD
ç
```

Figure 1.2. Cat Photograph Rendered in Image Viewer (on Left) and Text Editor (on Right). ***"Baby Blue" image and text, copyright Jody DeRidder, 2017.***

in current holdings, managers must test files for viability and monitor them throughout time. Whenever operating systems update or crucial software packages (e.g., Microsoft Word) offer new versions, managers should compare stored file formats and versions to those accessible via the new hardware and software.

There are three basic approaches to providing access to older content, and each requires slightly different preparation. We must begin by establishing a foundation for the original content, from which current access methods can be derived throughout time. That foundation may be the capture and documentation of the software and system properties necessary for rendering the files effectively ("encapsulation"); or it may be the development and maintenance of either original or imitation software and hardware to try to render the content as it first appeared ("emulation"); or it may be the "migration" of files to currently supported archival file types, which is then repeated throughout time. This usually begins with "normalization," which is the transfer of content to selected standard formats, so curators have fewer file types to manage on a long-term basis. Managers may choose to use one approach with certain forms of content and another approach with other types of materials, based on the costs, availability of software solutions, and available expertise.

When faced with a deluge of digital content, the only practical solution is to find or create software to handle most of the work; however, that is impossible without standardization of the content. Standardization is what enables us to build software to automate work that extends far beyond what we can do by hand with the resources at our disposal. There are multiple aspects to standardization, which answer a series of questions, for instance, "what," "why," "where," "when," "who," and "how" for content managers.

We need standardization of archival file formats to ensure the continued ability to render the content. "Archival formats" are those that have been identified by the preservation community as most likely to be accessible and usable on a long-term basis and retain the content in the best quality available. The fewer the number of types of formats we need to sustain and migrate on a long-term basis, the more likely we are to be successful. No one can effectively manage hundreds or thousands of different file types unless they have unlimited resources. Furthermore, by agreeing as a community to support selected archival formats for each type of material, we build the user base that will ensure continued appropriate migration paths, as well as the tools and workflow development that will be necessary to implement those migrations. This type of standardization will help archivists of the future answer the question, What kind of material is this?

We also need standardization of the various kinds of metadata: "Metadata" is simply information about something. For example, the title of this book is a form of metadata; so is the fact that this book has nine chapters. Descriptive metadata describes what something is "about" (e.g., the title of a book); structural metadata tells us how the object is composed of parts (e.g., the order of pages in the book and the number of chapters). Technical metadata tells us about the file formats used to store the book, the size of the digital content, the creation date, and more. And administrative metadata (e.g., the copyright on the book) tells us what we need to know to manage the content effectively.

Descriptive metadata provides context and answers the questions, Why is this important? What is it about? Standardization is important here, so that huge quantities of stored materials can be effectively searched. Experience with search engines drives home the need for the ability to index together the content of fields that mean the same thing. Development of portals, for instance, OAIster[22] (now incorporated into WorldCat[23]) and discovery search interfaces of library materials, has demonstrated some of the uselessness of search results if there is confusion among contributors about how particular fields should be used.[24] For example, if the "date" field is used for the date published, the date created, the date the record is created, and the date the material was digitized, user searching with that date value will be confused by the search results. If, furthermore, dates are formatted in multiple ways, users will be unable to retrieve all the content that actually has the same date. Example formatting for dates can range from "Monday, March 7, 2016" to "2016-03-07" to "circa 2016," "7 March 2016," "approximately the 7th of March in 2016," and so on. Without standardization of field meaning and field content, effective search and retrieval is almost impossible. And if we do not use standard methods of documenting structure, software delivery systems will not be able to reconstruct the content in a meaningful way. What use is a book with its pages out of order?

We need standardization of rights metadata to identify what is now out of copyright, what is still restricted, what is accessible only to certain user groups, and more. If this information is clarified in standard forms that computers can use, software can automate the access and the restrictions. This is what is meant by saying that the information is "actionable." This type of standardization answers the question, Who can use this and in what circumstances? Rights are a form of administrative metadata, which includes the types of information necessary for effective content management.

We need standardization of technical metadata as well, so file formats approaching obsolescence can be located and migrated to newer formats. Imagine having digital content stored in many separate directories across multiple servers, hard drives, shared drives, and media. When the manager first realizes the latest operating system or software update has made it impossible to open a stored file type or version with current hardware and software, she needs to be able to quickly identify any of those files across existing holdings. Those files will need to be migrated to newer formats or emulation methods established and then maintained throughout time to ensure continued access. By standardizing the information gathered and storing it centrally in a database, the manager can effectively answer the question, Where do we have this content? Furthermore, by standardizing the information stored in public registries, we support the ability of software systems to access that information and answer the questions, When do I need to take action? and How can I access this content?

By developing agreed-upon standards, registries, and best practices, we help make digital curation more feasible and practical for everyone. By adhering to standards that are supported by the digital curation community and building sharable tools and workflows around those standards, we tip the odds toward success in ensuring long-term viability of valuable digital content. These standards enable us to develop software to facilitate the digital curation of tremendous quantities of files. Such software solutions are currently in development in many areas of digital curation, as the preservation community and industries struggle to save access to crucial information.

KEY POINTS

1. Digital curation is a multifaceted *effort* to ensure both current and long-term access to and use of digital content.
2. As hardware and software evolve, older files and content become obsolete and may no longer be accessible.

3. The development of new knowledge depends upon the ability to access previous knowledge and information.
4. Only by documenting provenance and authenticity is it possible to verify for users that they are indeed encountering the original document.
5. Successful digital curation depends, in part, upon adherence to the decisions of the broader preservation community.
6. Standardization is key to effective management and future access.

2

What Can I/We Do?

As you read through this book, think about what policies and procedures you and your organization have established in each area of digital curation. Compare what you are doing to what others have tested or established before you and look for the gaps that need to be filled. You may find that a type of content your organization wants to collect does not yet have established best practices for identification, selection, metadata, ingest, storage and protection, or provision of access. You may find there is a certain area of digital curation in which you want to lead. And you may find that your area or region has a real need for your leadership in establishing a service or a collaboration that will help others effectively address the challenges.

As an archivist, librarian, or administrator, you have the opportunity to establish yourself and your institution as leaders in the field and your area or region. This leadership will help you obtain funding, attract skilled and talented potential employees, and increase your standing among your peer institutions. By sharing what you have done, how you did it, and what worked and what didn't, you help establish directions and guidelines. Others can then test your processes and follow safely in your footsteps, establishing new best practices, or you can be the ones to test others' processes or help lay the groundwork for developing standards from existing best practices. By working together and communicating effectively, we leverage our shared resources, talents, and skills to better effect.

Archivists and librarians are the people on the front lines. You have the most direct contact with researchers, students, faculty, users, donors, and the community. You are most likely to be asked to develop instructional material, perform outreach, identify and select materials for curation, digitize materials, or extract digital content from media. This direct contact with digital content, users, and creators makes you crucial to the future of digital curation. You are also best positioned to perform research, become involved in community efforts and best practices development, and develop articles and presentations to share with others. You are the communicators who have some of the greatest impact on the development of the digital curation community. Your advocacy in the field, work in the institution, and participation in listservs and collaborative efforts is key to our success. Thus, some of the areas where librarians and archivists can take an active leadership role include the following:

instructing
selecting and performing intake

digitizing or capturing digital content
building web services or assisting in digital humanities projects
serving as participants and leaders in community and collaborative efforts
acting as advocates in the field

The following sections will briefly address each of these areas, followed by roles for administrators and for institutions.

INSTRUCTION

The instructional role is particularly important, as this is an area largely disregarded at present. And yet, from where will all future digital content come? Creators, of course. Most of what is pouring into special collections and archives today is far from being preservation-ready. The only way to address that effectively is to educate content creators. The most brilliant writers are creating their content in Word documents and saving it on flash drives. Gifted musicians are capturing their compositions via proprietary methods and in formats that may not be migratable. The scientists of today are capturing research data with insufficient metadata, on proprietary databases that will likely soon be obsolete. Digital humanities projects and digitization projects are developed and placed on the web and forgotten, with no thought to how the content needs to be managed for long-term access.

This form of digital literacy is a new level of education badly needed in our culture. By teaching patrons how best to choose formats, capture information, organize and name their files, embed descriptions, or store content so it can be accessible and usable in the future, you will make an incredible difference: You impact what will remain, versus what will be lost. Some basic aspects of this education follow.

To begin with, advise content creators to use open-source software, rather than proprietary software, when creating their materials. OpenOffice, for example, is much preferred to Word or Excel. PDFs are another problem. A PDF can be created by a variety of software—and in many different forms. These PDFs may not comply with the standard, refer to external fonts or external resources, have recursive references, or contain scripts, all of which interfere with migration to archival formats. Research has shown that we cannot always successfully transform PDFs to PDF/A (the archival version of PDF), so content creators should be advised to generate a PDF/A version of their work to be stored on a long-term basis (preferably PDF/A-1a[1] or PDF/A-1b[2]) and embed descriptive and rights information. Then at least we are likely to be able to provide long-term access to some version of what they have created.

Databases should be exported in open formats for long-term storage; these may be simple CSV exports or complex XML exports with associated schemas. Content should be in ASCII text (American keyboard characters) or Unicode. CSV or XML exports should be accompanied by data dictionaries to describe the meaning of each field and how the fields in the tables relate to one another. As often as software changes, this is the current best-practice method of ensuring that the database itself can be reconstructed for use in the future. Any kind of digital content should be captured in uncompressed open formats, using the highest quality representation available, and with careful descriptive, structural, and rights information (more about this in chapter 6). Whatever the type of content, documentation is crucial to maintain the provenance of the content, ensure authenticity, capture descriptive information that will enable us to manage content throughout time and make it accessible, and capture the rights restrictions associated with the content.

Storage and protection of content is key as well. If the materials being created are within the scope for your organization to collect and store, two to six copies of valuable content should be kept in multiple geographic locations, with the associated metadata; however, most content creators must rely on their own resources. Advise creators that flash drives, CDs, and DVDs are not viable storage devices. External hard drives are better but are still subject to crashing, theft, damage, and loss. Encrypted cloud storage options may be the best option, but admonish creators to read the fine print and be certain they can retrieve what they store, no matter what happens (organizations lose content and fail, of course). Since no storage method is perfect and all are subject to bit loss throughout time, checksums should be captured prior to storage and the stored copies checked periodically against the checksum, so that damaged copies can be replaced with good copies while they are still available (chapters 6 and 7 discuss these issues further). Become familiar with checksum tools and teach creators how to use them as well and when to do so. Help them set reminders in their calendars and record where they have the most valuable content stored. Make sure the document that records this information is located somewhere it is easy to find and reference but also that it is backed up with the other documents.

To reach more content creators, you need to develop online instructional and reference documents that refer to such sites as the one the Library of Congress hosts for personal archiving[3] and the DPC Technology Watch paper.[4] An example site for content creators was developed by the author for the University of Alabama,[5] and a great one for agencies and local governments was developed by the North Carolina State Archives.[6] This is an extension of the "libguide" concept: Many of your patrons never receive personal instruction, so these online instructional and guidance tools are crucial. Be sure to educate creators on the use of rights metadata (see chapter 6), and encourage them to adopt and incorporate Creative Commons statements[7] into their works (see figure 2.1).

SELECTION AND INTAKE

If you are involved in selection and intake, you will need to familiarize yourself with not only how to identify what is of value (appraisal; see chapter 5), but also what is within the scope of your institution to curate, preserve, and make accessible. You will be the one to ask about and record the preservation, storage, migration, access, and usage rights. If the creators are not present, you will need to perform the research to try to establish this information, for without it, you cannot legally retain, preserve, and provide access to the content.

To obtain this information, begin by first trying to ascertain the creator. For content that arrives in physical media, of course you would examine the packaging for clues. The immediate source of the content may have information, whether that is a donor or a website owner. For the latter, look for an "about" or "contact" page for someone to consult. Failing that, view the source of the online site pages to see if any names or contact information are embedded there; if so, it is usually located near the top of the page in comments or embedded metadata and occasionally at the bottom. If schema.org markup is used throughout the page, however, it could be anywhere; use "control-F" (find) on your keyboard to look for "Person" markup or "contactPoint." Another option is to look up the website name and find out to whom it is

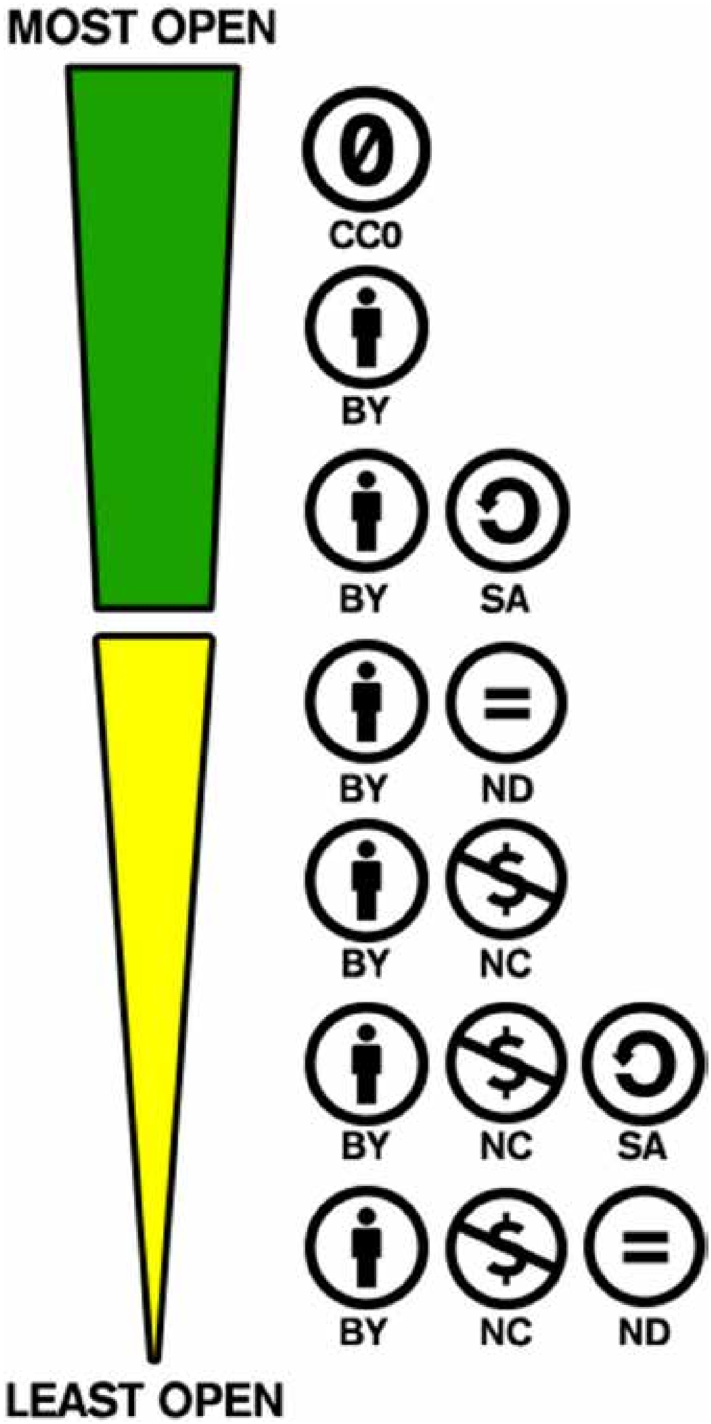


Figure 2.1. Levels of Rights Available in Creative Commons. *"Understanding Free Cultural Works" (http://creativecommons.org/share-your-work/public-domain/freeworks/), by Creative Commons, is licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).*

registered and how to contact them.[8] This is not effective if the owner has paid for privacy, however, since all you would find here is the company that provides the server support.

If no information is yet forthcoming, look for clues in the embedded information. For example, Microsoft Word automatically captures the name in which the software was registered and stores it in the file properties. If nothing there is useful, you may need to resort to web searches to locate other copies that may have more context or references to the content in hand that can provide you with more information. Locating content creators and obtaining permissions is nothing new for archivists, who have struggled with this task for many years; if you are a librarian, you may want to consult with archivists on how best to approach thorny problems. Again, if you are unable to obtain the rights to effectively manage the resources, you run some legal risk in doing so.

Rights metadata is not the only information you need to gather and document. To provide effective search and retrieval, and indeed establish the value of the content itself, you will need to look for, ask for, or create descriptive metadata as well. Key information includes the creator, the date of creation, location (particularly if it is crucial in understanding the content), topics, type of material (text, image, audiovisual, software, etc.), current file types, and relationships between files. These last go beyond descriptive information into administrative information you will need to manage the content effectively. For example, if the material cannot be fully understood when separated by component or surrounding files, the relationships between them is key to both preservation and future use.

The file types are important as well. As part of the contingent on the front lines, you must become familiar with which kinds of media you can no longer access and which kinds of files you simply cannot migrate or make usable. If your institution does not have or cannot obtain the resources to access content on older media (e.g., floppy disks, zip disks, and a host of more arcane media), you must redirect donors or content to other institutions, or, sadly, turn it away. You also need to become familiar with the technical capabilities of the support staff available to you and seek to expand their skills and tools to effectively manage as many different file types as possible, including those that are obsolete. Being on the front lines, you must put policies into action: You are where the rubber meets the road. There is, perhaps, no role more actively involved in digital curation than this one.

DIGITIZATION AND CAPTURE

Are you involved in digitization or capturing digital content from the web? Both are "capturing" activities, and, as such, you will want to understand the challenges of those who are performing selection and intake, as you must cover much of the same functions, to a more limited extent. If you are digitizing, you are the digital content creator, and you are responsible for capturing the most accurate representations possible in archivally sound digital file formats. You cannot trust your software to generate those files in a way that meets the specifications of the file format but that is necessary for effective digital curation and reuse. You must learn how to validate the file formats, regenerate them if necessary, and best capture and store metadata with the content (chapter 6). You may also need to migrate materials to newer formats (chapter 4), test their viability, and report on successes and failures.

For those of you capturing web content, it is important to familiarize yourself with the archival formats for this and the methods to describe, store, and provide access (chapters 4 and 8) to web content. In both cases, you will want to coordinate with the community to ensure that what you are capturing is unique, and not duplicative, to best leverage resources (chapter 9). It is likely that you will be most responsible for preparing content for long-term preservation, overseeing the digital curation process and implementation, and preparing content for reuse. Carefully study the methods and procedures that ensure long-term access (chapter 7) and master the technical metadata (chapter 6). Without your attention to these details, digital curation cannot succeed. You have the power to make or break the best of plans: Upon you rests the responsibility to manage these resources effectively. Your dedication and conscientiousness may make all the difference in ensuring that future users will be able to access and use this valuable content.

WEB SERVICES AND DIGITAL HUMANITIES

Those of you who are building web services or assisting in digital humanities projects are uniquely positioned to educate colleagues, faculty, students, and administrators about the best practices in usage of standards, accessibility, rights and access management, archival formats, and how best to store the content you are creating in a form that will withstand the tests of time. Most administrators and web developers are not mindful of the impact of time on web content; you need to be the exception and serve as a guiding light. We must move past the mind-set of "put it online and we're done"—this is not, and has never been, true.

As hardware and software change, as servers are upgraded or transitioned, web content often disappears or becomes forgotten or unusable in today's environment. Older web pages and databases are no longer secure, nor are they suitable for mobile browsers of diverse sizes. Older web software is hazardous if not maintained and transitioned as new security risks evolve. Now more than ever, it is imperative to capture valuable content before it disappears and store it and make it accessible in safe and useful ways, without security risks.

Track the websites and software within your domain. How old are they? What are the dependencies? What kinds and versions of software and databases were used? Who created it, and who is responsible for maintenance, if anyone? What types of files are linked in, and how old are the versions? Collecting this information will require some investigation on your part, as there are likely multiple sites and even systems in service at any one time. As new systems and software are developed, and staff turnover occurs, update this document, and ensure that multiple others can update it, too. Having an easily accessible document or database storing this information will inform you rather quickly when something is outdated (and likely a security risk, e.g., when it is dependent upon an older version of software or database). This information will also serve you well when updating the servers, as you'll know which systems will be impacted by recent changes in the underlying software libraries or the server itself.

If the file types or versions are becoming obsolete, arrange for migration to newer formats. If the databases are proprietary, export the content and load it into open-source databases and keep those current. If you are concerned about long-term access, you will want to export and collect snapshots of the content in archivally sound formats with appropriate metadata (chapters 6 and 7), prepare them for curation, and ensure copies are placed in secure storage. For those of you working with developing new websites and new functionality, advise creators on file formats, metadata creation, and software selection. Document as much as you can about the rights associated with the content being developed, as well as the decisions made in software development and implementation. This will be important for ongoing management and curation.

RESEARCHERS, PARTICIPANTS, LEADERS, AND ADVOCATES

As your understanding of digital curation issues deepens and your experience grows, your participation in community and collaborative efforts becomes more and more important. Your leadership and advocacy in the field will serve to educate and motivate others. As you try new tools and test methods for digital curation, document your findings and share them in conference presentations, articles, and listserv discussions. For example, the Code4Lib Journal[9] supports practitioners in sharing research projects about tools and discoveries. The Digital Library Federation (DLF)[10] and the Society of American Archivists[11] regularly invite presentation proposals by practitioners on what they are learning and developing in the field. When you see gaps in what has been established, engage your colleagues in exploring options and proposing new guidelines and best practices; this is often begun in professional organization sections and groups, but in smaller organizations, it may begin with presentations and grow organically, which is how DLF interest groups are formed.[12]

As you advance in the field, offer to serve on standards development working groups, for instance, the National Information Standards Organization (NISO),[13] which lay the groundwork for others throughout the country and the world for years to come. Your expertise and hands-on experience is invaluable for informing administrators, institutions, and policy development at every level. Never underestimate the importance of your voice, as we must learn from one another in this developing field. Gather data, and be prepared to justify and demonstrate your findings and recommendations. This is practical research that will help your career grow, while furthering the profession and our joint success.

ROLES FOR ADMINISTRATORS

Administrators play a strong guiding role in establishing and supporting directions for institutions. They not only lay the groundwork for effective programs, but also determine the extent of those programs, the overall policies and procedures, development of positions, and further education for existing staff. Administrators may seek funding both within and outside the organization. In establishing support, administrators develop cross-departmental and cross-institutional collaborations, and forge and support cross-institutional (and sometimes international) collaborations, which extend capabilities beyond available resources.

Administrators are well-positioned to advocate for funding and incorporate digital curation needs and priorities in organizational missions and strategic plans. This will ensure funding and support, which is necessary for sustainability. As an administrator, you should ask yourself the following questions: Are the digital records of your organization captured and managed effectively? Are the databases of content that must be legally maintained indefinitely curated and managed by digital curation professionals? Or is the content merely stored in proprietary formats, in proprietary databases, with administrators blithely expecting that content to be accessible and usable when it is needed, possibly dozens of years in the future?

The scope of what is to be digitally curated at an institution should be incorporated into the mission statement and the strategic plan. If your institution cannot effectively manage curation of software or video, for example, that should not be part of your mission. If, however, you determine that some variety of content is crucial to your mission, consider what the University of Indiana did in 2013, when they committed to a massive group effort and dedicated funding to extract and curate valuable audio and video content from obsolete and failing media,[14] and, in one fell swoop, demonstrated leadership that drew international attention. As Indiana's effort exemplifies, the funding to support digital curation must be established at the institutional level to ensure longevity and business continuity. If your institution has valuable content at risk of obsolescence, has invested thousands of dollars into digitization and web delivery, hosts digital humanities projects for faculty, or collects e-resources, it is in your best interests to protect those investments.

As an administrator, you can also negotiate with colleagues to develop collaborations. For example, you may consider developing low-cost digital preservation coalitions within your region; the successful Alabama Digital Preservation Network[15] began with a single conversation among administrators about shared concerns. Even better, build on these coalitions to arrange for business continuity agreements to ensure that valuable digital content will continue should you lose funding or if your organization fails.

As an administrator, you may apply for (or support) applications for grant-funded efforts that can help with preparation preservation, develop new models to solve existing problems, or establish collaborative efforts with other organizations. Major funders in the United States for these types of work include CLIR,[16] NEH,[17] NHPRC,[18] Mellon,[19] and IMLS.[20] Engage with others cross-institutionally to develop new models and platforms that will take digital curation efforts to better and more sustainable levels. Venues for this type of connection include such groups as the Coalition for Networked Information membership meetings.[21]

Administrators are empowered to establish policies, oversee the development of procedures, and select and hire the best managers and most skilled and talented employees. By setting the direction; communicating the vision; sharing your knowledge; and engaging your employees, peers, and others you can establish and support major changes in the field and the viability of our valuable digital content. Be sure to create appropriate job descriptions, support employee development, and set aside the funds necessary to ensure long-term access and effective content management. Arrange for cross-departmental digital curation advisory groups to build internal participation and support. Seek out role models and examples from institutions you would like to emulate; meet with faculty groups to determine their needs.

You may find that you want to establish and staff forensics laboratories so researchers can extract and access older content from media that has become obsolete. You will need to establish (or update) digital preservation and curation policies, and guarantee that digitization and digital humanities projects are appropriately safeguarded and prepared for long-term access. It falls on your shoulders to ensure that incoming digital content in special collections and archives receives appropriate attention and treatment, and that the collection policies reflect both what you can support for digital curation, as well as the mission statement of your organization.

As an administrator, you are responsible for advocating for necessary changes in your organization or institution. An example of this would be to engage other administrators to ensure the digital form of your institution's memory is captured, described, effectively stored, and continually accessible to meet legal and other demands. Institutional archives are a pressing issue at the time of this writing, as many forms of content that must be legally managed for long-term access are basically being ignored. Your leadership is key to future access and use of digital content created yesterday, today, and tomorrow. Without your support and guidance, even the most excellent staff cannot succeed in addressing these issues.

INSTITUTION ROLES

Have you ever noticed how some institutions lead the way, while others look to them for guidance? One of the long-standing approaches to dealing with change in libraries and archives is to seek out how other institutions have managed that same type of change. No one wants to be on the "bleeding edge" for long; it's too expensive. Only the most well-funded institutions can afford that level of research and development, and even they normally engage in collaborative efforts with other institutions to develop models and platforms—often with grant funding. But there is no denying that institutions that take a leadership role gain credibility in the sight of others, and this can pay off in terms of donor and grant funding.

Institutions build and maintain archives—of their own historical content, and often content from the community or areas of crucial interest to their constituency. The effectiveness with which these archives and collections are managed reflects on the standing of the institution and often the quality and prestige of the researchers and faculty the institution attracts. This provides a level of sustainability that is key to digital curation, particularly in the United States, which does not yet have dedicated support for this at the national level.

Collaborations are often established at the institution level, for example, the Big Ten Academic Alliance,[22] the Orbis Cascade Alliance,[23] and the Digital POWRR (Preserving (digital) Objects with Restricted Resources)[24] effort (several digital curation collaboratives are discussed in chapter 9). Institutions become known for their participation in collaborations, standards development, and international outreach.

Perhaps even more important to success is the "inreach" performed at institutions. As funding and directions change, the effort of engaging and involving employees, target audiences, administrators, and donors becomes an ongoing challenge. Effective multidirectional communications, staff support and involvement, and continued access to the necessary resources are paramount to maintaining morale in a continually shifting environment. Institutional success is dependent upon developing and maintaining a solid foundation. Employees must be aware of and involved with new developments in the field, and with the employees, donors, and target audiences that make them a success. The sustainability of your institution is key to the sustainability of your digital content. A collaborative effort of employees and administrators is necessary to keep your institution both effective and healthy. If you help your institution become a leader in the field, not only will your institution benefit, but also you and the future of digital curation.

COMMUNITY (ALL OF US!)

Digital preservation challenges are born of our changing culture and our very real responses to those challenges. Formats, software, and media change constantly, and the extent to which we have been managing valuable content across those many changes determines just how much work we still need to do. There are steps we can take together to establish safety nets that will help us move forward. For example, we can collaborate with other institutions to store copies of one another's digital content. We can determine the best way to protect the original objects, while working out the best way to transform or migrate copies. We can learn from one another's explorations, research, and testing to avoid making the same mistakes and duplicating coverage.

Make use of existing networks to share information, establish collaborations, and learn from others so that we do not all make the same mistakes. Work closely with colleagues to develop best practices and then to turn those into standards, as this creates a structure that supports both current and future access to valuable digital content. Some of the organizations that support developing best practices are the Society of American Archivists (SAA),[25] the Digital Library Federation (DLF),[26] the National Digital Stewardship Alliance (NDSA),[27] the Best Practices Exchange (BPE),[28] and the Digital Curation Centre (DCC) community[29] (more are discussed in chapter 9).

To translate best practices into standards, we must also become involved in working groups for such organizations as the National Information Standards Organization (NISO),[30] the National Institute of Standards and Technology (NIST),[31] and the International Organization for Standardization (ISO).[32] Begin by learning from listservs, publications, and conferences. Then start developing your own policies and procedures, testing out the best methods, and share what you learn, again via listservs, publications, and conferences. Work with others to develop new models and better methods of support. Build your "inreach" and your outreach; establish new directions and deepen your community involvement to sustain them. Bring issues to the attention of lawmakers and politicians: Lobby for change. When you can find no other way to protect valuable digital content on a long-term basis, offer it to the Internet Archive.[33] By working together, we improve our chances of success and the scope of valuable digital content that will survive into the future.

KEY POINTS

1. Instruction of content creators is crucial to ensure that valuable material being created today and tomorrow will be available and in a form that can be curated for long-term access.
2. Those involved in selection and intake need to be aware of the tools and skills needed, the media and formats their institution can effectively handle, and the parameters for selecting valuable content, including how to locate and define the necessary rights information.
3. Those performing digitization and capture must select the best archival formats and representations possible, collect necessary information in standardized forms, and verify that the captured content meets format specifications and is stored and monitored properly.
4. Those who are building websites and delivery platforms need to be mindful of the needs of sustainability: using open-source software, creating the necessary documentation (including of supporting libraries), selecting the best possible formats, and capturing snapshots for preservation.
5. Researchers should share their discoveries as they build on developments; participants can help develop best practices and standards, and leaders can establish working groups and collaborative efforts to better leverage resources.
6. Administrators should ensure that their mission statements encompass digital curation, develop inter- and cross-institutional collaborations, oversee policy and program development and hiring, and ensure that staff involved in digital curation receive ongoing training.
7. Institutions must offer support of digital curation, and inter-institutional collaborations are likely the most effective route for long-term sustainability of our efforts.
8. We can all better leverage our resources and discoveries by communicating effectively, learning from one another, and developing collaborations.

3

Models

Which Ones Do I Use When?

To the uninitiated, digital curation can seem like a vast and confusing wilderness—unknown territory. Those of us who venture into it face uncertainty, for we cannot foresee all the ways in which software and hardware will change in the future. We have limited resources with which to face these risks. In many ways, we are explorers entering largely unmapped wilderness. Still, we can learn from what others have discovered by examining the trails they have blazed and mapped out through portions of the territory. These maps are what we call models, and each one provides a slightly different perspective of the landscape. It is as if different explorers entered the wilderness, each climbing a different tree to get a high vantage point. Each explorer wanted to get a good sense of what is out there and how all the parts relate to one another.

We will explore three of these maps and look at how they overlap and how they differ. Just as looking at three different panoramas of the same territory can provide a better sense of the landscape, the different perspectives of these maps will help clarify what's important in digital curation. Knowing how they differ and why will help you to select the model, or map, that you need to guide your own work.

THE DCC CURATION LIFECYCLE MODEL (DIGITAL CURATION CENTRE)

The Digital Curation Centre's Curation Lifecycle Model is perhaps the most comprehensive "map" of what is involved in digital curation, as it covers the entire life cycle of digital content. Everything in the model revolves around the digital content, from the moment it is conceived until it is deleted or transformed into new content. We'll begin with the outer ring of the diagram and work our way in to the center. At the top of the diagram (see figure 3.1) is the word "conceptualize"; this is where the person who has not yet even created the content is conceptualizing what they want to create and how they want to capture and store it.

Ideally, librarians and archivists would work with information creators to ensure that the content they develop is captured in archival formats, with good descriptive information and appropriate rights information, and stored and protected properly. When generating new digital content, creators may incorporate or transform digital content that already exists, for instance, including images in a PowerPoint presentation or a research article. When digital content is

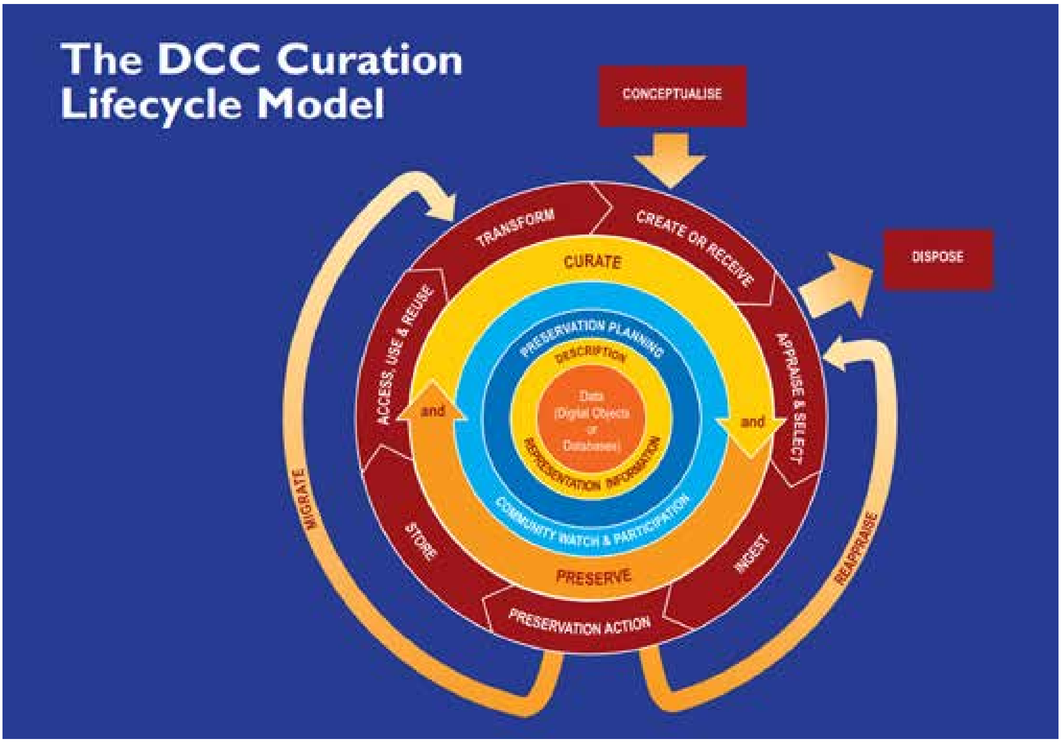


Figure 3.1. DCC Curation Lifecycle Model. ***Reprinted with permission of JISC Digital Curation Centre, copyright CC-By, available from*** **http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf.**

changed or incorporated into other digital materials, the result is something "new" to curate, from a digital preservation perspective. This is part of how the model covers the entire life cycle. The model also includes disposal of content that is not selected for long-term preservation or deselected during reappraisal. Throughout time, all content should be reappraised to determine whether it is still useful to the "designated community" and within the scope and ability of the institution to maintain. The "designated community" is the groups of users for whom the institution is selecting content. For example, the "designated community" for an academic research library may be the faculty and students at that institution.

If content is deleted, it leaves this life cycle model and no longer exists. This is happening now with many websites, for instance, and also with personal computers whose hard drives have crashed, without effective backups. Tremendous amounts of digital content are created and then disposed of before it can even be evaluated for digital curation. Finding ways to sift this content and make intelligent selections is one of the challenges we as a community need to address. If, on the other hand, someone appraises and selects the content, the material may then be "ingested," which in this model simply means it has been transferred to an archive or other custodian[1] (in other models, "ingest" means to upload the content into a preservation system). Once transferred, the new custodian will perform preservation actions (which "should ensure that the data remains authentic, reliable, and usable, while maintaining its integrity"[2]). Of course, the custodian then stores the content and makes it available for access and use.

That's the outer part of the diagram. In the second layer, the top half is labeled "Curate" and the bottom half is labeled "Preserve." In this model, curation incorporates appraisal, selection, obtaining content, and transformation. It goes hand in hand with preservation, which incorporates ingest, preservation actions, and storage. The management and administration of both of these aspects must be maintained continually throughout the curation life cycle.

The third layer, which is light blue, is labeled "Community Watch & Participation." The collaborative work of the digital preservation community to develop shared standards, tools, and software is crucial to our success. Within this layer is "Preservation Planning," which must surround digital content, as do "Description" and "Representation Information." The latter is any information that converts the raw data into something meaningful, for instance, the format of an image that makes it possible for software to render it properly.[3] In this model, "Description" covers all forms of metadata, which is key for content management, access, and use.[4] At the center of this diagram is the digital content itself.

The DCC Curation Lifecycle Model is content-centric: Every concept, every action, every viewpoint revolves around the digital content. This is the most complete "map" we have of the territory, from the perspective of the digital content itself. This model points out how important it is to engage with creators and participate in the preservation community, and how curation and preservation are central, ongoing processes to be managed.

OAIS (OPEN ARCHIVAL INFORMATION SYSTEM)

The Open Archival Information System (OAIS) was developed from the perspective of the digital repository, which manages the content from the point of ingestion to provision of access. This model, or map, does not attempt to cover the territory around the origination of content, the incorporation of some digital content into other digital content, nor does it cover the human tasks of reappraisal and destruction of content; however, this model is the most complete and most revered one for mapping out the basic functionality that must be incorporated into digital preservation systems. Therefore, the OAIS model is the one that almost every preservation system attempts to emulate and often claims to implement. Knowing this model will help you select, or develop, the best content management systems.

In the OAIS model, the focus begins once the producer, or creator, submits the content (see figure 3.2). For many of us, this happens when records are obtained in one form or another. While this model expects content to be prepared for submission by the producer, the truth is usually far from this ideal. Instead, most of us who manage digital content must generate most of the Submission Information Package, the SIP, for incoming content ourselves. This SIP should include descriptive information (e.g., title and creator), content information, and preservation information.[5] The content information consists of the content data object and the representation information, which clarifies how to make the data meaningful. The actual definition of these is dependent upon the repository and the material in question, but an example would be this: The content data object is an export from a database, and the representation information explains that this data is in comma-separated values (CSV), Unicode text, English language, and goes on to explain the fields in the database tables, how the fields in the tables are interdependent, the constraints on the fields, and so on. Without the representation information, it is not possible to make sense of the content data object. The preservation information includes reference information (e.g., assigned identifiers), context information (which explains why this content was created and how it relates to other content), provenance information (which tracks the history of this content and its custody), and fixity information (e.g., checksums, by which the content can be checked to verify it is unchanged).[6]

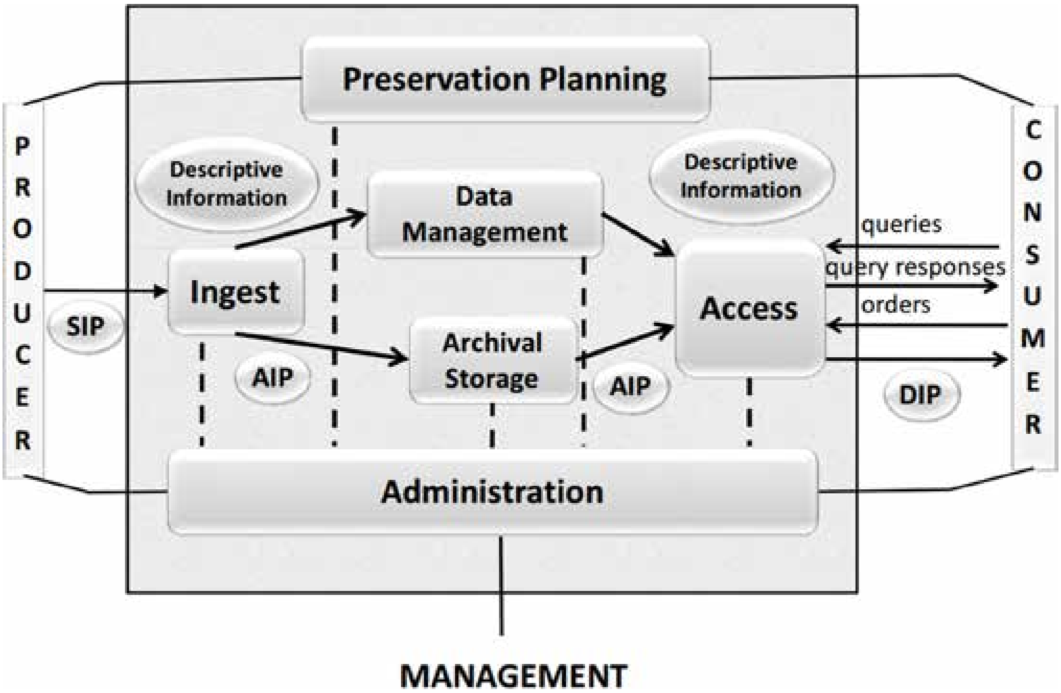


Figure 3.2. Open Archival Information System Functional Entities. ***Reprinted with permission of the Consultative Committee for Space Data Systems.***

Once this SIP is complete for an incoming information object, the OAIS (which is what a system is called if it implements this model) must then generate an Archival Information Package (AIP). This will contain the completed, and refined, standardized versions of the information from the SIP and also include information about this package, which can be used to reference this content from a finding aid or ordering aid (from which consumers can select the content they want). This descriptive information also goes into the administrative database for effective data management and handling of orders. (It is possible to combine multiple information objects within a single AIP for storage, which may be particularly useful for multipart content, e.g., research data from a large project.)

At this point, the information object goes into storage, where it is protected, duplicated, and transferred to newer media as hardware changes, and it undergoes fixity checks on a regular basis (whenever moved, and before backups, to ensure that good backups are not overwritten by bad ones). These verifications and any modifications to the content throughout time are recorded as part of the AIP, as this is provenance information, which helps ensure that users can be assured of the authenticity and completeness of the original.

When access to content is requested, the OAIS provides a Dissemination Information Package (DIP), which includes the submitted information (SIP) and may also include some of the information that was added for archiving (AIP).[7] More complex dissemination packages may provide subsets of the original content and combinations of multiple AIPs, or they may transform the original content data objects into versions of files (“derivatives”), which can be more readily used by consumers in the current day and age.[8] These requests and the management of this dissemination of content to users (“consumers”) is handled by the data management portion of this system. Notice that new or changed descriptive information may be added at the point of access, as the earlier descriptive information may no longer be sufficient in the current time period or for the current consumer.

At the top and bottom of figure 3.2 are “Preservation Planning” and “Administration.” This indicates that planning and administration of the work processes must encompass all aspects of the management of digital content. Administration will include establishing internal standards and policies, managing the system configuration, auditing content and possibly migrating it, updating the AIPs, and managing customer service and dissemination requests.[9] Preservation planning involves selecting the preservation and migration strategies, designing the package standards, monitoring changes in the preservation community standards and best practices, and monitoring technology changes.[10]

The OAIS can be viewed as having three different conceptual levels. On the top level, we have preservation planning, which encompasses every aspect of the system, and secondly, the administration of the workflows that have been planned. On the next conceptual level we have data management and archival storage of the content. And on the lowest, or most specific, level, we have the following:

submission of information, or receipt of it
ingestion of the content into the system
provision of access in some form

While it is certainly possible and ideal for a preservation system to perform many of these aspects of the OAIS model, it is not feasible for any one software system to fulfill all of it. Humans must monitor changes in the standards and best practices of the preservation community; humans must develop the preservation planning, the migration plans, and the package plans for the SIP, the AIP, and the DIP. Humans are required to audit migrated content to ensure that the result effectively represents the originals. Humans will refresh the hardware as needed (shift content from one type of storage to another and then verify that the transition succeeded). Humans will also select the software used for testing incoming content, migrating content, storing in a database, tracking information, and generating the information packages. It is commendable for digital preservation systems to emulate the OAIS model in as many respects as can software, but the human element is still very much a part of the OAIS model. Be aware that no software system can encompass every aspect described here, and that to a large degree, when you select a preservation software “solution,” you are entrusting most, if not all, of these decisions to other humans, or portions of this model simply are not included.

DPOE (DIGITAL PRESERVATION OUTREACH AND EDUCATION) MODULES

So let’s add one more map (or model) into the mix. Again, think of these maps as having been developed by explorers clinging to different treetops, and hence having different views of the landscape. The Digital Preservation Outreach and Education (DPOE) train-the-trainer effort was designed to provide continuing education to working individuals and organizations, and increase the capacity for effective long-term management of digital content.[11] Education is ideally provided via hands-on workshops, and representatives from regional areas are trained and, in turn, expected to train others. The DPOE modules consist of six topics,[12] which provide an overview of every aspect of digital preservation.

In the DPOE model, the modules include identify, select, store, protect, manage, and provide. Identify and select are in the center (see figure 3.3); everything else builds on them. Since effective preservation planning requires a knowledge of the scope of content to be preserved, identification and selection are crucial to the next steps. It would be impossible to determine the extent of resources needed for digital curation if the scope of content is not determined at the outset. Note that each of these processes (or steps) repeat, over and over, as time goes on, and new material comes in. (Most of the information in this section is taken from a three-part webinar series provided by the author on the topic of the DPOE modules, for the Association of Southeastern Research Libraries.[13])

Identification generally begins with building an inventory to collect information about the content that might be considered for curation. This content may vary widely, from various types of material already in the institution, to incoming donations and submissions, to content on the web, still in the hands of potential donors, or even content undergoing creation. This module describes the type of information to collect and how to go about developing the inventory to use it as a basis for developing digital preservation and curation policies, as well as decisions within the organization.

The initial inventory is expanded during the selection phase that follows. The DPOE module spells out how to develop your scope of selection, aligning it with the mission of your organization. Once the initial scope is determined, you'll likely need to prioritize where to begin and what makes sense given the current and expected levels of resources. This module

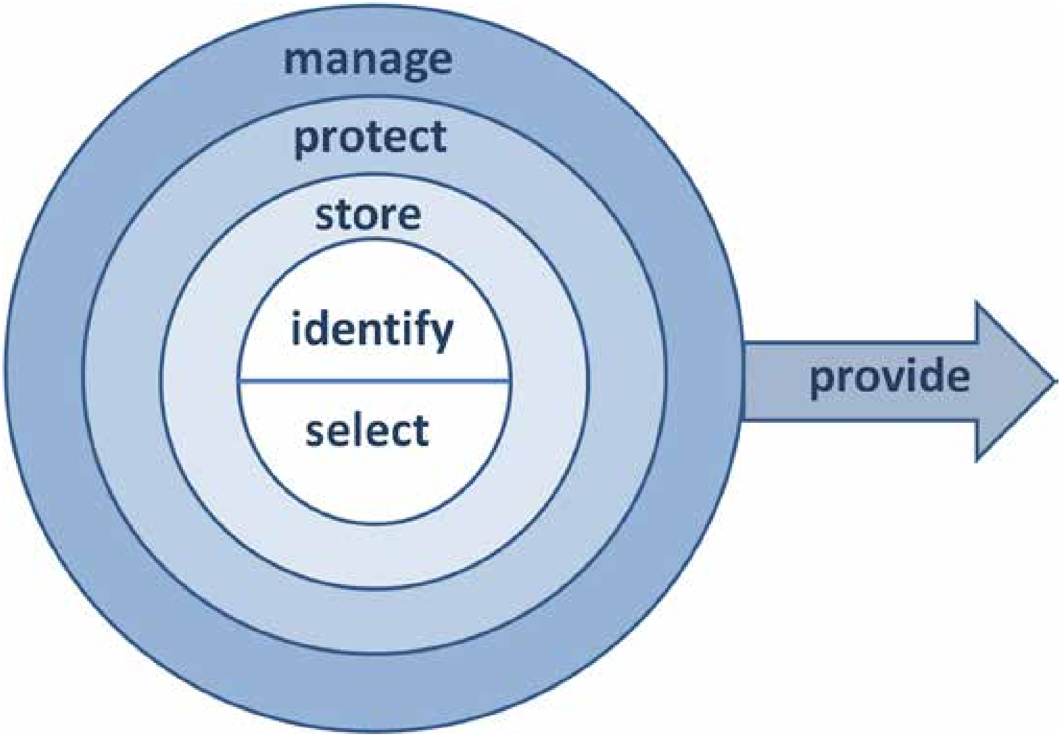


Figure 3.3. Digital Preservation Outreach and Education Modules. ***"DPOE Baseline Modules" v. 2.0, reprinted with permission of Digital Preservation Outreach and Education, the Library of Congress (http://www.digitalpreservation.gov/education/).***

provides a list of considerations that can help in deciding what is going to be most feasible for you.

Following selection is the module on storage, which describes what you store (file formats and metadata), as well as how you store it: the number of copies, storage media used, and how to go about selecting repositories. With storage comes protection, the next module. We often don't think much about protection, but we should. Who can access your content? What security measures do you have in place? What policies and procedures protect the content when there is staff turnover? There are other kinds of protection as well. Low-level ones include temperature controls in the server room, locks on the doors, and password encryptions. Mid-level protections include disaster planning against hackers, floods, tornadoes, bombs, and more. The highest level of protection would be that of business continuity: Who will take on your content and protect it if your organization is disbanded? The protection module provides excellent guidance about important topics we rarely consider.

Management is the next module, although in truth, management encompasses every phase of the process. Policies must be developed; procedures designed, tested, and implemented; and funding obtained and allocated appropriately. Management also applies to monitoring the content, ensuring the proper preservation actions are taken when needed, and transferring content to newer storage systems as needed (called "refreshing").

The last module is "access," or providing the content to users. Without the ability to provide access, there is, of course, no point to digital preservation. This module focuses on determining what form of access needs to be developed, identifying your "designated community" (those for whom you are curating the content), and clarifying levels of access when rights restrictions are involved.

In summary, the DPOE modules provide excellent guidance for those already in the field who need to develop practical skills and expertise to effectively curate and preserve digital content. Since these modules are normally taught in hands-on workshops, you will have the opportunity to ask questions and share ideas and concerns with others who have similar situations. These workshops are helping to build a network of digital preservation practitioners throughout the United States and beyond who can reach out to help one another and, together, build solutions most of us cannot attempt alone. The very approach taken by the DPOE initiative is one of building capacity and a supportive network, which may have a tremendous positive impact throughout time.

CHOOSE YOUR PERSPECTIVE (OR "MIX AND MATCH AS NEEDED")

You may have noticed by now that these maps share some basic concepts, for example, tasks for intaking, preparing content, and storing it safely, and also providing access. The more you look at various maps of the same territory, the more you start to recognize the landmarks (see table 3.1). Soon, you will have a better grasp on what is involved and where you are going. And knowing the perspective used in developing the map, or model, will help you determine when to use which one. What follows is a brief analysis of the overlap of the models, for instance, when you overlay one map onto another, to see what matches up, and which areas of the territory covered by each map are different.

In the OAIS model, the "producer" is tasked with identifying content and creating an inventory, but it does not give specific instructions that translate well to the cultural heritage field, as

Table 3.1. Aspects of Curation in Three Models

Aspects of Curation	*DCC Lifecycle Model*	*OAIS Model*	*DPOE Modules*
identification	create or receive	by producer	identify
creating/receiving	create or receive	ingest	select?
appraisal and selection	appraise and select	ingest	select?
ingestion	ingest	ingest	select and store
preservation actions	preservation actions	store in archives and manage data	store and manage
transformations	preservation actions and transformations	ingest and disseminate	store, manage, and provide
storage and protection	store	store in archives and manage data	store and protect
management and planning	preservation planning, community watch, and curate–preserve	preservation planning and management	manage
providing access	access, use, and reuse	disseminate	provide

Source: Copyright Jody DeRidder, 2017.

the model was developed for government records agencies. The DCC Curation Lifecycle does not address initial identification of content; instead it focuses on the creation of or receipt of content. Thus, if you aren't a records manager and need guidance on how to build that initial inventory, look to the DPOE module of identify. And while both the DPOE "select" module and the Digital Curation model include appraisal, in OAIS the appraisal is folded into the ingest process. Effective appraisal and selection are important for managing the scope of your work and ensuring that your digital curation dollars are well spent. Again, the DPOE "select" module will provide the most guidance on how to go about making that appraisal in the cultural heritage field.

We can map the "ingest" and "receive" part of the Curation Lifecycle Model stage of "create or receive" to the ingest phase of the OAIS model. In the DPOE model, actually obtaining the content falls somewhere before or after selection, and is not distinctly addressed. The methods for this process may vary tremendously. For example, web archiving will be an active method of obtaining content performed by technical staff. Soliciting content from donors would also be an outreach effort, likely performed by curators or archivists. Digital material already in the organization that needs to be collected into a digital repository for curation will require a different set of actions for "receipt" or "ingest," requiring workflows across units and departments, clear communication, and possibly administrative mandates. Other content arrives without notice. Some may come in by e-mail, on various forms of media and hard drives, or transferred via file transfer protocols or such services as Box or Dropbox. Moreover, organizations (or other parts of our own organization) may contact us and ask for our services. The latter is becoming fairly common in research institutions throughout the country, as awareness grows of the need of digital curation for important materials now in "silos." Oftentimes, libraries and archives lead the way either as guides or service providers.

All three models include preservation actions. In the DCC model, these are separated from storage but fall in the archival storage and data management components of OAIS, and the store and manage modules of DPOE. The DCC model also separates out transformation, which may happen as a preservation action or for access. If transformation is happening on ingest, it would then fall in the DPOE "store" module and the OAIS "ingest"; if occurring at the point of providing access, it would be part of the OAIS dissemination phase and the DPOE "provide" module. To confuse things further, “transform” in the DCC model may also refer to the use of something older in a new way. Thus, this mapping is not completely accurate but should provide some clarity nonetheless. Any way you look at it, the purpose of transformation is continued use.

The DPOE “manage” module matches up with the high-level preservation planning and management in the OAIS model. In the DCC Lifecycle, there is a preservation planning layer, but overall management likely encompasses the “community watch” component and the entire “curate–preserve” layer of the model. The DPOE “manage” module focuses on balanced overall management of your preservation approach, considering organizational requirements and objectives, technological opportunities and change, and resources, including your human resources. The DPOE protection module does not have a clear match in either the DCC Curation Lifecycle Model or the OAIS model, and it is likely incorporated into the storage component for each one.

Providing access is the entire point of digital curation and preservation, and has a clear component in all three models. The DPOE “provide” module is focused on providing access to users, which clearly matches up with the OAIS dissemination phase and would match up to at least the first part of the DCC Lifecycle stage of access, use, and reuse. The actual use and reuse of the content, however, is only covered by the DCC Lifecycle Model, which, of course, is considering everything with regards to the content itself. From the DCC perspective, curation should begin when content is being created. Ideally, content creators would provide archival format versions and appropriate metadata for long-term management from the point of creation.

Whereas the DCC Lifecycle Model views curation from the perspective of the digital content, the OAIS model views tasks from the point of view of the storage system itself, and the DPOE model takes the perspective of the curator and the organization or institution managing the content for long-term access. By looking at all the models, we get a clearer picture of the various aspects of curation, which can be summarized by this list:

identification
creating/receiving
appraisal and selection
ingestion (into a repository system)
preservation actions
transformations
storage and protection
management and planning
providing access

When you want to get a good picture of most of the aspects of digital curation to determine whether there are areas you need to develop or encompass, study the DCC Curation Lifecycle Model. As noted, the DCC Lifecycle Model highlights the involvement of creators, which is the best long-term solution for preparing content effectively for preservation. If you’re selecting a preservation repository system, you’ll want to study the OAIS model to determine how many of the necessary services are covered by the system and which ones you’ll need to find a way to provide. If you’re trying to determine how to get started, need to develop more expertise, or want to develop better content protection, consider the DPOE workshops as an excellent overall guide. All three models have their purpose, and all three are useful tools to help guide your decisions and those of your institution.

KEY POINTS

1. The DCC Curation Lifecycle Model is content-centric and includes the creation of content, as well as its destruction.
2. The OAIS model defines the primary requirements of a digital preservation repository: It starts with the intake of content and ends with dissemination of it to users.
3. The DPOE model provides guidance for digital curators and administrators in the field to gain expertise and develop useful networks and community connections.

4

Emulation, Migration, or Encapsulation?

Time marches on, continually driving the digital hardware and software we have used into obsolescence. The varieties of storage devices, operating systems, and software packages we use continue to change as vendors stop supporting older versions and offer for sale the newest, safer versions, with new bells and whistles designed to excite and entice. Yet, the files, databases, and content we developed or collected may have crucial dependencies on previous software or hardware. How can we ensure continued access to older digital content, despite this incessant change?

There are three primary methods adopted by digital preservation specialists, each with pros and cons to be considered: emulation, migration, and encapsulation.

1. Emulation is the effort to retain or imitate the original user experience of the content in question, often by recreating a semblance of the software and hardware environment.
2. Migration is the effort to transfer the content in the files to newer formats that are supported by current hardware and software.
3. Encapsulation is the effort to retain the information needed for future digital curators to reconstruct access to the content.

Dependent upon your resources and the types of content you are curating, you may select one or more of these methods to ensure continued access. Each one requires planning and effort from the outset, so consider them carefully, and do not delay in taking steps toward implementation. Laying the groundwork as you capture your content is ideal, for the underlying dependencies are a moving target, and you must keep your finger on the pulse of what is required by each new set of incoming material.

PROS AND CONS OF ENCAPSULATION

Encapsulation is the effort to document dependencies and collect necessary supporting software (and sometimes the hardware). The benefits of this approach are that the original content remains unchanged, and the user experience, when the hardware is retained, also remains the same. The difficulties are in the complexity and challenges imposed by the passage of time and the need for a complete understanding of the computing environment necessary for each type of material. There are two basic approaches—one for those who are unable to provide access at present and one for those who are.

In the minimal approach, managers try to identify and clarify exactly what is needed to access their content and save copies of software, supporting systems, and more, with the content, with clear instructions, in the hope of future reconstruction or emulation. For example, if encapsulating Postscript or PDF documents, first install a virtual machine (VM)[1] on your desktop computer, and on the VM, install a current version of Linux (an open-source operating system).[2] Then install an open-source software that effectively accesses Postscript or PDF documents, for example, Evince.[3] Install the necessary dependencies and test the software with a copy of your Postscript documents to verify that they display and function correctly. For audio files, you will probably want to install Audacity[4] or VLC media player,[5] plus LAME[6] for derivative creation; for images, you will want to install GIMP[7]; for video, FFmpeg[8] or VLC; for text or HTML, vim[9]; and for word processing and office documents, OpenOffice.[10] Save your modified operating system in your VM and shut it down, noting where the image is stored. Finally, copy this modified Linux operating system image from your desktop and store it with the content (and include documentation of how to install and use it). It is insufficient to simply store copies of the rendering software, as this will not include all the dependencies needed for installation, and those dependencies will need to be specific to the operating system on which the software is installed.

As you can see, this is not a simple solution, and it only forestalls eventual migration or emulation; however, if you do not have access to emulation services or the resources to host the content locally on the original hardware and software, this form of encapsulation provides a "life raft" to potentially provide digital access to the material in the future. Whatever form of file organization a manager chooses, the author recommends providing a descriptive "map" in plain text at the top level of the archive. This should describe both what is in the archive (and why it's important) and how to reconstruct the items (and collections, if items are stored within the collection directories) appropriately. Remember to note the operating system type and version, and preferably the "endianness" of the files. Endianness indicates whether data is stored right-to-left or left-to-write on the computer. The following operating systems are little-endian: AXP/VMS, Digital UNIX, Intel ABI, OS/2, VAX/VMS, and Windows. And these are big-endian: AIX, HP-UX, IBM mainframe, Macintosh, and Solaris.[11]

Institutions with more capacity take encapsulation further and retain the original software and hardware—the original computing environment. Users are generally allowed access to the computing environment in a secure environment, as the systems cannot be updated for security purposes and hence cannot be safely web-accessible. As time passes, access to parts for repair become impossible, so hoarding of old computers may be required to ensure use can continue (some call this the "computer museum" approach). Repairs to the software and the system require arcane knowledge, and the extent of the material available via each computing environment retained is limited to what was originally usable via that environment. For an organization to provide access to many materials throughout time would require maintaining many different computing environments. If, however, your organization has a large quantity of valuable materials or software best experienced in a particular computing environment, and your researchers are willing to come in person to your facility, this may be a useful approach for you.

PROS AND CONS OF EMULATION

"Emulation" builds on encapsulation, and it uses the knowledge of the original system, and the collected software and images to emulate, or recreate, the original user experience of the content in question. Normally, this involves documenting the original dependencies and underlying system, and then creating support for those dependencies and software expectations in newer systems, often within virtualized environments.

The benefits of this approach are that the original object is unchanged and the original computer environment is (as much as possible) rendered. The difficulties here lie in the complexity and costs of implementation, the extent of knowledge required of the original computing environment, and the limitations of information reuse as each environment is isolated.[12] On top of all of this, emulators must then be preserved to be useful in future systems.[13]

An example of a method of emulation is the Dioscuri project, which takes a modular approach to recreating each of the basic component parts of a computer, within a VM.[14] VMs can host completely different operating systems than the host computer and be installed on multiple platforms (e.g., Mac or PC). Use of Dioscuri requires "deep knowledge of fundamental computer hardware concepts,"[15] but is freely available on SourceForge.[16]

Recognizing that few people have the technical knowledge to successfully use such systems, the later KEEP Project (see figure 4.1) incorporated Dioscuri and six other emulators (Qemu, VICE, UAE, BeebEm, JavaCPC, and Thomson) to effectively emulate six platforms

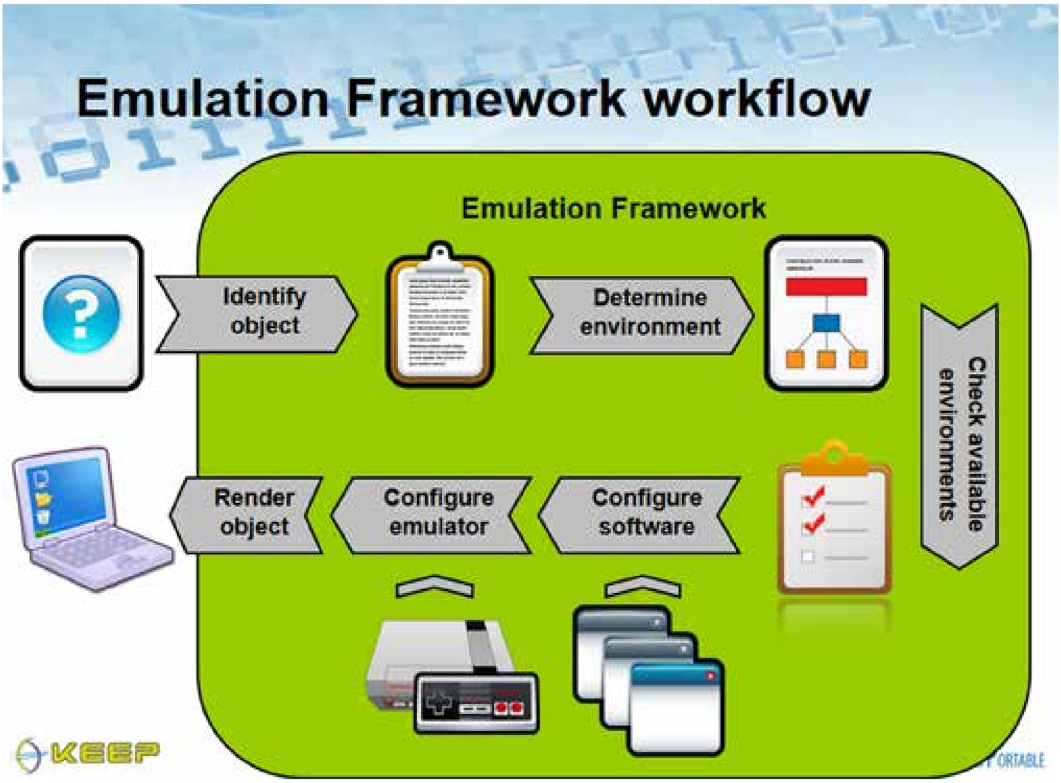


Figure 4.1. Model of the KEEP Emulation Framework. ***Reprinted with permission of the KEEP Project, copyright 2011, available from http://emuframework.sourceforge.net/about.html.***

(x86, C64, Amiga, BBC Micro, Amstrad, and Thomson T07) and more than 30 file formats.[17] This groundbreaking project automated the process of identifying the file, locating the required software and platform, matching dependencies, configuring the emulator, and injecting the file into the emulated environment, turning control of the emulated environment over to the user.[18] This openly available system[19] depends upon an emulator archive, a software archive, and a technical metadata registry.[20] (Details of the architectural design are freely available for the technically curious.[21])

Other developments provide emulation via the web. Olive, developed at Carnegie Mellon University, focuses on archiving software packages (e.g., games, simulation models, data visualization, dynamic websites, expert systems, and tutoring systems). Olive requires you to have a Linux system with installed client software that interacts with the Olive Web server.[22] After installation, you would use the client to request and then emulate the computing environment needed, and then you would install the (requested) software package you want to use.[23] Any changes you make are made to your local copy, so the preservation originals are unchanged.

The University of Freiburg developed bwFLA to provide "Emulation as a Service" (EAAS) in the cloud, so no downloaded software is necessary. Each request for a new emulation is assigned a VM, within which the required operating system and software are overlaid and the system is booted to run the software necessary to access the desired content.[24] The intent is to deploy individual emulation components in large-scale cluster or Cloud infrastructures. The EMiL (Emulation of Multimedia-Objects in Libraries) project (2014–2016) extended the bwFLA project to simplify usage within libraries and for library patrons.[25] Users select a digital object from the library catalog, and the catalog process passes the identifier to the EAAS framework, which characterizes the item and suggests an emulation environment (if multiple, the user can select one), and then the object is rendered in the user's browser using HTML5 techniques.[26] At present, bwFLA has emulation components for all major past and present desktop systems.[27]

Continued, effective, and widespread use of emulators is heavily dependent upon the development and maintenance of software repositories containing all the necessary components for use in a variety of emulations. Olive, for example, maintains an archive of executable software[28] as well as a repository of VMs (which for legal reasons are only available to their collaborators).[29] Technical file format registries are needed to track the types of software (and dependencies on that software) needed for rendering each type of file, and for using each type of tool; an example for this is "TOTEM" (Trustworthy Online Technical Environment Metadata Database), which was developed for the KEEP Project.[30] As of 2011, this database already supported the PC architecture, Commodore 64 architecture, and console gaming platforms.[31] Collecting the pertinent technical metadata about the computing environment for software and keeping it up to date is an incredible challenge that requires a collaborative approach; it is far more than any one institution or agency can manage.[32]

PROS AND CONS OF MIGRATION

Migration is the conversion of content from one format or carrier to another, to increase accessibility and usability in current and future computing environments. Normally, content is migrated to nonproprietary uncompressed formats that are well-supported in the preservation community,[33] are platform-independent, and retain the highest possible quality capture. This is the lowest-cost method for digital preservation, but it has some drawbacks. First and foremost is that, in any transformation, there is the likelihood of some loss of meaning, usability, appearance, or functionality. This leads to the thorny question of determining what the "significant properties" of the content are, how best to retain them in the migration, and the "acceptable loss" to the user community. How do you determine the "significant properties" of something? A full analysis would include the following:

the expected functions and behaviors of the object
the structural properties needed to achieve each function
the functions stakeholders perform when using the object
the quality thresholds needed for each structural property to perform the necessary functions.[34]

When in doubt, you must engage scholars who are representative of your designated community to evaluate migration results. The more complex the object, the more complex this process becomes, which is part of why migration is normally restricted to simpler digital objects and those without embedded content.

When you are selecting a format for migration of content, consider how your target audience would interact with the content. In visual works of art, colors may be crucial. In this case, you would need to compare the colors, lighting, contrast, and color saturation between the original and the migration result to evaluate success. If, for example, the original is an image of a text file, the significant properties of the content are most likely the textual information. In this case, the format selected for migration would be a text format, not an image format. This means you will need to find a way to extract the textual information, perhaps by using optical character recognition (OCR) and then correcting the errors by hand. Even then, the formatting and display of the text may be key to usage, for instance, when the text includes scientific formulas or the visual presentation of the text is crucial to understanding or experiencing what the writer or publisher sought to communicate.

In these cases, XML formats suitable to the display should be used, for example, MathML[35] or OpenOffice.[36] Even then, the result of the migration may not meet your expectations; formatting and display errors are common in migrations, and sometimes content itself is lost. In cases like this, migration is far from simple. If the significant properties of the content are not considered carefully, your migration results will be poor. Always, always, retain the original content, as the future may offer better options for migration and/or emulation. Also, never use your only copy for migration, as the process often destroys the original.

This last example points to another problem with migration: that of selecting, finding, or creating a viable migration process. Whatever methods you locate and implement now will likely be improved in the next few years, so keep an eye on developments in the field. The author recommends that you regularly review the proceedings and presentations of conferences that regularly feature new research, for instance, iPRES,[37] Archiving,[38] and the International Digital Curation Conference.[39] Migration methods and approaches are discussed further in chapter 5.

Migration of such interactive materials as websites and executable software adds layers of complexity that are generally best avoided. Any rewrite of obsolete software (assuming it is open source, and hence readable) will normally involve a deep understanding of the old programming language and the ability to effectively translate the functionality into a newer one. Since few programming languages perform the same function in the same way, the result has a high likelihood of differing from the original, even when successful. In general, migration is a tool for static files and is best suited for those with minimal complexity, suitable current formats, a known migration path, and a clear understanding of the significant properties to be retained.

CHOOSING YOUR APPROACH(ES)

Migration is the lowest-cost option, but it is primarily suitable for static files, not software or interactive systems. This is the method most widely embraced by the cultural heritage community, particularly for locally digitized materials, where the quality is high and file types involved are limited. Encapsulation is a way of postponing migration or emulation and is best suited for complex materials that cannot be effectively migrated and for which emulation is not yet available. The "computer museum" option of retaining and supporting the original hardware and software to provide local access and the original experience is a time-limited option, as equipment and media will wear out and eventually not be replaceable. Still, this is a valuable temporary solution for pockets of high-value content needed by local scholars when the material is not yet effectively usable via emulation or migration. If, for example, your organization has large quantities of a particular type of computer game or software, you may want to consider this option.

Emulation, which is ideal for software and systems, is the highest-cost option, but increasingly emulation services are becoming available, including via the web. If you have complex content or software for which your designated community needs access to the original experience, seek out emulation methods or services. Be mindful that the success of these services is heavily dependent upon multi-institutional and/or community support. Much of the emulation development has been driven by game enthusiasts,[40] with limited development and support by institutions and preservation-oriented organizations. Continued development and support of emulation for cultural heritage, business, scientific, and governmental needs will require broader support within the community and the coordinated engagement of experts and enthusiasts alike.

KEY POINTS

1. There are three primary methods of providing access to older digital content: encapsulation, emulation, and migration.
2. Encapsulation is the effort to document all dependencies and collect all necessary supporting software (and sometimes the hardware). This is difficult, and emulation and/or migration will still be needed at some point.
3. Emulation builds on encapsulation, seeking to recreate the original user experience. This is complex, but recent research is beginning to make this more viable as a community effort.
4. Migration is the transfer of the content from one format to another more archivally sound and current one. Success is heavily dependent upon whether the significant characteristics of the content were transferred, and results must be checked by humans. Most migration results in some loss of content.
5. Retain the original, as future methods of access may be far better than what we have now.
6. Choose your approaches based on your resources and the types of content you are curating.

5

How Do I Identify and Select Content?

If you want to develop an effective plan for digital curation, you will need a reasonable understanding of the scope and types of content to be preserved and curated. A useful approach is to first clarify the focus areas for collection, then identify the possible content for inclusion from within those focus areas, and then select from within the identified options. Identification and selection will be an ongoing process, so your ability to maintain clear communications and agreement among stakeholders and administrators about appropriate scope will be crucial to your ability to develop a practical, realistic, viable, and sustainable digital curation program. Scope creep is your enemy! No one has unlimited resources.

When an administrator or stakeholder suggests including a new set of content, you must be prepared to ask what content you should no longer accept or maintain, in exchange, or how you are to obtain the resources necessary to effectively manage the additional content. With that said, it is only realistic to expect your organization’s focus and values to change throughout time, and your selection priorities should always align with the mission of your organization. Thus, your selection criteria will change throughout time. It may also be subject to change when you realize that either you have taken on more than you can effectively handle or you discover that you can and should manage more. The person who has the deepest understanding of the ramifications of a change in priorities (likely you) is the person who must ensure stakeholders are involved in this process and convince others of the importance of careful and effective planning. With knowledge comes responsibility. Long-term curation of digital content is a commitment and not one to be taken lightly.

SELECTION PRIORITIES

Define and document your selection criteria to avoid confusion and justify decisions made to those who challenge you. This is a form of collection development. As selection criteria change throughout time, be sure to note the changes, the date, the reason for the changes, and the extent to which they are to be retroactive. If the new selection criteria preclude continuing to curate content already collected, difficult decisions need to be made, developed into written policies and procedures, and implemented. Those policies should state whether existing content no longer in your scope should be deaccessioned and either handed off to another institution for continued management or deleted. If previously curated content must be deaccessioned, ideally you should transfer it carefully to another digital curator to avoid loss of potentially valuable material. It may still be important in other contexts or to other groups of people.

In developing your initial selection parameters, begin by considering the mission of your institution: Are there limitations based on legal mandates? If you are working in a business, the focus will be on the materials that support, protect, and extend your business. If your institution is educational, perhaps the beginning scope will simply be materials that support research and education for the age ranges your organization serves. What defines the community of people for whom the digital material will be provided? This is called your "designated community." In higher education, your designated community may be faculty and students. In a state archives, your designated community may be legislators, state employees, public figures, and anyone needing access to governmental documents. In a public library, your designated community may be the general public in your county or city. By clarifying the groups of people for whom the content must be stored and made available, and then the types of information that they need, you begin to get a sense of the outlying boundaries of what you potentially should collect. Some of the potential guidelines you may want to use are as follows:

Your organization's collection or acquisition policy
Priorities or precedents already established
Significance of certain types of content for researchers and their interest areas
The extent to which the content is unique and cannot be reproduced
Historical or evidential value for your organization or stakeholders.[1]

Appraisal and selection is a complex process and requires many judgment calls. A useful resource is the National Archives and Records Administration's appraisal policy,[2] which provides numerous questions to consider, many of which are reflected in the aforementioned list. Whenever possible, develop clear policy guidelines for selectors to avoid as much confusion and error as possible.

Another important part of your initial selection policy will be determining who should be deciding what types of content are important and what should be ignored. If you are working in a government archives, these decisions may already be clarified by legal mandates; however, if you are working in a cultural heritage institution, you have both the opportunity and possibly the moral imperative to engage experts in determining selection priorities and values. After all, the person who selects what will remain of our history is the one who determines what is in that history, when viewed in the future. To whom should such an incredible honor and responsibility belong, if not to the experts in the field?

If your designated community includes researchers in the arts, sciences, and humanities, it behooves you to locate experts in those fields who would be willing to assist in clarifying what types of information you should be collecting. Ideally, these experts will also be involved in identifying potential content, if possible, and submitting lists of possibilities for consideration. It is only by developing a comprehensive network of scholars throughout the world who are willing and able to assist in selection that it will be possible for us to appropriately select what should be retained for future use. If you are such a scholar, assisting in this process with local, regional, or national cultural heritage institutions will help ensure that a fair representation of valuable knowledge will be retained and accessible. Those who select content are, in a very real sense, shaping our history and our future. An example set of general policies that can easily be modified to meet your needs is as follows. Content selected for digital curation should meet these guidelines:

1. It is highly significant and useful for our designated community, as determined by experts in the field.
2. It is unique, cannot be easily reproduced, and is unavailable from other sources.
3. It falls within our current acquisition policies.
4. We either own all rights to the content or are easily able to procure the rights for preservation, curation, and access.

Once the primary selection policy is both informed and developed, and initial selectors engaged, it’s time to begin to identify content that could be considered and then make the final selection from that list. This will be an ongoing process, and it’s best to specify who does what and how communications should happen between people who are filling different roles to ensure everyone is on the same page. Document processes and update them as necessary; make sure policies and procedures are easily available to those involved, as well as to administrators, to justify funding and resource expenditures.

IDENTIFICATION

Identification generally begins with building an inventory. Normally, you would begin creating the inventory with general information and add more specific information later, so it’s best to use a spreadsheet for this purpose. Start by asking the following: Where is the content we need to manage? On what media or servers, and where are those located? Most institutions will have digital content on a range of media, from flash drives and CDs and hard drives to servers, in multiple locations, and some may only be accessible via the web. How much space does this take up? Can you get an approximate count of the number of files? What types of content are there? This may be at least partially determined by file extensions, but the initial inventory is likely at a higher level and may be based on directory names or “institutional knowledge”—what employees know but have not yet written down in a form that is reusable.

In the beginning stages, “types” might be “genres,” for instance, “photographs,” “text documents,” “video,” or, more generally, “websites.” Are there any known rights issues with the material? Is there any descriptive information that might help determine the value and use of the content? Capture any information that may be crucial in determining what is worth keeping and what may have restrictions that prevent effective digital curation. Also note where this descriptive and rights information was found, so later investigations can build upon the initial inventory. Note when you logged entries into the inventory (and your name, since others may also be assisting). When possible, collect date information (how old is this content?), as this may determine priorities in later steps. Here is an example entry from an initial inventory, with the spreadsheet columns represented by the words prior to the colon on each line:

Category: Special Collections
Title/Description: Railroad Photographs, SE U.S.
Type: images, digitized
Format: TIFF
Extent: 242 GB; 2,250 images

Location: archival server in Room A, Central IT
Coverage Dates: early 1900s
Creation Date: January–June 2006
Inventoried: 12/15/2011, by Fred Jones[3]

By taking these initial steps to identify the content you might want to preserve, you are collecting the information you will need to develop your preservation plans and priorities. This inventory should be kept in a secure but shared location and updated regularly, as content moves and new content arrives or is found. This document serves as the basis for effective digital curation, so ensure that management of this inventory is assigned to responsible people and available to those who select the content for long-term curation (which is the next stage of the DPOE model[4]).

FINAL SELECTION

Once an inventory is in place and as it is updated, personnel who are intimately familiar with the capabilities of your organization need to make some practical decisions. The digital curator or administrator who oversees this process will need to pay careful attention to decisions made on this level to ensure they are aligned with mission of the organization, the extent of resources available, and the needs of the designated community. Sooner or later, there will be a conflict between expectations and what is realistic. At these times, it is necessary to make the case to upper administration or funders to obtain additional resources or expertise to fulfill the mission of the organization. If you are unsuccessful in obtaining the necessary resources, the selection policy must be modified to what is actually possible. Integrity is important. Say what you will do, publicly, and then do what you say. This builds trust in the community and among your designated community. It also clarifies for employees how they should be spending their time, and in the end it will save you unnecessary expenditures.

One of the first questions you should ask in making the final selection is this: Is it feasible for you to preserve the content?[5] Costs and technical considerations may make it beyond the capabilities of your staff and resources. For example, retaining old video games in a usable form will be far more challenging than curating images. Storing videos also requires a great deal of space and complexity of metadata, as videos are composed of multiple types of content: text, audio, and images. Curating software requires retaining extensive information about the dependencies and operating system, and will likely require emulation services to make it function in the future.

A second question to ask is this: Will it be possible for you to make the content available?[6] If it is already inaccessible, the answer is no; and if there are rights restrictions that preclude access and use, there is no point to keeping the content, unless you are willing to risk legal repercussions. While your policy for initial selection can mandate limitations on the types of content to be considered based on a number of parameters, your final selections will clarify where the rubber hits the road on these important issues. Be sure to create a feedback loop between those making final selections and those who manage the initial selection policy, as the mandates of reality may require modifications of collection scope to avoid unnecessary work.

The final selection of content will likely also involve detailed identification of content, and some of the information gathered at this stage should be used to update the inventory. After all, this is the beginning of the process of ingesting content into your repository. As described by Helms and DeRidder, the ingest stage includes the following functions: “(1) acquire and document, (2) write block media, (3) virus and malware scan, (4) image disk or copy files, (5) validate and establish fixity, and (6) inventory.”[7] Their survey in 2016 sought to identify the tools most useful for content identification (see figure 5.1), among other tasks.[8] Once the content to be obtained is identified, someone has to acquire it, if it is not already in hand. That may mean pulling it from the internet; extracting it from media, computer hard drives, or servers; or obtaining the content via e-mail or file-sharing software like Dropbox.

In all cases, protect the computer from viruses and malware, and if possible have the intake computer offline to prevent infections of your network from downloaded “worms” and other wicked parasitic programs. Content should be scanned with updated malware and antivirus software for potential danger but also protected from inadvertent modification. Remember, if possible, we want to keep the original content untouched. Use write blockers (hardware versions are better than software) whenever possible, to prevent modifications of the original material. If the content is unsafe, you need to have in your policy statement whether you will allow the dangerous content to be quarantined or deleted (hence modifying the original), or whether this status precludes inclusion of the content in your repository. Once you are certain the content is safe, obtain checksums before transferring or copying the content; you can then verify the checksums in the destination location to ensure the copy process completed without any data loss.

Context is important and should be captured along with the content being selected. There are two important types of context in the digital environment. One is the type of operating system used to store the originals. This can impact how the files are stored and whether a

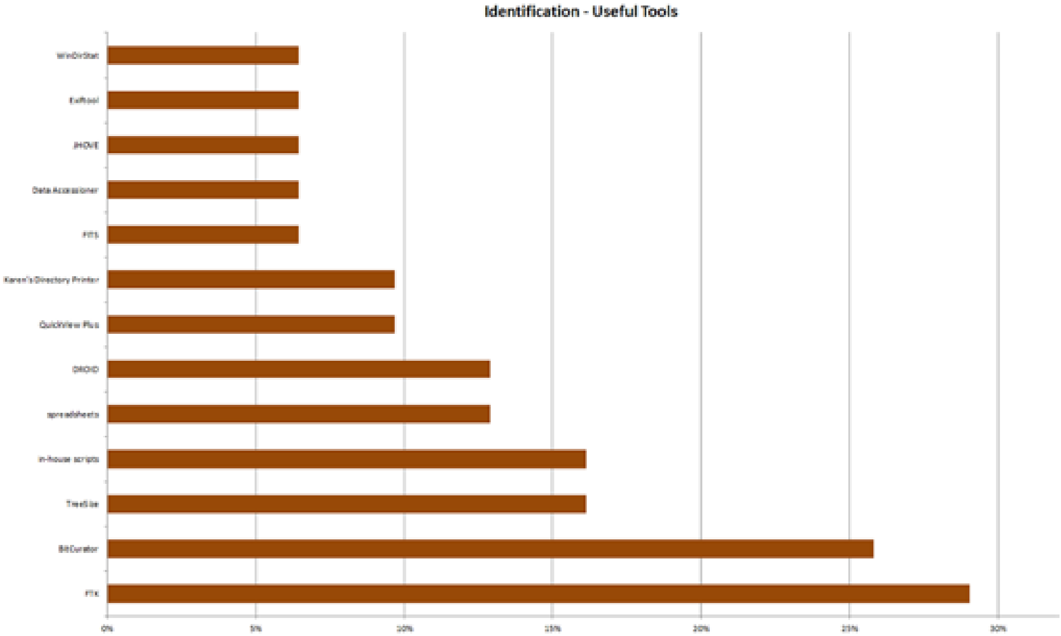


Figure 5.1. Useful Tools for Content Identification. ***“Useful Tools for Identification” from “Intake of Digital Content: Survey Results from the Field,”*** **D-Lib Magazine** ***22, no. 11/12 (November/December 2016), doi:10.1045/november2016-deridder. Reprinted with permission of the authors, Jody L. DeRidder and Alissa Matheny Helms.***

file should be "read" from right to left, or left to right, by the software that opens it. (This is called "endianness." Macintosh computers store files differently than Windows computers.) Secondly, the organization of the files and directories can provide clues as to the value and content of the material, and may be linked or even part of the intellectual whole of the item. For example, videos are normally composed of multiple files; the same is true for software. Extracting a single file from the set of files necessary to render the video would leave you with a fairly useless digital object.

When faced with large quantities of incoming content from a single source (e.g., a hard drive or server), the first steps of appraisal and selection will involve some practical measures, for instance, deduplication, bulk inclusion/exclusion, filtering options, and more. Thus far, the best single set of tools available to manage this work is BitCurator,[9] developed with Mellon funding by the University of North Carolina at Chapel Hill and the Maryland Institute for Technology in the Humanities. Also, the DigiPres Commons[10] Community-Owned Digital Preservation Tool Registry (COPTR)[11] collaborative effort has collected many links to tools, organized by function. This is a wonderful resource. How best to use those tools and when one is preferable to another would be best practices we have yet to create. When in doubt, reach out to others in the field and ask what they have done or recommend (see the appendix for resources and recommended listservs).

SPECIAL CONSIDERATIONS

Content in the Wild

Since online digital content is constantly changing and updating, harnessing it into something that is feasibly preservation-ready in a form that enables effective use is a lot like harnessing wild animals. Think ahead to what you want to provide to your users, but also think about what levels of scope you can handle effectively. For example, social media and websites link to or embed external content; where will you draw the line on what you collect? How do you plan to deal with the variety of formats and the rights issues involved for content from so many different sources? When collecting websites, your crawler may be capturing one part of the site while another is being updated; it will be a challenge to capture snapshots that are accurate representations at any given time. How many variations of content do you plan to harvest, and how will you make this useful and usable for your designated community? If collecting e-mail or social media, will you attempt to collect conversations and exchanges with others and track them in threads? This may require getting access to many accounts and permissions from many people. It may also require some creative methods of structuring the captured content or linking and metadata to recreate the experience for users. This section reviews some of the primary concerns for "live" digital content and also recommended strategies.

There are two basic approaches to web archiving. Micro archiving is the collection of a limited number of websites in a limited period of time, generally by those with limited resources and technical expertise, and often for particular research purposes.[12] Macro archiving is carried out on a large scale and during long periods of time, usually by institutions, to capture cultural heritage. Although you may plan on macro archiving, begin with micro archiving as a pilot test to determine how best to proceed. You will want to test the functionality of your capture software, examine the results, and determine what you will need to do with those results to ensure long-term access and the type of usability you have in mind for your designated audience. Most web archiving efforts select content based on the domain (e.g., .gov or .edu), event or topic, genre, and media type,[13] so identify your scope and then begin to select your initial (or "seed") URLs. Jeremy Floyd describes the initial workflow as follows:

1. Set seed URLs and scoping rules.
2. Run a test crawl.
3. Review the crawl; if needed, modify the seeds or scope and return to step two.
4. Set frequency and run a crawl.
5. Review quality assessment reports.
6. Enter descriptive metadata.
7. Monitor for changes and new URLs to capture.[14]

To capture an entire website, you would need to capture not only the structure of the site and its elements, but also the actual and possible movements between and in the elements of the structure.[15] Since most complex sites are database driven, this is virtually impossible without access to the underlying database and software or, at minimum, a current sitemap that contains every useful link on the site. Hence, most archiving software tools only seek to capture web pages and sometimes linked documents and subpages, and generally have to be configured to a set "depth" of capture. For example, if the main web page links to subsidiary pages, that would be a depth of two captures; if each of those is linked to more pages or content, that would be a depth of three captures, and so on. The number of pages harvested from each site can thus grow exponentially, especially if capturing video, audio, and other media. Moreover, each page captured may link to resources on other sites, which may or may not be crucial to the usefulness of the website. If you want to limit capture to only content available on the targeted website, examine your selected capture software to determine if it allows for screening of the base URL on embedded links and content. Also check to see if your software avoids capturing the same links multiple times, as several pages in a website will likely link to one another. Since many web pages interlink, the software must be able to avoid endless loops (traversing links back and forth between the same pages endlessly).

Web archives are normally captured in the form of a WARC[16] (Web ARChive) file format, which is a concatenation of one or more WARC records.[17] Each WARC record consists of a record header containing date, type, length, two newlines, and a content block that contains the resources in any format.[18] There are eight types of WARC record: warcinfo, response, resource, request, metadata, revisit, conversion, and continuation.[19] WARC was developed from the Internet Archive ARC format[20] to improve the harvesting, access, and exchange needs of organizations, and it includes related secondary content (e.g., assigned metadata).[21] Both file types use lossless data compression.

Captured websites provided to users will not function as do websites in the wild, as searching and other capabilities supported by server-side software will no longer be available; however, users will be able to search within the page and browse links to the extent that the targets of those links have also been captured. As stated by the Library of Congress, "The aim of web harvesting is not to be able to rebuild the website with all its functionality, but to capture web pages as the user viewed them, to the extent possible."[22] Still, the appearance of captured sites may differ from online sites due to lack of access to style sheets, JavaScript, and other server-side functionality that is usually prohibited to crawlers by the robots.txt file on the site in question. This file is used by website managers to prevent search engine crawlers from accessing particular parts of the site or particular files, but this prevents the capture of information that may be crucial for correct display and functionality, and thus curation.

To overcome this barrier, in early 2017, the Internet Archive announced[23] they would henceforth ignore the directives of sites' robots.txt file. Thus far, the Internet Archive is the clear leader in capturing our online cultural history, forging a path that requires the least cost and effort for the greatest amount of gain. Their success since 1996 is impossible to deny, with more than 279 billion web pages, 11 million books, 4 million audio recordings, 3 million videos, and 100,000 software programs captured since 1996, and these figures are growing daily.[24] The Internet Archive Archive-It team has even developed a life cycle model[25] (see figure 5.2), which echoes many aspects of the Digital Curation Centre Curation Lifecycle Model[26] (discussed in chapter 3).

Large-scale web archives usually approach rights issues by offering to take down content to which there are objections (the "opt-out" model).[27] The alternative is to seek permission

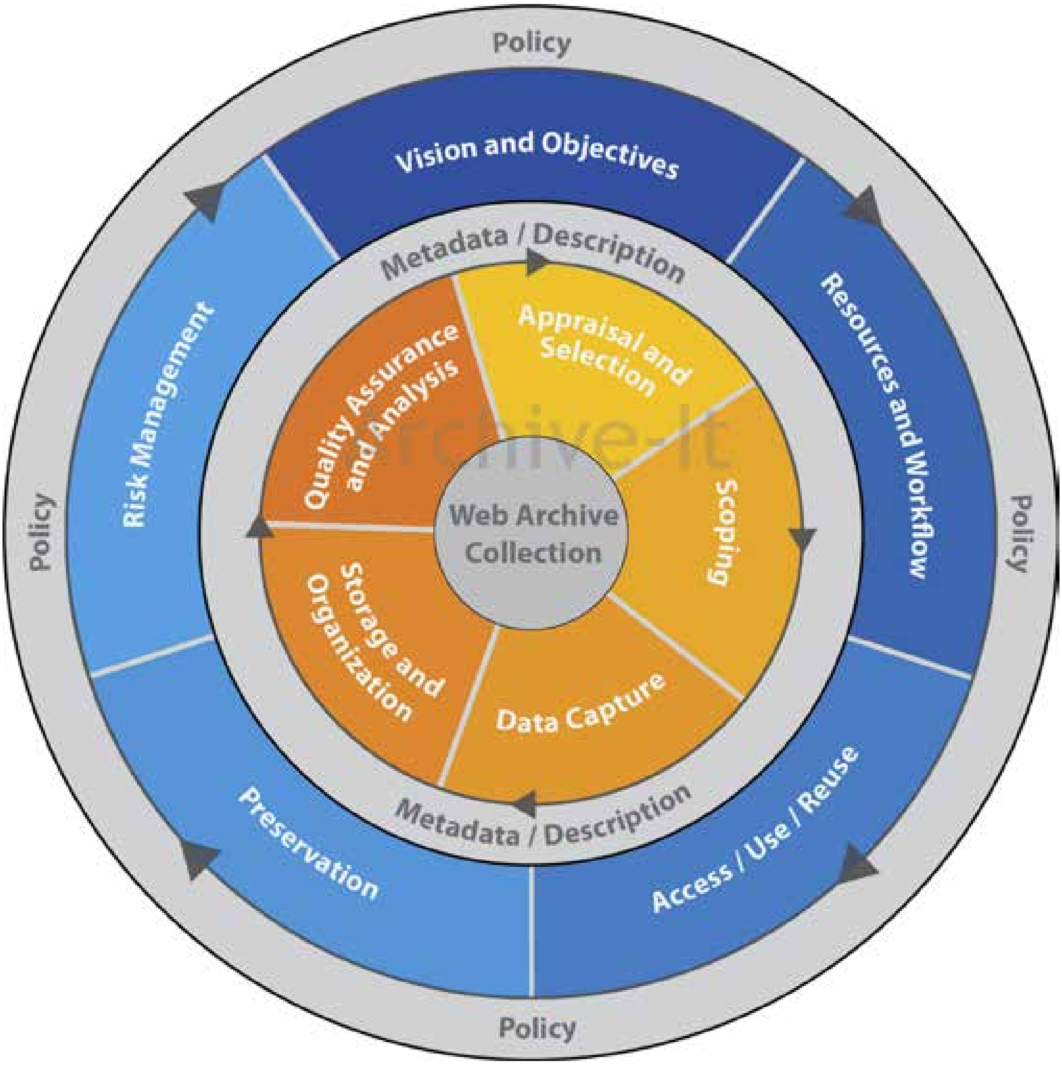


Figure 5.2. Internet Archive Web Archiving Lifecycle Model. ***"The Web Archiving Lifecycle Model," Archive-It Team, Internet Archive, March 2013, available from http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. Reprinted with permission of the Internet Archive.***

from copyright and intellectual property rights owners of the web content (the "opt-in" model) prior to capture. The latter method is time and resource intensive, and more feasible for micro archiving, where far fewer contacts need to be made. Efforts to contact all rights owners and obtain appropriate permissions are often hampered by an inability to identify those owners or locate current contact information.

Since the point of web archiving is to capture representations of the websites as they existed at particular points in time, formats of files captured are retained for access in the form found, rather than normalized or migrated.[28] Provision of user access to captured websites is dependent upon software ability to render the content effectively, and the International Internet Preservation Consortium is working to develop an open-source method of providing that access ("OpenWayback").[29] Future methods of access will likely need to involve emulation.

Social media capture is a subset of web archiving focused on user-generated online conversations and interactions in such proprietary environments as Facebook, Twitter, LinkedIn, YouTube, Flickr, Tumbler, and more. These platforms carry their own rights and access restrictions, and have their own methods of providing archiving (or preventing it). An excellent synopsis issued by the Digital Preservation Coalition in 2016[30] outlines the primary challenges in capturing and providing social media for research. While some of the platforms provide application programming interfaces (APIs) for access to collect content, these come with policies and agreements to restrict sharing and reuse.[31]

Since social media content is user-generated, analysis (particularly when combined with data from other sources) may reveal personal information about users that could potentially violate privacy or ethics and raise legal challenges.[32] The social media landscape continues to evolve, and the challenges inherent in both capturing and obtaining rights permissions are complex. In 2013, the National Archives and Records Administration issued a white paper on best practices for capture, in which they recommend methods of capture suited to each social media provider[33]; to date, there have been no additional current best practices published.

Due to the tremendous volume of available content, selection and scope are particularly important when web archiving or collecting social media. At this writing, the Library of Congress is still struggling with how to manage and make available the Twitter donation that began in 2010.[34] As one wry commentator described the situation, "The Library of Congress finds itself in the position of someone who has agreed to store the Atlantic Ocean in his basement."[35]

Once content is captured, you will still need to work out how to make it searchable, browseable, and accessible. For whichever type of content you seek to harvest, carefully review the restrictions and seek out how others are providing access, as the methods are still very much in development. For general web archiving, the Digital Curation Centre provides a list of web archiving software and services that may be useful.[36] The easiest option is to enroll in a subscription web archiving service, for example, Archive-It, which offers browsing by URL and full-text searching.[37]

Content from In-House or Donors

If you already have digital content within your institution that is valuable and must be made available throughout time, start by making an inventory of this. Remember to consider materials that may be on CD/DVD, flash drives, zip drives, floppies, external hard drives, and other media, stored in drawers, cabinets, shelves, and closets within your organization. These items may or may not have been adequately documented in your catalog or content management system. Some digital content media may be included in books, for instance, an added DVD or CD, and locating it may be difficult. Digital content may have already come in via e-mail, Dropbox, or other methods, and may be stored in multiple accounts or on multiple desktops. Digital content on servers and Windows share drives may include electronic theses and dissertations, research papers, digitized material from special collections and archives, purchased e-books, maps, geospatial content, databases and spreadsheets, research data, audio, video, images, text, e-mails, and even software. Dependent upon the needs of your institution or audience, web content may need to be included.

Online guidance for donors is highly recommended. A beginning point is demonstrated by a web page provided by the Indiana University Born Digital Preservation Lab,[38] which spells out which file formats are acceptable, what types of storage devices can be used to transfer materials, and online transfer options. Ideally, your website (and available handouts) will also provide guidance as to the scope of your collection policy (as does the University of Wisconsin–Madison Libraries site[39]) to avoid unexpected donations of materials far from your field of focus. Since it is sometimes difficult for donors to understand that ownership of material does not convey intellectual property rights or copyright, it is also helpful to provide some explanation and links to resources, for example, the excellent Council on Library and Information Resources report *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*.[40] This report explains privacy and intellectual property, as well as the content reviews and key stages in acquiring materials, including safe transfer and the risks of including unexpected files.

Ideally, a curator will work directly with a donor to review content prior to the transfer. The curator can collect information about exactly what should be retained and what should not, as it is common for hard drives and other storage devices to contain a wide variety of content. Reviewing the material together will allow the curator to educate the donor about sensitive information and privacy considerations (social security numbers, bank accounts, medical and personal information), intellectual property rights (was any content created by others, or do images or videos portray other people?), and deleted or cached content. This is also an excellent time to determine whether the content truly falls within the bounds of your collection policy and gather initial metadata, which will be useful in assessing the value of the content and eventually providing access. Be sure to document which content will be donated that may require restrictions or redactions, since this is an extra layer of work. Remember that you will still need to sift out proprietary and common software and system files. Don't forget to review e-mail (including attachments), social media accounts (ask the donor to archive them using the vendor-provided software if possible), external media storage, and any cloud-stored data (e.g., Google photos, Dropbox, or backup subscriptions) that may need to be downloaded prior to transfer.

Content in the Making: Working with Creators

Preferably, selection of content will start when meeting with the content creators so you can learn more about the material itself and have the opportunity to help them learn how to make their content usable throughout time. Few people realize that how they create materials, name them, and store them can make all the difference between total or partial loss and the ability to access and share their creation for years to come. The author firmly believes that educating creators is a primary duty for educators, librarians, and archivists, to increase the likelihood of valuable content surviving the next few years. The Shakespeares of today are creating their masterpieces in proprietary file formats and losing them on their hard drives because they can't find them, and they are also losing valuable content because it is not properly backed up or stored. Simple guidelines may include the following:

Use open-source software (e.g., OpenOffice) instead of proprietary software (e.g., Microsoft Word or Excel) whenever possible.

If saving files in PDF form, choose the option to save as PDF/A-1a or PDF/A-1b whenever possible. Even if one of these does not capture all the relevant information in your document (e.g., embedded software), capturing a version of your file in this archival format will ensure that at least part of your work will survive throughout the years.

Create images, audio, and video in the highest quality form available and save in uncompressed formats or by using lossless compression so that content is retained.

Use short descriptive file names with no spaces and no punctuation except for hyphens and underscores. This will help you locate the content on your hard drive and make it usable on different types of operating systems.

If saving multiple versions, use a consistent method of documenting the version number, for instance, adding a date (in the form YYYY-MM-DD) or version number (v1, v2, v3) to the file name.

Whenever possible, embed metadata into the file. Crucial information includes the date of creation, creator name, contact information, keywords, and any rights restrictions.

In documenting how people should be allowed to use your work, assign Creative Commons licenses.[41]

If creating databases, export snapshots regularly in CSV format and document carefully what each table and field mean, and how they interrelate.

If creating software, document thoroughly and retain source code and copies of any dependencies. In a plain text file in the top directory, include information about the operating systems (and versions) on which this software works, dependencies, necessary organization of files, system requirements, installation and use directions, and, of course, a rights statement (ideally from Creative Commons[42]).

Use regular backup and storage systems, preferably multiple copies in different locations. Do not depend on external media like DVDs, CDs, and flash drives, as they are not preservation media and will lose bits of data.

Develop a systematic method of organizing your files and folders. Librarians and archivists can be helpful in providing suggestions on how best to organize content based on the needs of the creator.

As media and systems change, be sure to move your stored content to newer media. Many creative endeavors have been lost to the march of time, as they were stored on floppy disks, zip disks, Jaz drives, and other external devices or outmoded operating systems.

As you work with content creators, become familiar with their creations, potential audiences for their work, and the possible need for long-term preservation and curation involved. Many will be developing work under copyright or for sale, but even so, most of these guidelines are still helpful. Multiple research institutions have had to develop digital forensic services to try to help creators access older content that is no longer in current formats or on current media. By assisting creators in taking proactive steps to prepare their materials for future use, you are doing a service to the creators, digital curators, and researchers and educators in the future.

KEY POINTS

1. Begin by clarifying your focus areas for collection based upon the strategic priorities of your institution and the needs of your designated community.
2. Working within those focuses areas, identify the possible content for inclusion, and begin to build an inventory.
3. When possible and feasible, engage experts in the particular field to identify the content most valuable for inclusion.
4. Select from within the identified options based upon available and expected resources, expertise, rights and privacy restrictions, your ability to access the content, and the quality and value of the material.
5. During capture, protect both your computer and the original content.
6. Collect metadata and context during receipt.
7. Be aware and respectful of rights restrictions and privacy concerns.
8. Educate creators so they can help ensure content will survive long enough to be preserved and curated for the future.

6

What Foundational Work Will Prepare Content for Preservation and Access?

Everything in life tends toward chaos. Untended fields become tangled underbrush and sprouting forests. Closets and garages become confused catchall piles of forgotten belongings. Gardens fill with weeds and suffer the depredations of insects and hungry animals. Throughout time, metal rusts, machines degrade, material rots, and food decomposes. Those of us who do not exercise, eat right, and get enough sleep will find our quality of life dissolving as we speed our own destruction through a simple failure to pay attention. Without careful tending, we cannot retain order, structure, and usability in ourselves or our surroundings. It can be argued that this is what librarians and archivists do for information: We create order, structure, and usability out of chaos and seek to manage this effectively throughout time.

For digital content, we must first create order by collecting and organizing information about the content and its structure into standardized form for reuse. Without standardization, we have no order to build upon and cannot impose useful organization of content or leverage automation processes to effectively handle tremendous quantities of information. But the information about the content is not the only part that needs to be managed effectively; the digital content itself poses its own challenges. Dependent upon whether we want to migrate or emulate, we need to standardize the content itself or else collect and document the software and hardware necessary to manage long-term emulation (to "simulate" the original access experience, regardless of changes in hardware and software throughout time). Even if all we choose to do is "encapsulate" our content for later recreation, standardizing and ordering our materials and information will be necessary to facilitate the unearthing and recomposition of materials. In this chapter, we'll dive into the various types of information, or "metadata," about the content, including options available, and then discuss the options for stabilizing and ordering the content itself.

Once digital content is identified and selected (as described in chapter 5), workflows for effective management content can be described as having four basic stages: ingest, process, preserve, and access.[1] Chapter 7 discusses preservation (which was not covered by AIMS), and chapter 8 covers access. Both preservation and access are dependent upon the effectiveness of the ingest and processing described in this chapter.

RIGHTS ISSUES

Rights issues should be first and foremost when considering management of digital content on a long-term basis. There are several kinds of rights that may impact the decision to select or retain digital content, and the rights may change throughout time. For example, intellectual property (IP) rights extend for periods of time to protect content and must be documented and tracked. Content under IP protection can include patents, trademarks, copyrights, industrial designs, inventions, literary works, and trade secrets. If the material you are considering is under IP protection for the next fifty years, the cost of preservation and management of that content must be weighed against the expected value of the content to users in fifty years, when perhaps it may be made accessible.

Hirtle et al. provide an excellent review of copyright for digitization[2] that is broadly applicable. Be aware that ownership of content is not the same as having IP rights to the content. For example, many people have copies of books, movies, and digital music in their homes that they have purchased; yet, the IP rights to that content belong to creators and publishers. It is illegal for one of these content "owners" to publish the content, sell it, or use it for profit. It is also illegal to make and preserve copies without permission from the copyright owner.[3] Any use of the content beyond that allowed by purchase requires additional permissions.

The U.S. Copyright Office makes available a freely accessible database where it is possible to look up a copyright (after 1978) by title, name, keyword, and so forth.[4] Content for which the copyright owner cannot be identified or is not locatable is called an "orphan work." The challenges for how to manage orphan works are daunting, as the uncertainty of potential legal action can prevent effective use or even preservation of content.[5] Although the European Union offers a database for looking up orphan works,[6] the author has been unable to locate a similar service in the United States. The Orphan Works Act of 2008 was suggested to reduce the risk of use of orphan works, but the bill did not become law.[7] In 2009, the Society of American Archivists released a statement of best practices,[8] which recommended a "diligent search" and complete documentation of the search to protect those seeking to use or preserve orphan works. This appears to be the most practical approach, should you be unable to determine the IP owner.

IP rights are not the only concern. Other rights we need to document include allowed uses for the content and any restrictions on access. If we cannot obtain the rights to provide access to our users, there is no point in preserving the content. And if indeed we want to preserve it, we will need the rights to make copies, generate versions of the content for web access, create descriptions so it can be found, and even the rights to preserve the content. If this content is undergoing digitization, we also need to obtain the rights to digitize from the IP rights owner. It is because of this broad collage of rights issues that most cultural heritage institutions have focused primarily on collecting material that is not under copyright. Unfortunately, this precludes preservation of a tremendous amount of our cultural history, which is not being preserved by publishers and IP owners.

To address this challenge, a few brave leaders have begun to collect content that is particularly at risk of obsolescence, for example, audiovisual materials, regardless of the possible consequences. Notable in this regard are the efforts of the Internet Archive, which seeks to capture snapshots of key web content and collect cultural heritage materials whenever possible. Their policy[9] is to "take down" any content to which the IP owners object, and thus far they have been successful.[10] Nonetheless, whenever possible, IP owners should be contacted and requests for permissions made.

How should rights information be captured and stored? Rights management standards are still evolving, so this is not completely clarified. Some materials may be assigned Creative Commons licenses to clarify the allowed uses of the content. A creative approach to documenting the known status of the copyright and reuse privileges of content is taken by Europeana (the European Union digital platform for cultural heritage materials) and the Digital Public Library of America, at RightsStatements.org.[11] At the time of this writing, they offer 12 rights statements varying from "In Copyright" to "No Copyright" to "Copyright Undetermined."[12] If the selected rights statement link is placed alone in a metadata field designated for this purpose, it is possible to make these links "actionable" by software to help prevent unauthorized use of the material. This means that software can be written to use the specified information to prevent downloads, copying, or other unauthorized use. At present, this is the best available method for documenting the rights for access and use, and is recommended, as well as the addition of a user-friendly entry in a separate metadata field to explain the rights statement for human readers. The Society of American Archivists provides guidelines on selecting appropriate options at RightsStatements.org and points out that these are only needed if Creative Commons licenses do not exist for the content in question.[13]

For other types of rights, we need additional methods of documentation. Thus far, the broadest and most complex method of encoding rights has been developed by the W3C (World Wide Web Consortium), and it is called the Open Digital Rights Language.[14] If technical capabilities are available, this standardized encoding of who may do what, when with this content would be ideal; however, for those with limited technical resources, it may suffice to create a document for IP owners to sign and date, to keep on file before accepting the digital content. Here is an example, developed by the author:

The Donor grants [*your institution name here*] and its agents the right to:

- Digitize all submitted content and create derivative representations for web access
- Reproduce and distribute reprints or derivative representations for noncommercial scholarly purposes
- Augment or create metadata to enhance accessibility and management of content
- Electronically view, present, and display the full digital content to others, including providing open access via the web

Check one of the following:

___ The Donor hereby warrants that she/he is the intellectual property rights holder of all content deposited and has the legal right to make this authorization. The Donor also represents that the content does not, to the best of his/her knowledge, infringe on or violate any rights of others.

___ See attached sheets for listing of the intellectual property rights holders for content deposited. For the sections for which the Donor is listed as the rights holder, the Donor hereby warrants that she/he is the intellectual property rights to said content and has the legal right to make this authorization. The Donor also represents that said content does not, to the best of his/her knowledge, infringe on or violate any rights of others.

This authorization constitutes a nonexclusive, perpetual license, and the Donor retains all other rights to this content to which she/he, as an intellectual property rights holder, is entitled.

Additional space and instructions should be provided for the donor to specify any exceptions, and name, contact information, and date must be clearly provided, along with the signature of the donor (and, ideally, that of a witness). This document should be stored with the digital content in the preservation repository and a copy kept on file in a centralized location with other legal agreements. If possible, have your legal department or an IP lawyer review your document prior to use.

DESCRIPTIVE METADATA

The point of digital preservation and curation is to ensure continued long-term access to digital content. Yet, that access cannot be provided without effective description. Descriptive metadata is the information that describes the content, and it is necessary to provide context and to enable search and retrieval. Without descriptive metadata, which can be indexed by search engines or linked into a browse option, locating the content may be impossible. No matter how great the video is that one uploads to the web, if there is no title, description, or keywords to describe it, no one will locate the video to view it.

Descriptive metadata is key to discovery, and to use it most effectively, one must plan in advance just how users should be able to find your content. For example, if the intent is to develop a collection of online materials, organization of those materials into categories for browsing would require incorporating those standardized browse terms into the descriptive metadata. Without a “key” that will connect like materials, indexing services cannot identify the materials that should be grouped together. Few organizations have the resources to remediate metadata later to provide better findability. For practical purposes, it is most effective to consult with the target user group (when possible) and plan the access points and methods before creating the descriptive metadata. One cross-institutional archaeological project the author was involved in failed to do this, and stakeholders were frustrated to find that users could not effectively retrieve similar content, because some was described with the word “rock,” while other content was described with the word “stone.” Not all search systems are capable of supporting synonym disambiguation, so be sure to clarify any content-specific terms that need to be consistent, as well as the categories (and their definitions) prior to beginning description.

For the same reason (effective search and retrieval), other forms of controlled vocabulary are extremely important to incorporate into description fields. Ideally, leverage existing standardized controlled vocabularies; there is no point in recreating the wheel. Particularly important fields for controlled entries are subjects, names, genres, dates, and locations. Appropriate subjects and genres may be drawn from the Library of Congress Subject Headings (LCSH),[15] Faceted Application of Subject Terminology (FAST),[16] the Art and Architecture Thesaurus (AAT),[17] or the Thesaurus for Graphic Materials (TGM).[18] Controlled values for names may be specified in the Library of Congress Name Authority File (NAF)[19] or the Virtual International Authority File (VIAF),[20] which assigns a specific URL[21] to each name for disambiguation (this URL should be included in the metadata record).

Location values may be drawn from the Thesaurus of Geographic Names (TGN)[22]; if latitude and longitudinal information is available, capture that as well in a separate field, in a standardized format, as this will provide the data necessary to locate the source of the content on a map (“georeferencing”). If the materials you are working with are primarily geographical (e.g., maps), consider supplementing the descriptive metadata with fields from the Geography Markup Language[23] to standardize how the information is captured. If you are working with information specific to a research field and need extensive, reusable metadata appropriate to that discipline, seek out the correct standard and engage experts in the field in implementation. The DDI (Data Document Initiative) standard is used for describing data generated by observation or surveys in the social, behavioral, economic, and health sciences.[24] EML (Ecological Metadata Language)[25] is used for ecological datasets and their content. Life sciences are likely to use Darwin Core.[26] Geospatial metadata standards include the Content Standard for Digital Geospatial Metadata (CSDGM) and the International Standards Organization (ISO) 191xx series of standards.[27] Mathematical and scientific content on the web may require encoding in MathML.[28] Standardization ensures both usability and reusability.

How should this descriptive metadata be stored? Most delivery systems are changed out or transformed every few years, and each delivery system has its own restrictions on what metadata can be entered and in what form. Additionally, even when exports are provided in appropriate encodings, the information exported may not contain all the information that has been input, and even Unicode[29] information may be lost. For effective digital curation, the best possible metadata record should be stored in Unicode (UTF-8[30] is preferred) in a standardized form of XML (eXtensible Markup Language) with the digital content (see sidebar); extracts from this can be used for delivery systems or linked data applications as appropriate. If the descriptive information is stored in a database, export the database in CSV[31] (comma-separated values, using quotes to clarify text that contains commas) or tab-delimited form (also called TSV,[32] tab separated values) to store with the digital content. Be sure to provide extensive data dictionary documentation about each field, how tables relate, and dependencies between fields.

Databases also change throughout time, and the metadata should be stored in a nonproprietary format with the digital content. Each digital object and its corresponding metadata record(s) should contain the same system-wide unique identifier to avoid confusion. The author will not soon forget starting a new job and finding that more than 30 collections of content there did not have identifiers, which would allow matching up metadata to archival files. To prevent chaos and later headaches, keep the metadata with the files, and assign locally unique identifiers with which files are named and included in the corresponding metadata records. For example, if the digital content comes from manuscript collection 3045 (MSS 3045), the finding aid for that collection might be named mss3045.ead.xml and should contain that identifier (mss3045) in the file as well. The filenames for the digital content that comes from that collection may begin with mss3045, for instance, mss3045_1 .tif, mss3045_2.tif, indicating the first and second items in the collection. If an item (e.g., a letter) has multiple pages, simply add another segment to the identifier to indicate the page numbering: mss3045_3_1.tif, mss3045_3_2.tif (indicating the third item in this collection, with two pages). If descriptive metadata has been captured for this third item, it should contain "mss3045_3" in an identifier field and would be best named with the identifier as well, for example, mss3045_3.xml.

Since it is likely that at some point, the locally unique identifiers may need to be used for reference (as "ID") within XML files, it is highly recommended to follow the required syntax rules for XML Names.[33] (As stated in the XML standard, "Values of type ID must match the Name production."[34]) For simplicity's sake and to avoid problems with different operat-

More Info: What Is XML?

XML stands for eXtensible Markup Language; it was designed to both store and transport data in a form that is readable by both humans and machines.[35] The name of each field is contained in brackets and is usually written in camel case, for instance, <titleInfo> or <physicalDescription>; these are usually called “tags.” The contents of the field follow and are followed by another set of brackets, containing the field name again, but this time prefaced by a backslash to indicate that this is the end: <digitalOrigin>reformatted digital</digitalOrigin>. (Examples used in this sidebar come from MODS metadata.[36])

Fields may be embedded in other fields, but the open and close tags must line up. For example, the following is valid XML:

```
<titleInfo><title>Letter from Henry L. Abbot to Fred Jones, 1904-02-11</title></titleInfo>
```

However, if the title close tag is not inside the titleInfo tags, then it is no longer valid XML:

```
<titleInfo><title>Letter from Henry L. Abbot to Fred Jones, 1904-02-11</titleInfo></title>
```

Thus, it is important to ensure that XML is “well-formed,” which means that the tags are correctly embedded and every opening tag has a matching close tag, and the top of the file includes a line that indicates which version of XML is used and what encoding. For example, <?xml version=“1.0” encoding=“UTF-8”?> indicates that version 1.0 of XML is in the file, and the contents are encoded in UTF-8 (“Unicode Transformation Format,” using eight-bit blocks to represent each character).[37]

Attributes may be included in a field tag to provide additional information about the contents of the field. For example, the attributes in these two language fields clarify what form of information should be included:

```
<languageTerm type=“code” authority=“iso639-2b”>eng</languageTerm>
<languageTerm type=“text”>English</languageTerm>
```

The first entry specifies that the field contains a code, rather than a text entry, and that the code comes from the International Organization for Standardization set of codes specified in ISO 639.2 (bibliographic).[38]

The rules that specify what fields may be used and must be used, how the fields may be organized, and what type of information may be included in the fields and their attributes are called XML schemas. By recording these rules in machine-readable form and including the link to the schema (called a “namespace”) in the top of the XML file, computers (and humans) can locate the rules and understand how to decode and use the information in the file.[39] Every metadata standard has its own schema. For example, the MODS (Metadata Object Description Schema) metadata schema specifies that the <title> field must be within the <titleInfo> field, and that the <title> field may contain text, numbers, and punctuation. The schema can be used to validate the file and create a file from scratch. By specifying the link to the schema when creating a new XML file, some software systems (e.g., XMLSpy) will provide helpful prompts as to which fields are allowed where the mouse cursor is located. XML files should always be validated against their schema(s) to ensure that they meet the constraints necessary to make the files reusable and so they will properly display.

The extraction of fields from XML files and translation of them to other fields or web display is normally performed using XSLT (XSL Transformations).[40] This is a rule-based system, and many software systems use this to translate XML files to HTML for web delivery. It’s also useful for extracting fields from one type of metadata to another. For example, to extract a simple <titleInfo><title>Letter from Henry L. Abbot to Fred Jones, 1904-02-11</title></titleInfo> from a MODS metadata file and use it to create a <title> field in an unqualified Dublin Core metadata record, the following in an XSLT file would suffice:

```
<xsl:for-each select=“mods:titleInfo”>
<title>
<xsl:value-of select=“mods:title/text()”/>
</title>
</xsl:for-each>
```

The “for-each” statement ensures that every titleInfo entry is processed separately. The “value-of” statement specifies that the text within the title field be extracted. The <title> and </title> tags will be output into the file being created, with the contents of the title between those tags, resulting in <title>Letter from Henry L. Abbot to Fred Jones, 1904-02-11</title>. XSLT is a handy language for librarians and archivists to master, both for web delivery and display modifications, and to extract and transform metadata to be reused in new or different standards.

ing systems, use ASCII letters (American Standard Code for Information Interchange)[41]: In other words, use those on the standard U.S. English keyboard, numbers, underscores, and hyphens,[42] and do not exceed 244 characters.[43] If uncertain about how to develop locally unique identifiers that will scale for your holdings, consider the types of information collected thus far and the types under consideration for the foreseeable future. Does it make sense to include genre or type as an indicator to assist in grouping materials, or is related content of mixed types? Also consider the number of contributing sources and whether it would be useful to group content according to where it originated. If so, it may be sufficient to begin each identifier with a letter or two to identify the source and complete the identifier with the date and time in a standardized form. This will not preclude duplication of content, however, which may be a concern.

Some file naming schemes use the first few characters of the checksum for identification, but this may not be effective for grouping multifile documents like software and videos. In the author’s current institution, content is named according to source, general type, originating collection, and then sequentially by intellectual item. Additional segments in the file name indicate sequential page or subpage numbering. For example, a file named w2_14_12_3.tif would be from the Williams repository (w) rare books (the assigned code for this type of material is 2), the 14th collection, the 12th intellectual item, and page 3. A second version of this file would be named w2_14_12_3.v2.tif, and the OCR (optical character recognition) text extraction file would be named w2_14_12_3.ocr.txt and so forth. Again, this doesn’t prevent an image of the same content existing elsewhere in the holdings, but in the digital preservation perspective, every digital object is unique, however related to other objects. This type of file naming allows scripts and humans alike to easily identify where the file belongs, what metadata applies to it, and how to order the online presentation, with both simplicity and clarity.

At some point, assignment of globally unique persistent identifiers[44] may be appropriate, and these can be used locally as well; however, registration of these identifiers generally comes at a cost. Common current types available include the Digital Object Identifier System (DOI),[45] Archival Resource Keys (ARK),[46] Persistent URL (PURL),[47] and Handles.[48] Granularity of assignment of global identifiers should be considered: Will users need to be able to cite a portion of the data, or will retrieval of that portion need to be managed by computer? If so, a single identifier for a grouping of content related to a publication may be insufficient. Moreover, be aware that delivery software may claim to assign “unique identifiers” to content, but those identifiers are no longer viable after the digital material is moved to other software. Ideally, the unique identifiers, once assigned, will never change. This means that throughout time, someone must be responsible for updating the redirects handled by these identifier registries to ensure that the links will always forward users to the correct content, even though that content has been moved from system to system, server to server, or even to other institutions.

Managing persistent URLs is a commitment not to be taken lightly. Yet, they are crucial to continued usability. What use is a citation or reference that no longer points to the original content? We have all experienced clicking on links that go nowhere. This is a major and growing problem on the web[49] and one that the support of persistent URLs exists to ameliorate, at least for scholarly materials.

How should descriptive metadata be encoded? The most common descriptive metadata encodings for item-level description records at this writing are Dublin Core[50] or qualified Dublin Core, which includes terms,[51] and MODS, which was developed from MARC[52] records, specifically to provide enhanced features for describing digital content. Eventually, it is hoped (by those who want to better leverage the web for search and discovery) that all bibliographic descriptions will be shifted to linked data formats,[53] the foremost storage encodings of which are Resource Description Framework (RDF)[54] and JavaScript Object Notation (JSON).[55] It is beyond the scope of this work to explain how best to generate linked data, but tutorials are freely available.[56] In linked data, each entity or relationship has a single authoritative link assigned to it, where centralized information can reside to disambiguate this entity or relationship from all others. The purpose of linked data is to provide standards for publishing content on the web that will enable users to locate related concepts and related content.[57] This is the most discriminating, most standardized, and most complex method of storing descriptive metadata but also the most flexible and reusable.

The problem with simplicity is that one can never effectively translate it to more complex records later, nor can one effectively document how each field is used in each record. If Unqualified Dublin Core is selected (one of the simplest types of records), then to be practical, limit how each field is used (and thus limit what information is selected), and document this clearly and publicly, as well as in the preservation repository. For example, the “date” field may be used in multiple ways, as discussed in chapter 1. To ensure these records are sortable, the form of the date should be in the ISO 8601 format,[58] for instance, 1987-04-03 for April 3, 1987; however, to ensure searchability, a second date field might contain “April 3, 1987.” The date to be represented should be documented clearly: Is it the date created? The date described? The date of publication? The date of issuance? When using a simplistic metadata standard, document and clarify how each field will be used and ensure consistency throughout the content. Consistency in all fields is highly recommended across all holdings, as this will improve the user experience, support migration to newer systems throughout time, and facilitate metadata transformations in the future.

If the digital content is specific to a particular discipline, consider incorporating appropriate metadata standards from that discipline, as mentioned earlier for geospatial materials. It is possible to generate standardized XML records that contain multiple types of schemas (or structured plans) if a single metadata file is desired. To do this, include namespaces for each schema in the top of the record,[59] with different assigned prefixes, which then are used for the specific elements (an example of mixing Dublin Core elements, Dublin Core terms, and Open Digital Rights Language schemas is available to demonstrate[60]). The simpler and more practical method is to just have separate metadata files if needed for these types of content and ensure that every record has a basic standardized metadata record with the same encoding. One practice for managing these multiple files is to name all metadata files for the digital content unique identifier and include the type of file prior to the prefix, for example, "w2_14_12.mods .xml" and "w2_14_12.odrl.xml." This clarifies the type of metadata standard without opening the files and allows for storage of multiple types of descriptive files in the same directory.

A key consideration for practicality purposes is the level of granularity of description. In the archival profession, where content sometimes comes in by the truckload, the "More Product, Less Process"[61] approach recommends a minimal approach: "The goal should be to maximize the accessibility of collection materials to users. . . . What is the least we can do to get the job done in a way that is adequate to user needs, now and in the future?"[62] In a 2016 survey of methods used for intake and management of digital content,[63] some respondents noted that they treated entire hard drives or disk images as collections in and of themselves.[64] This may be a practical approach when inundated with incoming digital content.

If it is not practical to provide item or folder-level description, then at minimum, create descriptive information at the collection level. For this approach, the Encoded Archival Description (EAD)[65] is likely the best option for a description encoding standard, as it was developed for collections of content and supports several layers of hierarchy. If it is feasible to logically or intellectually subdivide the "collection" into "series" by topical area or types of content, the EAD supports description at each level. For example, if the creator of the original content clearly kept images in one directory ("My Pictures") and personal writings in another ("My Documents\Writings"), then each of those could become a "series" within the finding aid, with an appropriate description. Subseries and other levels, with or without additional descriptions, are supported as well. This may be sufficient to meet user needs and facilitate access to the content without the use of extensive resources in preparation.

Consider what is more important for your organization: quick and simple or more complex, flexible, and reusable—or somewhere in between? Weigh carefully the quantity of content to be described, the granularity of description, the human and technical resources at your disposal, the knowledge and capabilities of the workers, and the amount of time available for the work. Try to strike a practical balance that will provide effective access on a long-term basis.

A final note about using controlled vocabularies: Be sure to note the authority type (e.g., LCSH or VIAF) in the record if possible. EAD (for collections) and MODS (Metadata Object Description Schema) support this documentation in an "authority" attribute, but to do this in Dublin Core would require using qualifiers, or terms, and XML namespace references.[66] This will enable software extraction and transformation at a much more specific level and can be used to later support the development of linked data implementations. Even when you do not have the resources to currently develop complex descriptive information, be as specific as possible. Plan ahead for when perhaps in the future, the information you generate now can be leveraged to improve access. The type of descriptive metadata standard selected can greatly simplify the work of description but at the expense of improved usability and future transformation support. It is recommended that librarians and archivists do the following:

1. Decide on the level of granularity that is practical and sufficient.
2. Select the metadata schema that most clearly defines the fields that are appropriate to capture for the content in question.

3. Whenever possible, use the same metadata schema for all content in one's holdings.
4. Assign unique identifiers to the digital content.
5. Leverage controlled vocabularies whenever possible.
6. Store the metadata in standardized, nonproprietary forms, using UTF-8 Unicode.

Standardization and clear definitions of metadata fields are key to future use and access.

STRUCTURAL METADATA

Structural metadata is the information that captures the structure of multifile materials to ensure that the correct organization of content is retained to enable future use. For example, many books are digitized one page at a time, including the front and back covers, and sometimes the spine. For a 300-page book that will contain about 310 archival quality images (counting title page, table of contents, publisher information, and so on), if OCR is used to extract text from those images for indexing, that adds more than 300 more digital files to the item. A metadata record (or catalog record) for the item adds another file. If thumbnails and web-size images are generated from each of the archival captures for online use, that could be 620 more images. If technical metadata is generated when testing the archival files to ensure they are properly encoded, that would add at least 310 more files to the item. At this point, one could easily have more than 1,540 files to manage for a single intellectual item: a book. To provide online display effectively, the visual images must be delivered in sequential order. To provide online indexing effectively, each OCR page needs to be associated with the correct image so that a search on a specific term will take users to the correct page. For effective long-term management, each of the technical files must be associated with the correct image as well. How does one effectively manage so many files?

There are three basic approaches taken to answer this question. One approach is to use structural metadata, which documents how these files relate to one another and in what order. The second approach is to leverage file system organization to physically organize the content appropriately. The third approach is to simplify the original, at a loss of information, to reduce the complexity of long-term management.

The most well-known standard for structural metadata encoding is METS (Metadata Encoding Transmission Standard),[67] usually for text and images, and there are less well-known ones used particularly for moving images, for instance, MPEG-21 DIDL[68] (Digital Item Declaration Language) and, more recently, MXF (Material eXchange Format).[69] Each of these standards are complex and require technical expertise and often the services of a computer programmer to implement, particularly in quantity. When choosing this option, review the variety of file organizations necessary in your holdings and develop a hand-encoded example to document each variation. Use these to inform the programmer and sift the types of content. When the example is complete, it is advisable to generalize each example into a "profile" of the rules followed and publish these publicly so that others will be able to decode your files. Refer to the namespace where you will publish the profiles in each file so the files can be "validated" by software to ensure they fit the designated profile. Also, if using METS, it is advisable to register each profile with the METS registry.[70] Others can learn from your efforts, just as you can learn from others by reviewing the registry before you begin. As an additional note, descriptive metadata tends to change throughout time; thus, if the standard allows it (as does METS), it is advisable to link to the most current external descriptive metadata files that may be replaced as needed and only embed the original description within the METS wrapper.

The second approach is to leverage the file system and file naming conventions to reflect the hierarchy of the materials, grouping of files, and sequence for delivery. This simple, straightforward approach is low-tech, without loss of information. Described by the author[71] and demonstrated at the University of Alabama,[72] each item has a unique directory, reflected by the file name; pages each have their own appropriately sequential directories within the item directory. Deconstruction of the file name by replacing underscores with slashes (forward or backward, depending on the file system) defines the directory so that scripts can easily relocate files and verify that content is in the correct directory. All metadata resides at the level to which it applies, grouped in metadata subdirectories with the type of metadata included prior to the extension (e.g., ".mods.xml," ".ead.xml," ".mix.xml," etc.). All transcript material resides in transcript subdirectories at the level to which they apply, named for the image file they reflect, with ".ocr.txt" extensions for OCR files and simply ".txt" for transcriptions.

Another elegant method of leveraging the file system for storage is the "Pairtree" method.[73] In this specification, file names are "cleaned" by converting any characters that are problematic to file systems, and the cleaned string is split into a file system path, with two characters for the name of each subsequent subdirectory, until the last directory has either one or two characters. This system helps prevent the risk of having too many files in any directory, which may slow the system's ability to locate or process content. This may be an issue for "collections" that have potentially millions of digital items.

The third primary method for managing this complexity of files is to reduce the complexity, although this method generally loses information. An example would be to generate PDF

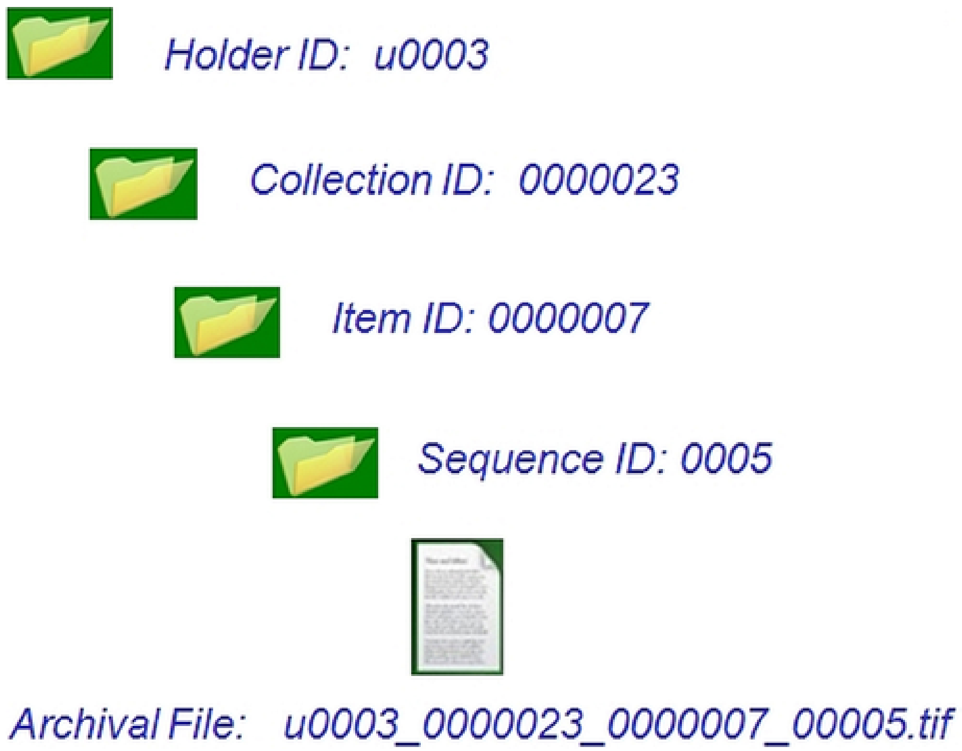


Figure 6.1. File System Organization Reflected in File Name

files (preferably PDF/A-1a[74]) from the archival files, compressing them in the process (which usually loses information), and then discarding the archival files. Some PDF creation products will extract OCR from images and embed it as the PDF is made. A thumbnail might still be generated for the cover image and a metadata record still created, but now only one technical metadata record needs to be generated, as the other files are compiled into a single file of type PDF, which is then used for web delivery, with copies stored for preservation and backup.

TECHNICAL METADATA

Types

There are numerous kinds of information that are considered technical metadata, including the size of the file, location of the file, file extension, date of creation, date of modification, and more. Just how crucial some of this information is depends upon the needs of the user in the future. Consider this: If this document were the equivalent of a handwritten letter by Abraham Lincoln, how important would it be to the researcher to know when it was written, on what kind of material, with what kind of ink? Technical metadata provides clues to the meaning and importance of the document, as well as the person who created it. Technical metadata can also verify that the digital document being examined is indeed the original. How?

One of the most common and most important types of technical metadata is the checksum. A checksum is a unique signature derived from the encoding of the file, the testing of which can verify whether the file has changed. The most commonly used checksums for file storage are MD5 (“Message Digest 5”) and SHA1 (“Secure Hash Algorithm 1”).[75] It is important to capture the checksum as soon as the file is created or obtained, as this will be key to determining whether the file is the original, throughout time. The checksum should be verified every time the file is moved, and prior to each backup if possible, to avoid overwriting good backups with bad ones. Even the best of storage media suffers from bit loss and failures,[76] so certainty that the file is the original and has not become corrupt can easily be provided only via checksum verification. The checksum verification is also called a “fixity check.” Wikipedia offers a quick overview of many of the options available for capturing and testing checksums.[77]

There are other types of technical metadata that are quite important. For example, the encoding of the file may require repair before it makes sense to spend time and energy preserving it for long-term use. Even well-trained professional digitization teams can generate badly encoded archival files, which the author discovered to her dismay. The author and her team documented known TIF (Tagged Image File) errors (and their apparent causes) after testing several thousand stored archival files at the University of Alabama in 2014.[78] Any software that opens or renames the file can modify or damage it, and some software products produce badly encoded files. Thus, the primary purpose of checking the technical metadata before storing content is to ensure that the files are correctly encoded and of a format that is viable for long-term archiving. Although the file extension would seem to provide enough information, it cannot be trusted; files are often inadvertently saved with the wrong extension. Nor can the file extension tell you what version of encoding was used, so you can’t know how recent the encoding is and therefore how viable it is for preservation purposes. What is meant by encoding?

Each type of file format has specific documented instructions as to what information must be where within the file and what is allowed within that information entry, much as a MARC record has specific instructions as to where to put the title or the author name, and AACR2 or RDA provide restrictions as to how that information may be entered. As software is updated, the file formats update as well, including new modifications; this is a newer version of the file format. Some types of files can be generated by multiple software systems, and some of those systems do a better job than others. For example, there are many ways to generate a PDF file. Not all of the different methods result in a properly encoded PDF file. Although Adobe Acrobat is quite generous in ignoring errors in current PDF formats, throughout time the ability to compensate for errors diminishes as more and more new types of errors are introduced by various PDF generators.

To ensure that the file can continue to be opened, it is important to make sure it has been properly encoded. Technical metadata can provide that information, as well as much more. Some of the information provided may be crucial to reuse of the file, particularly if emulation is desired. For instance, technical metadata can specify the software used to generate the file, the lighting (for images), the speed and direction (for audio recording), and much more. Ideally, technical metadata will be extracted and stored in appropriate standards (to be discussed later in this chapter) for later automation use, and some information may be stored in a database for easy access to assist in management. Thinking ahead, if at some point your institution will have hundreds of thousands of a specific file type stored, wouldn't it be helpful if software could locate the versions that need updating and perhaps perform the transformation? Standard encodings and consistency can make this possible, and for the long term, it is the most practical approach.

Tools and Standards

Two types of information are important to look for when testing a file to determine if it's worth preserving: whether the file is well formed and whether it is valid. As defined by JHOVE[79] developer Gary McGath, "In general, a document that isn't well-formed is one that's unusable; it will break an application that makes the assumptions in the spec. An invalid document is one with errors that reduce its functionality."[80] By "spec," McGath means the specification for the encoding of the file: the rules that explain what information should be stored where and in what form. It simply doesn't make sense to spend resources on preserving content that isn't well-formed and valid; and clearly, "well-formed" is the most important quality.

At the time of this writing, there are four primary tools for extracting technical metadata and testing file types, which are described here. Software systems (e.g., BitCurator[81] and Archivematica[82]) that perform testing and analysis (prior to upload of content to a preservation repository) incorporate one or more of these tools and may also use others. Unfortunately, new file types are developed each year, and there are no tools that effectively identify and analyze all of them. This is another reason why it is recommended to limit the file types included in your scope of content management and migrate ("normalize") incoming files to those chosen file types. If you are unable to verify that a file is well-formed or even what type of software is needed to open it, preservation of the file is likely of limited value, and effective digital curation is potentially impossible.

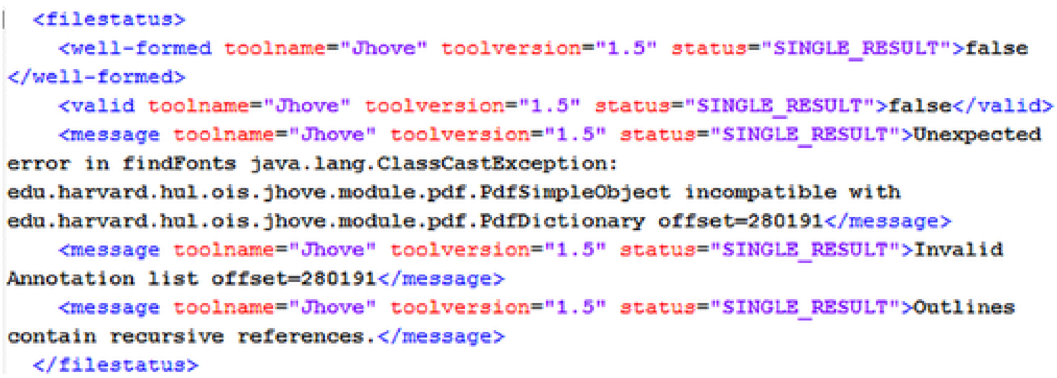
DROID (Digital Record Object IDentification)[83] is a file format identification tool developed by the National Archives (United Kingdom) that works hand-in-hand with their PRONOM file format registry[84] to identify files primarily by extension and the initial ("magic") characters; it does not verify whether the file encoding is well-formed or valid. But this tool does provide a CSV report of files with such basic information as location, size, checksum, media or MIME

(Multipurpose Internet Mail Extensions) type,[85] format, and format version. This is handy for collecting a "map" of content on a hard drive and locating duplicate files, and can form the basis of an initial inventory of content to assist with selection, particularly if complete hard drive images are not to be provided as the point of access for users.

JHOVE (JSTOR/Harvard Object Validation Environment)[86] performs format identification and validation, and provides a list of known characteristics (e.g., size, location, media type, checksum, and more). JHOVE can be run both in GUI (Graphical User Interface), as well as on command-line (which makes it easy to incorporate into scripts for traversing multiple files and/or directories). The software can be instructed to use different modules to better identify specific types of files, and the output can be in text or XML. The ability to check well-formedness and validity of files makes this tool invaluable, for a specific set of types of file formats. Also extremely valuable is JHOVE's ability to generate technical metadata in community standard schemas appropriate to the format. Schemas included are MIX[87] for images, AES[88] for audio, TextMD[89] (for plain text, XML, HTML, and related files), EBUCore[90] (which supports semantic web implementations for broadcasting), and DocumentMD[91] (for such formatted text as Microsoft Word and PDF). The information output in these schema-compliant files may not contain all the information desired for preservation, but to date this is the best method of compiling technical metadata for these formats. At this writing, the JHOVE standard modules cover these formats: AIFF and WAVE (for audio); ASCII, HTML, UTF-8, and XML (for text); GIF, TIFF, JPEG2000, and JPEG (for images); and PDF. But the support for PDF Module cannot test conformance to the PDF/A to the extent required by the pending standard.[92]

FITS (File Information Tool Set)[93] was also developed by Harvard to incorporate and extend the capabilities of JHOVE to effectively test a wider variety of file formats. In the default FITS output file, the type of file and version will be found in the "identification" section, and the crucial information as to validation and well-formedness may be found in the "filestatus" section. For example, the lines in figure 6.2 indicate that the file needs repair or redigitization.

Due to the complexity of TIFF and PDF files, however, the current version of FITS (1.0.7) may not include the JHOVE output in the filestatus section. To ensure this information is available so you can check this, modify the configuration file for FITS (fits.xml) to set <display -tool-output> to "true" to view the JHOVE status field, which will be found in the toolOutput section of the results file. For example: <status>Well-Formed and valid</status> indicates that the tested file is good. On the other hand, <status>Not well-formed</status>" indicates a severe problem with the file.



```
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">false
</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">false</valid>
    <message toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">Unexpected
error in findFonts java.lang.ClassCastException:
edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject incompatible with
edu.harvard.hul.ois.jhove.module.pdf.PdfDictionary offset=280191</message>
    <message toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">Invalid
Annotation list offset=280191</message>
    <message toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">Outlines
contain recursive references.</message>
  </filestatus>
```

Figure 6.2. FITS Test of PDF That Is Not Well-Formed or Valid

At the time of this writing, FITS includes both JHOVE and DROID, and supplements them with Apache Tika, MediaInfo, ExifTool, FFIdent, Windows File Utility, and the National Library of New Zealand Metadata Extractor. FITS will use all the tools at its disposal to test the file in question and report on any disagreement between tools. For example, if ExifTool reports that the file is a TIFF version 6 and JHOVE reports the file is TIFF version 5, this will be noted in the XML output. Some tools are more effective with certain file formats than others. For example, MediaInfo was designed to identify and extract technical metadata from video files. FITS can be run on command line (useful for setting up scripts to test all the files in one or more directories) and can also be set up as a web service. Since FITS incorporates both JHOVE and DROID, it is, in many ways, the most useful single software available for testing files. Even so, it is not yet capable of effectively testing and validating the archival form of PDF, known as PDF/A.

VeraPDF[94] is the software currently in development for testing and validation of PDF/A and is thus an excellent partner for FITS in the manager's toolkit. VeraPDF can be accessed through a GUI, via command line, or as a web service. This is the only tool in this listing that changes the source file, so managers who seek to retain a copy of the unchanged original file are warned to not test the original. Current modifications include removal of the PDF/A flag if the file does not conform to the standard and repair of broken embedded descriptive metadata. The software generates an extensive report of which encoding specifications are met by the file and which are not, as well as any fixes that have been applied. Managers may only want to review the top of the report for the "validationResult," as this is the most crucial information, specifying the type of PDF and whether it is compliant with the expected encoding.

You may notice that there are no current technical metadata standards mentioned here for databases, software, or video. Because most databases are proprietary, exports are recommended to be in CSV or XML, and recommendations for documenting the fields and tables vary with the discipline (see the descriptive metadata section earlier in this chapter). Software varies tremendously, as do the variety of languages in which it is written, and the methods for documenting are still in development.[95] Video and moving pictures standards are also in development, and as the encoding includes images, text, and audio packaged together, the contents can be extremely complex; however, a new metadata standard for video (International Press Telecommunications Council "IPTC Video Metadata Hub") was released in late 2016, which includes descriptive, rights, administrative, and technical metadata.[96] At the time of this writing, it has not yet been incorporated into openly available software for testing video files.

Recommendations at this time are to test archival files and review the results to ensure the files are well-formed and valid, and repair those that are not; then store the technical metadata generated in the standard most appropriate to the type of file format. Repairs are best made by opening the file with the creating software (if possible) and then resaving; be aware that this will overwrite or modify existing technical metadata, so keep an original copy of the file if authenticity is important. At the time of this writing, FITS is the best overall tool for file testing, as it incorporates and builds upon JHOVE and DROID to provide broader testing capabilities, an analysis of validity and well-formedness, and exports in appropriate technical metadata standards; however, for PDF files that fail to conform using the FITS software, test files using VeraPDF to separate out those that validate as PDF/A format, and store at least the synopsis section from the output with the validating files. Remaining PDF files need to be ideally migrated to newer formats, which may involve opening in Adobe Acrobat Pro and saving as PDF/A.[97] Then retest. Be aware that tools and software are constantly updating, so it is necessary to stay abreast of changes and developments in the digital preservation community to inform and update local practices, policies, and procedures. Digital curation is still a work in progress.

ADMINISTRATIVE INFORMATION

Administrative metadata encompasses rights and information (discussed earlier in this chapter) but also includes provenance information, preservation actions (e.g., migration) taken on the content over time, and any other information needed to effectively manage the content on a long-term basis. For example, administrative metadata would include entries in a database of the location of every file with documentation of file format and version, so a manager will be able to identify which files need migrating when (and this is, of course, technical metadata). Provenance information would document the chain of custody and change history of the object or file. Where did it come from? Has it been modified throughout time in any way? If so, by whom and in what ways? In 2013, the W3C developed recommendations (called PROV[98]) for how to encode this information, but it is not yet a standard. Since provenance is a pressing area for managing scientific data effectively, a more recent effort (2016) by the DataONE[99] Cyberinfrastructure Working Group[100] is extending the previous work. Check out the current status of ProvONE[101] if you're curating scientific data.

Provenance clearly overlaps with "preservation actions," which also include documenting when checksums were verified. Administrative information can be confusing and extensive. PREMIS (PREservation Metadata Implementation Strategies)[102] was developed to document administrative metadata, which includes some elements of both technical metadata (checksums, file sizes, etc.) and descriptive metadata (to identify the contents, the access and use permissions, and relationships between objects). But PREMIS does not specify how the information should be stored or encoded; instead, it offers different levels of "conformance" to the standard, which vary from the ability to map stored information to required (and possibly optional) PREMIS fields to standardized encodings already implemented for each file.[103] Should the manager want to encode PREMIS for exchange with other organizations, the most commonly used method incorporates extensions to the METS standard.[104]

Required fields in PREMIS include the type of identifier (e.g., a locally generated one or a DOI) and the value of the identifier, the object category, and the format.[105] Object categories include intellectual entity, representation, file, and bitstream[106] (see figure 6.3). Intellectual entities are the things one might catalog, for instance, a book, an article, or a website (and yes, intellectual entities can contain other intellectual entities). A representation is all the files necessary to construct the intellectual entity; for example, videos are commonly made up of multiple files, as are books (when not created as a single PDF). Bitstreams are data that require additional information (e.g., headers and formatting) to be understood as a file by an operating system.[107] For example, information transmitted via the internet arrives in packets, which, when compiled in the correct manner, become files that an operating system understands. Also, operating systems store information from large files in multiple locations on the hard drive and create a "lookup table" that allows them to collect these bitstreams and reconstruct the complete file in the correct order.

Highly recommended optional fields include the checksum and the checksum type (called messageDigest and messageDigestAlgorithm, e.g., MD5), the creating application name, version and date, the location of the file, the assigned preservation level, and the significant properties. Everything but the significant properties and preservation level can be extracted or generated by FITS. These two qualities bear further discussion.

The preservation level is assigned according to the extent to which the organization commits to preserving the content. No organization has the resources to curate and preserve everything; boundaries must be drawn to effectively make progress. Some file types may be proprietary

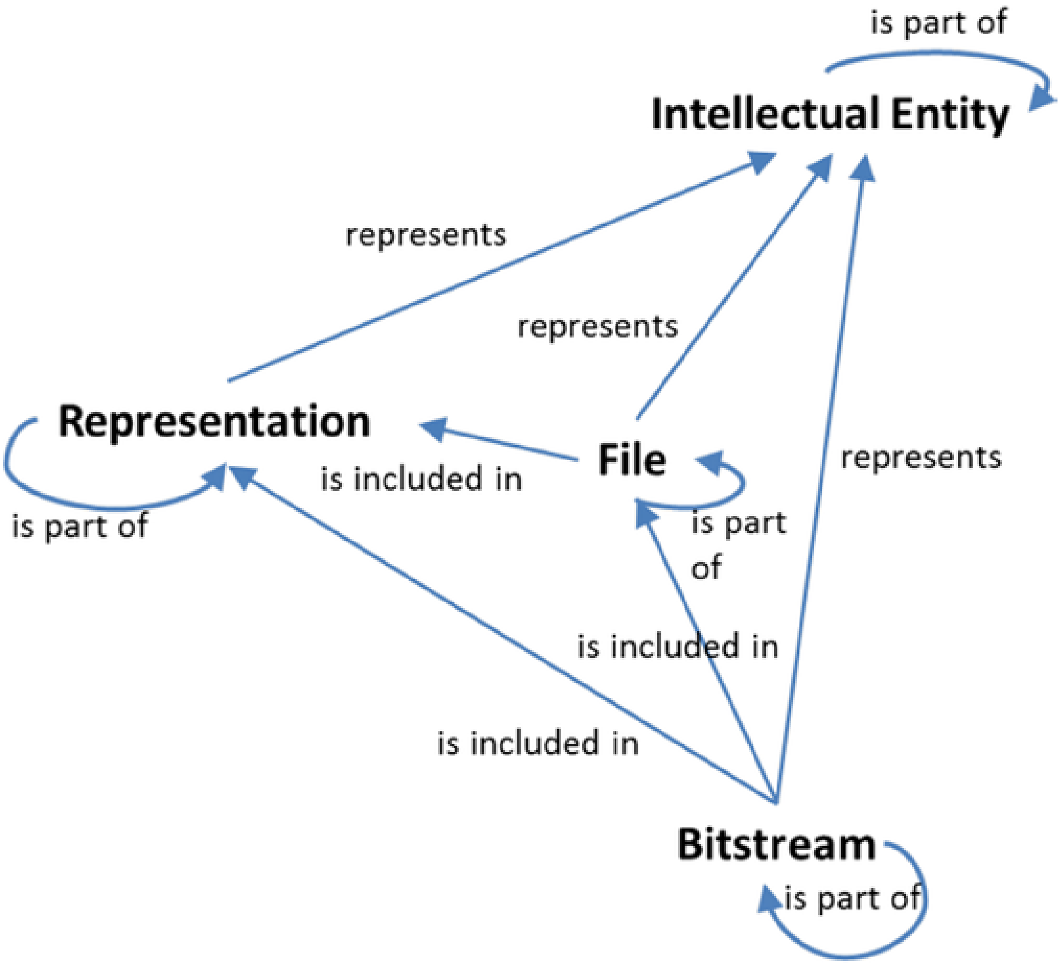


Figure 6.3. PREMIS Object Categories and Their Relationships. ***Reprinted with permission of the PREMIS Editorial Committee, available from http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.***

and perhaps cannot be "normalized" or migrated effectively to archival formats. It may make no sense to retain these files. Some types of content, for example, video and moving images, may be too space-hungry for an organization with limited resources to manage effectively. Some content may be protected by copyright for an extended period, which may not make it worth the risk to retain until it can be managed and accessed effectively; therefore, it makes sense for an organization to develop "levels of preservation" and document what each level means, from temporary bit-level storage to a full commitment to providing effective access to the content throughout time (this concept is explored further in chapters 7 and 8). If this information is stored, the fields in PREMIS that apply are the type of level and its value, the role (which can be used to indicate if this is an aspiration level, partially met, or achieved), the date the level was assigned, and the rationale used for assigning that level. Clearly, the level of preservation assigned, the extent to which that has been achieved, and the rationale may change throughout time.

A practical approach would be to group content by factors that impact the organization's ability to manage it effectively and assign levels to the group. For example, a higher-education research library may commit fully to long-term curation of the content they digitize and their curated electronic theses and dissertations (ETDs), and yet decide to only commit to temporary storage of other content submitted to their institutional repository, unless it is in archival formats. As mentioned previously, some institutions may decide they simply cannot effectively manage the complexity of video or the file sizes for long-term storage. Leading research institutions, notably the University of Michigan,[108] often base their levels of preservation on the file formats, as this is the simplest method to determine what they can effectively manage on a long-term basis.

The significant properties of an object are those that must be retained for the object to be understood and experienced effectively. An image may be of an oil painting, in which case the color quality is a significant property, or an image may be of a book page, in which case the textual information is a significant property. The behavior of a software package would be significant, and the volume and fidelity of a sound recording would be significant. The determination of what is significant is somewhat subjective. Every time content is migrated to new formats or even emulated, it may lose structure, appearance, or behavior.[109] Determining the significant properties in advance will prepare the manager to select the best possible options for future access and can be used to judge the effectiveness of access in newer formats or an emulation system. If the color is wrong in the newer version, or the contents are unreadable, or the software does not behave properly, the significant properties have not been maintained. In PREMIS, the manager determines the significant property type (e.g., content, fidelity, or behavior) and specifies what is appropriate for the value (e.g., textual content only, the audio sampling rate and range, or expected software function). For the most part, significant properties can be selected or identified by type or genre of material and assigned accordingly. The trick for the manager is being able to locate the content by significant property, when the time for migration is near. Thus, it is helpful to store in a (protected) database an indication of the type or genre of material (at minimum), along with the location, assigned preservation level, checksum, file format, and file format version, to assist in effective content management.

CONTENT PREPARATION AND STANDARDIZATION

Imagine having hundreds of different formats of files in your holdings and for each of those formats having multiple versions. For example, it's not uncommon to have PDF versions 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, PDF/A-1a, and PDF/A-1b among collections of electronic theses and dissertations in research libraries. Microsoft Word has 15 versions for Windows dating back to 1989, 14 versions for Macs, and 14 other versions for other operating systems, at the time of this writing.[110] Then start adding on all the other types of file formats and their versions, and you begin to see what a quandary it is just to keep it straight what you have where, much less what you need to do with each format type and version.

Many formats, for instance, Microsoft Word and Excel, are proprietary with closed specifications, which means the encoding is not openly published, as that would give away trade secrets to competitors. When the software undergoes updates, it may or may not continue to provide support for previous versions of the files. If the business that creates the software goes under, it's possible that no other software will ever be able to correctly render the files; therefore, it's in the digital curator's best interest to "normalize" files to openly available formats (e.g., OpenOffice, which is an open-source effort) and limit the types of formats retained (PDF is proprietary, but the specification for it has been made openly available, in the interests of broad adoption).

"Normalization" is migration of file formats to select formats to simplify long-term management and curation. Even if the manager intends to provide access via emulation, the restriction of the types and versions of formats will greatly simplify development and management of the emulation environments. In either case, it is best to keep a copy of the original file, since it is possible that later methods of migration or emulation may improve on current capabilities.

Ideally, particularly if emulation is not intended, the manager will select a single archival format for each type of file. What is an archival format, and how does one decide what is best? An archival format is one that both meets the primary needs for capturing and storing the best quality available of the type of content in question, and which is likely to be accessible and usable for a long time to come. The Library of Congress provides guidelines and specifications that can help in selecting archival formats, considering such sustainability qualities as whether the specification is openly available, how broad adoption is for the format, whether it may be patent-protected, and the extent to which the use of the content has external dependencies.[111] The broader the adoption in the preservation community, the more likely that migration patterns and access software will be developed and sustained.

As pointed out, if the specification is not open and available, the ability to migrate or even access the file in the future is critically impaired, as it is completely dependent upon the company who developed the file format. External dependencies are problematic, as that complicates the number of things that must be managed effectively for continued access to and use of the file. Additional considerations for selecting archival formats include whether it involves compression (which usually results in loss of information) or encryption, contains internal documentation about its creation and basic metadata, and contains internal restraints to prevent use of and access to the content.[112]

According to a recent survey, the most widely used archival formats are TIFF for images, followed by JPEG2000; PDF/A for text, followed by PDF; WAV files for audio; and MP-4 for video.[113] For web archiving, WARC is the most commonly used file format at this time, and the Library of Congress recommends ESRI Shape files for geospatial data.[114] The content manager is advised to stay abreast of changes in the field in terms of what archival formats are currently in favor; monitoring the Library of Congress site on sustainability of digital formats is recommended.[115]

How does one migrate a file? Sometimes it is possible to open a still-accessible file with a form of software that will save as a newer type of file. For example, it may be possible to open an older PDF in Adobe Acrobat Pro and save it as a PDF/A or open an older image type with Adobe Photoshop and save it as a TIFF. This doesn't always work, however. The author has attempted this method of PDF transformation to PDF/A and received such errors as "font not embedded" or "circular references in internal hyperlinks." If the fonts used during file creation are not available on the computer that is attempting the migration, it will not be possible to generate a PDF/A, as one of the requirements is to embed all fonts (avoiding external dependencies). Other requirements of PDF/A include forbidding encryption, JavaScript and executable files, audio and video content, and colorspaces that are not specified in a device-independent manner; therefore, even this simple approach to migration doesn't always work. Even if the migration appears to be successful, content may not have the same appearance as before, and this can be crucial to some types of materials, for example, documents containing mathematical or scientific formulas, tables, and images. Ideally, the person performing the migration will compare an untouched original (never migrate your only copy!) to the new file, page by page.

The simple approach won't work for everything, nor does it scale for larger quantities of files. The method of migration and the software used will vary with the type of file and the source and target formats, and there are efforts underway to develop the best possible migration paths for particularly problematic content, for instance, the ePadd software package[116] for e-mail. A list of available tools for file format migration (and other aspects of digital curation management) is maintained by COPTR (Community-Owned Digital Preservation Tool Registry).[117] Also, a collaborative project between UIUC (University of Illinois at Urbana–Champaign), NCSA (National Center for Supercomputing Applications), NARA (National Archives and Records Administration), and ISDA (Image Spatial Data Analysis) has provided a web interface that enables you to enter your original format and your target format, and it will provide you with a list of software that can perform the migration for you.[118]

Migrations will need to be automated as much as possible for larger quantities of material; however, the likelihood of loss of functionality, usability, appearance, and even content requires managers and staff to review the results of the migration to evaluate the extent to which the migration was successful. Even if all the technical characteristics of the original content are retained, the results may be a phenomenal failure, as documented by the Planets' XCL (eXtensible Characterization Language) Project, which seeks to document the extent to which migrations succeed.[119] Thus, quality control review is crucial to every migration, particularly to evaluate the extent to which (especially) the significant properties of the file are retained. Be aware of this issue when offered migration by any software, particularly if migration is part of the upload process to a preservation system. Without manual review and comparison to an untouched original, you cannot be certain migration retained the significant characteristics of the content—or that it succeeded at all.

The reader is by now far more aware of the challenges involved in preparing content for long-term storage and future access. The quantity and variety of information to be managed, tested, validated, and tracked can be overwhelming. The best solution to prevent chaos is to limit the types of files managed, adopt standards and leverage controlled vocabularies whenever possible, develop and document policies and workflows, and, ideally, learn some form of scripting (if you do not have someone available with that skill already). Scripts are a form of software that do not need to be "compiled" into a machine-readable form. The ability to use a script to "wrap" the use of existing tools is invaluable in automating the management of massive quantities of files (see sidebar). For example, a script can traverse a set of directories one file at a time and use a call to the operating system (called a "system call") to apply a tool to each file to generate a checksum, validate each file, or perform other useful functions.

More Info: A Sample Script

The following is a sample Perl script (also available for download[120]) for using FITS to test the TIFF files in a directory, writing the FITS files to another directory, then testing the FITS to make sure the TIFFs are valid and well-formed. Errors are written to an output file. This script can be run on Windows, Mac, Linux, or Unix. The first line should indicate the path to your installed Perl library. Otherwise, information after a "#" sign indicates a comment to explain what the script is doing.

To use this script, you must have Perl[121] and FITS[122] installed on your computer (you may need to reboot after installation). Then follow these steps:

1. If testing PDFs or TIFF files, open the fits.xml file in the FITS xml directory with a text editor.
2. Change <display-tool-output>false</display-tool-output> to <display-tool-output>true </display-tool-output>

. Save the file.

3. Copy the information below into a text editor and save as “simpleFitsTest.pl” in a specified folder on your system.
4. If on Linux/Unix/Mac, modify the first line in the script to point to your Perl installation executable (after the #!).
5. Whatever your system, modify the $script value with the path to the FITS script on your system (using fits.sh for Linux/Unix/Mac and fits.bat for Windows); if in Windows, use quotes around the path if there are spaces in directory names and use a complete path (e.g., $script = “C:/FITS/fits-1.0.7/fits.bat”).
6. Also modify the $here value with the path to the directory where you have placed this script.
7. In the folder containing simpleFitsTest.pl, create a “Files” directory where you place the files you want to test; also create a FITS directory at the same level, where the FITS files will be written by the script.
8. Place the files you want to test into the Files directory. Remember that the JHOVE part of this will only test validity and well-formedness on AIFF, WAVE, ASCII, HTML, UTF-8, XML, GIF, TIFF, JPEG2000, JPEG, and PDF (not PDF/A). All other files will be reported by this script as “not well-formed and not valid.”
9. If on Linux/Unix/Mac, change permissions on simpleFitsTest.pl to that of an executable (chmod +x simpleFitsTest.pl) or use the command perl simpleFitsTest.pl to run the script instead of typing in simpleFitsTest.pl. In Windows, if you have properly installed Perl, then you need only click on the simpleFitsTest.pl to run it.
10. Run the script and examine the output file to locate which files have problems that need correcting. If none, the output file will be empty, unless the call to the FITS software failed; in that case, there will be an error message with instructions to check the FITS installation and the paths you corrected in the script.

Here’s the script:

```
#!/usr/bin/perl
# set up paths to FITS script, where to put the FITS, where to find the scans, and an error file
$script = “C:/FITS/fits-1.0.7/fits.bat”;
$here = “C:/FITS/”;
$fitsDir = $here.”FITS/”;
$filesDir = $here.”Files/”;
$output = $here.”fitsOutput.txt”;
# open the error file to write to in case we have problems
open (OUT, “>“.$output) or die “can’t write to $output\n”;
# open the Files directory, so we can look through it
opendir(FILES, $filesDir) or die “can’t open $filesDir\n”;
while ($file = readdir(FILES)){
    if ($file =~ /^\./){ next;} # skip dot files, which are hidden system files
# find each file, and capture its name in $id, so we can use it to name the FITS file
    if ($file =~ /(.+)\.[a-z0-9]{2,4}$/){ # this excludes the file extension
    $id = $1;
# put the location of the file in $myloc
    $myloc = $filesDir.$file;
# create a path and file name for the matching FITS file, named for the original file
    $fits = $fitsDir.$id.”.fits.xml”;
# here we use the operating system call to run the FITS script;
# -i flag indicates input file; -o flag indicates output file
      $val = system("$script -i $myloc -o $fits");
# clear these variables of what they had for the last file tested
      undef $wf;
      undef $valid;
# now test the FITS output; is the file well-formed (wf) and valid?
# first we open it
      open (IN, $fits) or die "can't read in $fits\n";
# then we read through it line by line, looking for the entries we need
      while ($line = <IN>){
      chomp $line; # this removes the return at the end of the line
#following pattern match looks for "<well-formed"...">true<"
      if ($line =~ /<well\-formed.*?>true<\//){ $wf = 1;}
# following pattern match looks for "<valid "...">true<"
      elsif ($line =~ /<valid .*?>true<\//){ $valid = 1;}
      elsif ($line =~ /<status>Well-Formed and valid<\/status>/i){
      $wf = 1; $valid = 1;
      }
      }
      close(IN);
# here we output errors into the output file if there's a problem
      if (! $valid && ! $wf){</indent>
      print OUT "$file: NOT VALID and NOT WELL-FORMED\n";
      }
      elsif (!$valid){ print OUT "$file: NOT VALID\n";}
      elsif (!$wf){ print OUT "$file: NOT WELL-FORMED\n";}
      }
      }
close(FILES);
close(OUT);
# then we notify the person running this script of results. If output file has a size, give
   warning.
if (-s $output){ print "\nUh oh. Check the output file for errors!\n";}
else{ print "\nGood job! They all check out!\n";}
sleep(10); # sleep 10 seconds before exiting to make sure the user sees the error message
exit;
```

Generally, one can open a text editor (e.g., NotePad++[123] on Windows or VIM[124] on Linux/Unix), write a script, save the file with the correct permissions and extension, and run it. The most commonly used scripting languages for creating batch processes, metadata extraction, and transformation are Perl,[125] Python,[126] and Unix/Linux shell scripts.[127] Other common languages include PHP,[128] Ruby,[129] and, for metadata transformations, XSLT.[130] Most of these (not the shell scripts or XSLT) can be extended to build graphical user interfaces (GUIs), which make the scripts more user-friendly for nontechnical staff. Perl is fast and excellent for pattern-matching, which makes it useful for locating specific text or correcting file names. Python is much slower but much easier to read, and it retains much of Perl's ability to pattern-match and manipulate files. Unix/Linux shell scripts are perhaps the least readable of all, but they are the fastest, as they communicate most directly with the operating system itself. Online tutorials and resources are readily available to assist the beginning programmer. A simple script that saves hundreds of hours is powerful indeed, so learning a little scripting is a practical approach and an extremely useful tool to have. Anyone who is detail-oriented, analytical, and capable of understanding logical procedures can write software, as the author can attest through experience and several years of training new programmers.

Since each file format is going to require different types of normalization and technical metadata, and may have different significant properties to document and consider, it becomes clear that developing pipeline workflows for each type of format supported will simplify and clarify what needs to happen when and how. When developing these workflows, however, it's important to retain the original context of the content, as this may be key for researchers in the future. Consider how scholars have studied the processes by which great artists and musicians have developed their works or how researchers seeking to recreate and expand on previous studies need to know the details of how those studies were carried out. If the original content is split into multiple workflows and separated, how does the content manager retain the original context for future researchers? Much of this information is captured during the initial inventory of incoming content.

For example, the differences in creation dates on different versions of a musical score may provide useful information to a scholar, perhaps indicating how long the composer needed to develop the version changes. The documentation that specific audio files were found in the folders where the scores were created may indicate that the composer either was inspired by these other musical pieces or generating some of her own while testing and developing her current masterpiece. The ability to bring to future scholars the original order of the content and information about its organization may be key to future discoveries. This is something archivists know well, as they try to capture and document content in the original order as generated by the content creator. Even if it is not feasible to recreate the original arrangement of the content for users during access, retaining information about each file's relation to the others when collected and incorporating that into the descriptive and/or administrative metadata will serve the same purpose.

The Mellon-funded AIMS (An Inter-Institutional Model for Stewardship) Project[131] provides a framework for the intake and access for born-digital content: collection development, accessioning, arrangement and description, and access.[132] Within that framework, the tasks described in this chapter fall within accessioning. But how does one organize these tasks into ordered pipeline workflows? The author recommends mapping out the steps for each type of content and developing and documenting policies and procedures for each step. Wikis are particularly helpful for this, as they support uploads of graphs, images, and text documents, and also allow easy modifications and access for staff. Begin with the method of content intake, and document everything from the use of antivirus and malware to protect the local computer to the use of write-blockers to protect the incoming content. The basic steps are as follows:

1. Safely access the content.
2. Inventory the content, collecting a map of what is where, size and types of files and creation dates, and known rights issues.
3. Identify and locate types of software that will be needed to access content.
4. Collect checksums on original content and copy content to a secure location for processing.
5. Verify the checksums and store them for future reference.

6. Assign locally unique identifiers.
7. Test file formats and extract technical and other administrative metadata; store appropriately.
8. Normalize a copy of the content to select archival formats.
9. Generate or create appropriate descriptive and rights metadata, and store appropriately.

Managers seeking more advanced software that combines tools and provides more comprehensive capabilities for technical metadata testing, documentation, and handling quantities of materials for ingest are advised to consider BitCurator[133] and Archivematica. BitCurator wraps multiple open-source software packages, was developed to handle especially incoming hard drives, and has deep and broad forensic analysis capabilities. This includes the ability to locate information in the files that may need to be protected, for instance, social security numbers and bank account information (personally identifiable information, "PII"). BitCurator comes packed inside a virtual machine (with a Linux operating system) to be installed on Windows using VirtualBox. Screencast tutorials[134] and a user group[135] are available for assistance; and two-day hands-on courses in BitCurator use are provided via the Society of American Archivists.[136] Archivematica[137] also bundles multiple open-source tools, expects incoming content to be placed in a single directory (preferably on an off-line computer with a Linux operating system), and provides an interface to select the desired tools to apply. If any of the content in the directory fails to meet the necessary testing requirements, the content must be tweaked or removed, and the processing starts again; however, once content survives the test, it should be ready for ingestion into a preservation system (pending review of any migrations). Some preservation systems will extract technical metadata from the files upon upload (but rarely check to see if the files are valid or well-formed) and even perform migrations automatically. Nonetheless, as noted, migrated content needs to be compared to untouched originals, and the original should also be preserved for three reasons: to document authenticity, to compare against migrated versions, and for the potential improvement in migration or emulation tools in the future.

KEY POINTS

1. There are multiple types of information ("metadata") that should be gathered about the digital content: rights, description, structure, technical, and administrative.
2. To avoid confusion and to match up metadata to the files, assign unique identifiers for each digital object.
3. Use appropriate standards for encoding metadata and leverage controlled vocabularies whenever possible.
4. Files should be tested to ensure they are valid and well-formed according to the format specification; FITS and VeraPDF are recommended tools for this purpose.
5. Collect checksums on files as soon as they come in; then verify those checksums prior to backups to ensure files have not changed and to verify authenticity.
6. Assign levels of preservation support to content to simplify management on a long-term basis.
7. For easy reference and effective content management, store in a database the type of material, location, assigned preservation level, checksum, file format, and file format version.

8. Use archival formats whenever possible, and normalize incoming files to those selected formats to minimize preservation costs on a long-term basis.
9. Compare migrated or normalized files to the original to verify that significant properties of the content remain.
10. Retain a copy of the untouched original to compare against migrated versions, to verify authenticity, and in preparation for later, potentially better, migration or emulation capabilities.

7

Storage, Protection, and Monitoring: What Do I Do?

As the years go by, digital curators must successfully navigate the constantly changing software and hardware environment to ensure that the content they manage continues to be usable. Curators must also be aware that digital content can begin to degrade on media almost as soon as it is written, depending upon the quality of the media, the temperature, the humidity, and other stresses. Many people don't realize that digital content is inherently fragile—far more fragile than paper. Loss of bits of content has been documented even on the best of servers,[1] and many of us have accidentally deleted or modified digital files without even realizing it. How do we safely store this content, protect it, and monitor it to make sure it hasn't changed and can still be accessed? And what kind of planning and protection do we need in place to ensure our hard work will not be in vain?

KEY CONSIDERATIONS

The primary considerations we need to address are safety and security, continuity, and protection of authenticity and the ability to access and use the content. Safety and security are basic; if we cannot ensure these, there will be no content to access and use, nothing to hand off to others to support (continuity), and no way to protect the authenticity of the originals. If we cannot continue to verify the authenticity of what we have stored, the usefulness of the content also comes into question. And if the content we store can no longer be opened or used, there's really no point in continuing to preserve it. Moreover, funding may fail, and of course we won't always be the ones to manage content intended for many years of use. We may find it uncomfortable to consider, but a practical and responsible manager will plan ahead. Every organization has turnover and changes in priorities and funding, and few last forever. Even institutions that last many decades may undergo merges or other transitions that can put our preserved digital content at risk; therefore, continuity planning is also a key consideration.

Safety and Security

Safety and security encompass multiple challenges beyond those of continuity, which we'll address separately. A good synopsis of the possible threats includes the following:

media, hardware, and software failure
media, hardware, and software obsolescence
failure of network services and transmission errors
operator error
natural disaster
external and internal attacks[2]

This may seem overwhelming, but like any mountain, it can be climbed one step at a time. First we need to understand what the threats are; then in the next section, we'll discuss how to address them.

Any type of storage system can fail, for multiple reasons. Materials wear out, component parts may be of poor quality, mishandling is common, and many types of media and hardware are sensitive to changes in temperature. For example, a study on the longevity of CD-ROM storage showed that as the temperature and relative humidity increase, the life expectancy of the media decreases.[3] Many of us have suffered hard drive failures, flash drives that do not work, CDs or DVDs that skip, and the "blue screen of death" on our desktop computers. Nothing lasts forever, and every type of storage has weaknesses and vulnerabilities.

Software also is known to fail. All software is written with a narrow view of how it should be used. Programmers develop software based on what they have been told to expect, but no one can foresee all possibilities. As situations and needs change, the likelihood increases that the software will fail to meet the changing expectations of users. When software is used for something it isn't programmed to handle effectively, it will fail. Often this is called a "bug," and updates to the software ("patches") will be written to address the oversight. Also, all software is built on top of existing "libraries" of code, which have to be updated regularly as hardware improves and the libraries themselves are improved. If the underlying libraries change, the software may fail, as it has not been updated to accommodate the underlying support systems. These "libraries" are often considered to be part of the operating system, for example, Microsoft Windows, MacOS, or Linux. As the operating systems are updated or changed, the software also needs to be updated or changed. Many types of software have to interact effectively with other software packages, and an update in one will often require changes in the other. Also, as hardware configurations are updated or modified, the ability for the software to communicate effectively with the underlying hardware may be lost or impeded. This is one of the reasons software has to be regularly updated.

Have you had more than one computer in your life? Odds are you have worked with laptops, tablets, desktops, and handheld devices. As new hardware becomes common, older hardware becomes obsolete, and the software for the older hardware becomes obsolete as well. How many times have you updated your operating system? Many of us remember Windows XP and Windows Vista, for example. From 2000 to 2015, nine different versions of Microsoft Windows[4] and thirteen different versions of the Macintosh operating system were released.[5] As newer versions of software arrive, older ones are no longer supported and become obsolete.

The external media on which we store content becomes obsolete as well. Floppy disks and zip disks are almost never seen anymore; VHS and Beta videotapes are now obsolete, and CDs and DVDs are on their way out as well. Many computers available today do not have CD or

DVD drives, and some do not even have USB ports, expecting users to obtain everything they need via the internet. When hardware is no longer built to access external media, users are at risk of never again retrieving the digital content they have stored.

As we become more and more dependent upon the internet, the impact of transmission errors and failure of network services grow. This can be as simple as a local router failing, or it can be as major as a power outage or denial of service attack taking out a major network hub, impacting hundreds of thousands of customers. As it pertains to digital preservation and curation, you may lose your ability to access the software necessary to open your files, the emulation services you needed to make your content usable, or the cloud storage where your precious digital content resides. Oftentimes, tape backups are performed across the network, so local outages and interruptions can damage your backups or destroy them. Medium to large files being sent across any network are divided into in "packets," which are then reassembled at the destination into the "original" files. The larger the file, the more packets generated, and the greater the risk that one or more packets will be lost or damaged along the way. This is the reason why it is so important to verify checksums on content after transmission across a network or via any media.

Operator error is not something anyone likes to talk about. We all make mistakes, but owning up to them in public can be humiliating for not only ourselves, but also our institutions. It's simple and easy to accidentally drag content or folders somewhere it shouldn't go, overwrite good files with others that have the same file name, or delete content without intention. Sometimes our errors are simply oversights. At a previous employer, after a coworker inadvertently lost a directory of archival content, the author discovered, to her horror, that the backup systems did not have copies. Why? Because the person overseeing backups never checked to find out how effectively the software had been set up: Only the first two layers of directories were backed up, and the subdirectories had been missed. Mishaps tend to occur in clusters, and losses due to operator error are more common than you would expect.

Natural disasters are what we normally think of when going through disaster planning. In 2011, a tornado swept through a half-mile swath of Tuscaloosa, barely missing the buildings that housed the digital preservation holdings and the original archival content that had been digitized. For several days, thousands of people were without electricity, telephone service, or access to running water; the author spent the first forty-eight hours following the tornado just trying to locate coworkers and employees. Even more damaging was Hurricane Katrina, which caused catastrophic damage to the southeast in 2005. Earthquakes, fires, storms, and floods are capable of wiping out infrastructure we depend upon daily. While we cannot foresee all possible natural disasters, there are commonalities we can indeed plan for and for which we can be prepared.

Last (but not least) are the risks of both external and internal attacks. Disgruntled employees can wreak havoc in a preservation system quite easily if they have access. If there is network access to the digital storage, hackers can gain entry and damage or destroy content, or hold it for ransom. One of the more recent types of attacks are called "ransomware," in which a hacker locks down access to the computer system until the owner pays the demanded fee, which is usually exorbitant.

Continuity

There are four types of continuity for which we need to plan: continuity of storage, continuity of the storage systems themselves, continuity of employees assigned to manage the content, and continuity of business operations. These are all tied to funding, but also planning and effective oversight. If we are seeking to store content on a long-term basis, we must ensure that funding will continue to be available to pay for that storage and that it is, in fact, obtained and used, which is a role for administrators. Whatever form of storage we choose will change throughout time or need to be replaced with newer hardware. Transferring content from one storage system to another to provide updated hardware is called "refreshing."

Someone needs to manage the refreshing, the monitoring, and other curation tasks. Both reorganization and staff turnover can impact the continuity of employees assigned to manage or perform digital curation functions; without clear prioritization and oversight, necessary tasks may be forgotten, and content can be put at risk or completely lost. And then we need to think about the possibility that our organization may no longer be able to tend our digital content. Business continuity is when you plan for other organizations to take over the management of the preserved digital content and curation tasks in the event of loss of funding, institutional failure, or any other situation in which the organization managing the content can no longer effectively do so.

Protection of Authenticity

Those who access your stored digital content in the future will want to be assured that it is either the original digital document or the original undamaged representation of the analog item that was digitized. This is particularly important for historical documents like letters written by Abraham Lincoln or Martin Luther King, or photographs taken of notable events, for instance, the shooting of President Kennedy or George Wallace's infamous "stand in the schoolhouse door" in an attempt to prevent desegregation in Alabama. Our ability to prove the authenticity of a digital object is dependent upon checksum verifications (to prove it is unchanged) and documentation of the provenance, or all those who have owned or managed the object and all events that may have impacted the object. It is possible to change a file simply by opening it and moving it from one device to another. Some changes may be intentional, for example, migration to newer formats so the file can be opened with today's hardware and software. Such events should be monitored and documented to provide a clear provenance back to the origin of the digital object.

Protection of the Ability to Access and Use the Content

Keeping the bits of the digital file intact and protected is vital but insufficient for digital curation. Curators must also monitor files to ensure they are still accessible with today's hardware and software, and that when files are opened, the software used still translates the digital content in a form that makes it usable. If opening the file displays unreadable or uninterpretable content, the file is not usable. Since hardware and software are constantly changing, this task of monitoring is crucial. If the curator allows multiple upgrades to user systems to pass before checking whether the stored content is obsolete, it may be too late to retrieve the preserved material. Curators who are dependent upon emulation must ensure that the available methods are still able to render the stored content. If the curators are also curating the emulation software systems, there is the likelihood that support software must be modified to adapt to the new hardware and operating systems. Only curators who have retained the original hardware and software required for the usability of their content may be untouched by the continual change in our computing environment. The price they pay instead is that of obtaining and maintaining the different versions of hardware and software needed by the digital content they collect (ideally offline to prevent security risks).

OPTIONS AND METHODS

We all want to make sure we're doing everything we feasibly can to address these risks to our precious digital content. Concern in the digital preservation community led to the development of a 2002 report clarifying what a "trusted digital repository" (TDR) should encompass.[6] In January 2007, a group of four preservation institutions agreed on ten basic characteristics for digital preservation repositories. In brief, they are that the repository does the following:

commits to maintaining the content for their identified target audiences
demonstrates the ability to fulfill its commitment
meets required contractual and legal rights and responsibilities
has an effective set of policies
obtains digital content based on the criteria that match its commitments and abilities
maintains the integrity, authenticity, and usability of the content
creates and maintains the appropriate metadata, including that of provenance
fulfills the required dissemination of content
has a strategic preservation planning and action program
has the technical infrastructure necessary to carry out these tasks, including security[7]

This set of guidelines was followed by the publication of a checklist[8] by which the repository's compliance could be judged, and in 2012, a variation of this checklist[9] became an International Organization for Standardization standard.[10] At the time of this writing, only six repositories have been certified by the Center for Research Libraries as having TDRs: Canadiana.org, Chronopolis, CLOCKSS, Hathitrust, Portico, and Scholars Portal.[11] Storing your content in an existing TDR would be ideal; however, digital curators are recommended to use the checklist as a guide and for comparison to identify areas in which they need to improve their own methods.

If you want a simpler set of guidelines against which you can compare your efforts, consider the National Digital Stewardship Alliance Levels of Digital Preservation (see table 7.1).[12] They create four progressive stages of implementation for storage, file fixity checking, security, metadata, and file formats ranging from simple basics to full support. The following is a brief and practical overview of what best practice guidelines involve.

The best methods of storage involve multiple copies in geographically diverse locations, preferably on different types of hardware and software systems. If disaster occurs in any one location, other copies still exist, and if hardware or software systems fail, not all copies will be damaged or lost. One of the lowest-cost methods of implementing this type of storage is by participating in a community that is implementing the Stanford-developed LOCKSS software, which stands for "Lots of Copies Keep Stuff Safe."[13] LOCKSS is the underlying software used by the CLOCKSS Trusted Digital Repository, which stores e-journals and e-books. In this software, copies of each institution's content are stored by all the nodes in the network, and the software checks for any changes in the content provided by the owning institution. Examples of private LOCKSS networks for cultural heritage materials include the Alabama

Table 7.1. NDSA Levels of Preservation

	Level One *(Protect your data)*	*Level Two* *(Know your data)*	*Level Three* *(Monitor your data)*	*Level Four* *(Fix your data)*
Storage and geographic location	Have two complete copies that are not collocated. For data coming in on heterogeneous media (optical disks, hard drives, floppies), get the digital content off the medium and into your storage system.	Have three complete copies. Have at least one copy in a different geographic location. Document your storage system(s) and storage media, and what you need to use them.	Have at least one copy in a geographic location with a different disaster threat. Start an obsolescence monitoring process for your storage system(s) and media.	Have copies in geographic locations with different disaster threats. Have a comprehensive plan in place that will keep files and metadata on accessible media or systems.
File fixity and data integrity	Check fixity upon ingest if it has been provided with the content. Create fixity info if it wasn't provided with the content.	Check fixity on all ingests. Use write blockers when working with original media. Virus check high-risk content.	Check fixity on transformative acts. Check fixity of sample files/media at fixed intervals. Maintain logs of fixity info; supply audit on demand. Have the ability to detect corrupt data. Virus check all content.	Check fixity of all content in response to specific events or activities. Have the ability to replace corrupted data.
Information security	Identify who has read, write, move, and delete authorization for individual files. Restrict who has those authorizations to individual files.		Maintain logs of who has accessed individual files.	Maintain logs of who performed what actions on files, including deletions and preservation actions. Perform an audit of logs.
Metadata	Inventory content and its storage location. Ensure backup and noncollocation of inventory.	Store administrative metadata. Store transformative metadata and log events.	Store standard technical and descriptive metadata.	Store standard preservation metadata.
File formats	Encourage use of a limited set of known and open file formats and codecs.	Inventory file formats in use.	Validate files against their file formats. Monitor file format obsolescence threats.	Perform format migrations, emulation, and similar activities.

Source: "NDSA Levels of Preservation," working draft 3.0, 2017, reprinted with permission of the National Digital Stewardship Alliance (https://wiki.diglib.org/NDSA).

Digital Preservation Network,[14] which lowers costs by having partners involved in the process itself, and the MetaArchive Cooperative.[15] Each of these groups openly shares their models so that they may be emulated elsewhere, and both are open to new members.

If you are not yet ready to participate in such an effort, consider periodically capturing copies on external hard drives (with checksums) and trading those hard drives with other geographically distant institutions in a shared effort to provide some of the same security. Other options include storing multiple copies in the "cloud," which basically means your content is spread across multiple computers elsewhere. If you choose this option, be sure to encrypt your content, and read the small print closely. Normally the cost of retrieval or any change to content is high, and usually the company assumes no liability for loss or damage. Additionally, should funding fail, your content will certainly be lost. Another consideration is that of network outages and disasters impacting the service, during which your content will not be accessible. For example, the five-hour Amazon S3 Service disruption[16] in February 2017, took down a sizable portion of the internet.[17]

Storage systems themselves need not be pricey or fast, since access to stored archival content does not need to be speedy and, in fact, should be restricted to specified personnel with secure access. RAID (redundant array of independent disks) storage is recommended, as this provides more internal security, but since anything impacting the system will likely damage all of it, this does not replace the need for multiple copies. Backup systems should be in place, with rotation off-site. Timing of backups should be far enough apart for the curator to verify checksums before good backups are overwritten by bad ones, particularly if additional copies are not stored elsewhere. And, of course, the room where the storage systems are located should have controlled cooling, with generator backups in case of failure, and it should also be secured and available to only specified personnel. To increase security, the storage servers can be kept offline. Incoming content should be tested for malware and viruses, and the servers should be protected from hackers, as well as local threats, mishandling, and inappropriate access.

Continuity

Planning for continuity is an administrative function and one that is heavily dependent upon both the commitment of the organization to digital curation and the assurance of continued dedicated funding to support the process. The best way to manage this is to align digital curation efforts with the mission of the organization and build curation goals and objectives into the strategic plan. Employee curation work expectations will then fall within the mission of the organization and contribute to the success of the strategic plan. Similarly, this supports the assignment of a budget line of funding to support storage, backups, security, and curation tasks and oversight. Managers can estimate the amount of storage they will need to add and how quickly those additions will need to take place by monitoring the quantities of incoming or digitized content throughout time. Typical hardware support agreements last only a few years, and the end date for support can be used to plan for regular replacements ("refreshing").

As mentioned earlier, the employees who have access to stored digital content should be limited and their access secure. Each employee who serves a crucial role in the digital curation processes should also have at least one backup person trained to fill in as needed, and all tasks should be adequately documented, and the documentation kept up to date. To do this, it helps to outline the key tasks for each of the basic steps of digital curation and document (and effectively communicate) who is responsible for each task or role, and who is assigned as a backup person. As a reminder, in chapter 3 these tasks are summarized as follows:

identification
creating/receiving
appraisal and selection
ingestion (into a repository system)
preservation actions
transformations
storage and protection
management and planning
providing access

Once assignments are made, develop policies for what is expected and procedures for how tasks should be carried out. Then arrange for training of backup personnel and set up regular "checkpoints" where staff implementation of expected work is reviewed and verified periodically. Plan ahead for how to manage staff turnover and loss and failure due to disgruntled employees or staff who suddenly are no longer available.

It's also important to have planned out how to recover from disasters of all kinds and at every level. If you stop to consider what the impact of various types of disasters might be, you will find that they have commonalities, or scenarios, such as these:

software and/or hardware loss
loss of online deliverables
loss of content undergoing transformation, intake, or curation procedures
loss of archival content
loss of geographically distant copies
loss of backups
loss of metadata
loss of crucial staffing
loss of staff working area or equipment
network or power outage

For each scenario, outline what needs to be done, who is responsible for the tasks, and who the backup people are for the assigned tasks and roles. Discuss the plan with everyone concerned, and ask for feedback; don't be surprised if you failed to consider some important aspect. By communicating effectively and engaging stakeholders in disaster planning, you ensure the buy-in necessary to both map out viable options and involve all participants in ensuring a successful outcome. Make sure the finalized plan is documented where everyone can access it, even when the power is out and the network is down. Exchange emergency contact information with everyone involved, and, ideally, issue a small printed card with a list of contacts that can be kept in a wallet or purse. Ask each person involved to add the contacts to their cell phones. Then revisit the plan regularly (at least once a year) to update it, and update the contact list twice a year or any time staff turnover occurs.

By working through this process, you will become more aware of risks that you need to ward against. For example, the author's team came to realize that our key documentation was on a wiki and that we needed to regularly capture copies of it that could be accessed and used offline (and that serve as backups). Moreover, our working area (a Windows "share" drive) only provided "shadow" backup copies of files that had been recently accessed: Crucial documents we needed to access infrequently had no backups at all, nor did archival files in collections that had stalled during digitization or processing. The difficulty we had in contacting one another in the wake of the recent Tuscaloosa tornado pointed out the necessity of having current and alternative phone numbers available in hard copy and on cell phones. When stakeholders, curators, and managers become aware of the weaknesses and hazards in workflow processes, and help plan for the different issues that can arise, your team and your digital curation system will be stronger and more viable.

A succession plan would be what you need if your organization fails, loses funding, or changes strategic directions that would preclude continued support for digital curation and preservation functions (this is called "business continuity"). If you are involved in a LOCKSS partnership or another arrangement where multiple copies of archival content are stored at different institutions, the simplest method is to develop succession agreements between those who already have copies of the archival files and metadata. Failing this, seek out a larger institution whose user community would value the content your organization is collecting and ask for their consideration. As a last resort, consider donating all content to the Internet Archive. Although the file formats and metadata are not ideal, this will ensure that multiple copies remain, and remain accessible, as long as the Internet Archive is able to obtain sufficient funding. Be aware, however, that all content shared will most likely be automatically placed online, which may not be appropriate for copyrighted material or content containing information that should be redacted (e.g., social security numbers).

When a potential successor has been identified, additional information needs to be shared about the organization of content, what the value of the content is, the targeted user community, and any necessary information on how to access content or reconstruct multifile objects or collections. Succession plans should be reviewed by a lawyer if possible, clearly documented, and signed by the involved parties (and witnesses). How and when content handoffs occur should also be documented clearly. Since administrations, content, and methods change throughout time, the plan should be revisited at least annually and updated as needed.

Protection of Authenticity

As mentioned previously, checksums should be verified on digital content every time it is moved and prior to each backup. Any system failure can case loss or corruption of files (see figure 7.1), and transferring content from media to media or across the internet risks loss or damage. If the checksum remains the same from the time the file was first created until the present day, you can be assured that it has not been modified. This is crucial in documenting the authenticity of the file; but we must also document its passage from the custody of the creator to the current holding institution. Has it changed hands? Did the checksum remain the same during the transition? Throughout time we can expect for content to move from one repository or storage system to another, as new and better methods of storing content are developed and administrative partnerships wax and wane. Also, as time passes, the content may have to be migrated to newer formats that are compatible with current hardware and software. The extent to which this new form of content reflects the original will be heavily dependent upon the extent to which the "significant properties" of the original are well represented.

For example, if the original was an image of a hand-painted statue and the new version of the content is a three-dimensional representation that distorts the appearance or compromises

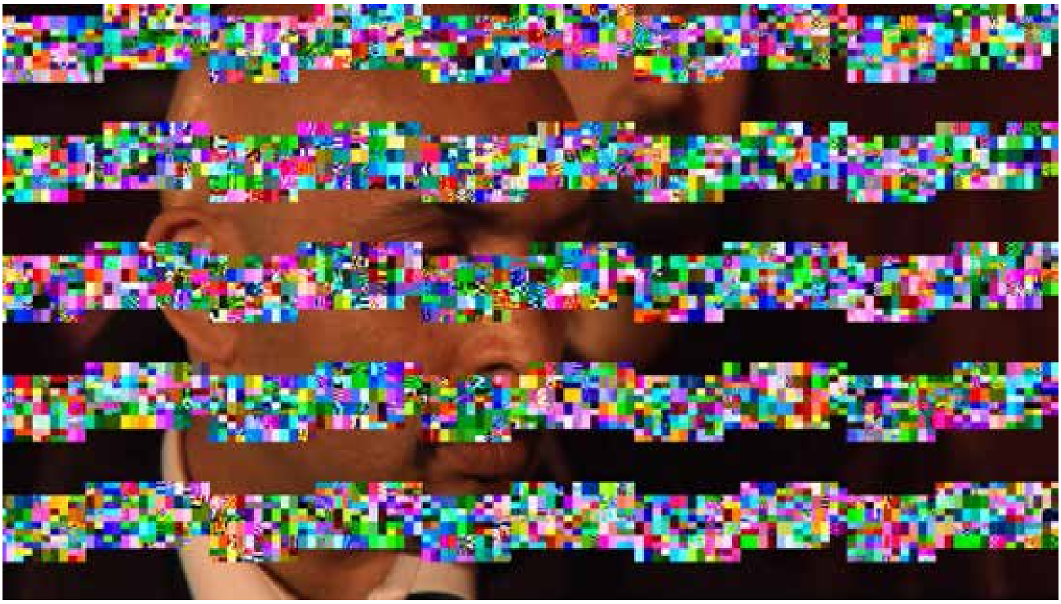


Figure 7.1. Example of a Corrupted File. ***Reprinted with permission of Djonny Chen, available from https://i1.creativecow.net/u/146058/sample_damaged_file.jpg.***

the original colors, the new version is not a viable migration. All migrations must be reviewed and documented, and applied only to copies, and the original digital material must remain untouched. There is an excellent chance that future options for migration or emulation may be far better than any available at present. Keep the original, document every event that touches it, along with the date, person responsible, and why the event occurred. Most of all, verify the checksum at every juncture and prior to backups. No transfer of content is ever considered successful until after the checksums have been verified.

Protection of the Ability to Access and Use the Content

How do we make sure that the content can still be accessed and used? After all, throughout time we are likely to collect many forms of content, on many types of media. Ideally, all incoming content would be copied to a secure, offline computer from external media (after capturing checksums, write-blocking the media to avoid changing it, and testing for malware and viruses). Here is where content will be analyzed for access in its current form, metadata collected and documented, and decisions made about significant properties that would help determine what archival file formats should be used for long-term storage. Unless your organization is prepared to emulate everything in its original form (which is highly unlikely), original material should be "normalized" to archival formats, especially if the original is in a proprietary form. As mentioned earlier, proprietary file formats are not long supported. Failure to migrate them upon ingest is likely to mean failure to access the content in the near future.

Select archival formats according to the type of content, as described in chapter 6. By limiting the types of archival files your organization manages, you increase your ability to effectively curate the content on a long-term basis. The author advises storing some information in an easily accessible (but protected) database, for example, format type and version, and where each file is stored. This will make it possible for the manager or curator to generate reports of all file types and versions stored, and where they are located. For each archival format, document what open-source software will effectively render the file (see table 7.2). Add links to the software, and, if possible, also document any software dependencies. For example, Evince software depends upon an installation of libspectre for postscript rendering, and upon Poppler and XPDF for PDF rendering. These links and the dependencies, and even the existence of this software, may change throughout time. Ideally, a copy of the necessary software would be stored with the archival content and updated as new versions are made available. Every time your organization changes operating systems or replaces desktops, a staff member should review the types of files stored, the links to the software, and the extent to which the software effectively renders the appropriate file types on current hardware and software.

This is another situation where awareness of current practice in the digital preservation community is helpful. It is possible that although the manager's software and hardware currently do not support the file type or version, other software and hardware are available that will do so. In this case, the action to migrate or emulate may be postponed. Conversely, it is possible that the manager's software and hardware are already behind the curve and the content in question is already widely obsolete. In this case, action is already required.

Table 7.2. Format Reference Table

Description	*Extensions*	*Category*	*MIME Type*	*Software*
TIFF	tiff, tif	supported	image/tiff	Gimp
PNG	png	supported	image/png	Gimp
JPEG	jpeg, jpg	supported	image/jpeg	Gimp
JPEG 2000	jp2	known	image/jp2	
GIF	gif	known	image/gif	Gimp
WAVE	wav	supported	audio/x-wav	LAME
AIFF	aiff, aif, aifc	known	audio/x-aiff	LAME
FLAC	flac	known	audio/flac	Audacity, VLC
MPEG Audio	mpa, abs, mpeg	known	audio/x-mpeg	Audacity, VLC
OGG	ogg	supported	application/ogg	ffmpeg
MXF	mxf	supported	application/mxf	ffmpeg
MPEG	mpeg, mpg, mpe	known	video/mpeg	VLC
MARC	marc, mrc	supported	application/marc	MarcEdit
Text	txt	supported	text/plain	vim
XML	xml	supported	text/xml	vim
HTML	htm, html	supported	text/html	vim
SGML	sgml	known	text/sgml	vim
Postscript	ps, eps, ai	known	application/postscript	Evince
Rich Text Format	rtf	known	text/richtext	OpenOffice
Adobe PDF	pdf	known	application/pdf	Evince
Microsoft Word	doc, docx	known	application/msword	OpenOffice
Open Office Text	odt	supported	application/vnd.oasis.open document.txt	OpenOffice
Microsoft Excel	xls, xlsx	known	application/vnd.ms-excel	OpenOffice
OpenOffice Calc	ods, sxc	supported	application/vnd.oasis. opendocument.spreadsheet	OpenOffice
Microsoft PowerPoint	ppt, pptx	known	application/vnd.ms-powerpoint	OpenOffice
OpenOffice Presentation	odp	supported	application/vnd.oasis. opendocument.presentation	OpenOffice

Source: Developed by Jody L. DeRidder, 2017.

Consult with others in the digital preservation community to determine what the current supported archival file formats are for different types of materials. Review such registries as PRONOM,[18] DROID,[19] and the Library of Congress site[20] to help in selecting new archival formats, and use the Conversion Software Registry[21] to identify what software may be used to convert a copy of your current content to the chosen target format(s). Use your database (mentioned earlier) to locate the files that are impacted by the change, and test the migration process fully before implementing. Then document what you have done, when, why, and to which content, for the sake of provenance and future curators' sanity.

WHAT MATTERS MOST

Remember that the entire purpose of digital curation is to ensure future access and use. That use may be compromised if we cannot prove that we have the untouched original or if we cannot document who has had custody of the content since it was created or obtained. If we lose the content due to security failures, inadvertent errors, malign intent, disasters, or loss of funding, our work is to no avail. The constant change in hardware and software requires continual monitoring. If we allow the content to be stored too long without review or fail to store copies in archival formats, then as time passes, that content may become obsolete and totally inaccessible. Continued funding is crucial to ensure continual protection, storage, and effective staffing. Conscientious managers and curators will plan ahead, prepare for possible disasters, and also plan for business continuity. Life is about change. Effective digital curation is a responsibility we accept in service to both the present and the future of access to valuable knowledge and information. This mantle of responsibility is key to the development of future knowledge. Focusing on why we do what we do will help us to clarify how best to proceed, and in the end, it will make all the difference in the world.

KEY POINTS

1. Keep the original content untouched.
2. Verify checksums before and after every move or transition, and before each backup.
3. Keep multiple copies in geographically disparate locations, preferably on different types of systems.
4. Protect and secure your storage systems.
5. Plan who should do what (including backup personnel) for each stage of digital curation and each disaster scenario.
6. Keep plans current, and monitor task completion.
7. Document what types of files are kept where, and limit the archival file types managed.
8. Document all changes and transitions.
9. Review content after each hardware and major software upgrade.
10. Align your curation work with your institution's mission; build it into the strategic plan, and obtain a line item in the budget to support digital curation.
11. Plan for business continuity.

8

How Do I Provide Access Throughout Time?

The entire purpose of digital curation is to support continued use of the content throughout time. This is a point that seems to elude many who are involved in digital preservation, believing that backups are sufficient, or even that effective storage is sufficient. At some point (if not continuously), versions of the stored content need to be made available to the targeted user audience, but in what form? As time passes, the types of hardware and software in use change, and so the methods of access to content must change as well, regardless of the original formats and original media. Some consider this the primary difference between "digital preservation," which may overlook access entirely, and "digital curation," which spans the life cycle of the digital content, including its creation, use, and reuse.

Remember, the focus of librarianship is to provide the user with the resources needed, in the form needed, at the point of need. The form of resources changes throughout time, as does the point of need: Is this document required via a mobile device? Is it needed in a 3-D visualization? Does it require geocoordinates for effective use in this instance? The context of user needs must inform our methods of providing access, which means we must look beyond the materials themselves to how they would or could be used. In this section, we address methods for assessing user needs, considerations for granularity of access, and derivative selection and assessment to meet the needs of current hardware, software, and users.

WHAT DO USERS NEED?

Fifty years ago, the personal computer did not yet even exist.[1] General user needs were focused almost exclusively on access to analog content: the printed word, printed images, hard copies, and original documents. The height of convenience was to have documents delivered, and it was often sufficient to have access within a few miles' travel. Today, however, users expect digital content to be accessible on demand, at the touch of a finger, or as the result of a simple query, delivered to desktop computers, laptops, tablets, televisions, handheld devices, and more. Such a tremendous change in fifty years. Already, computer operating systems are available in watches[2] and eyeglasses,[3] and "virtual reality" 3-D headsets have entered the general market.[4] We cannot possibly imagine all the changes the next fifty years will bring, much less the next hundred. How can we possibly keep in touch with the constantly shifting landscape?

The basic methods of user needs assessment include interviews, focus groups, observations, voluntary feedback, and surveys, and we will address these in turn. Aside from these, additional information can be gleaned from system logs or tracking software, particularly to obtain information about the type and version of browser or type of device accessing a site, and the extent to which users of this type of device are effectively navigating the content you have online. When, for example, your weblogs[5] (or Google Analytics[6] or other analytics software) indicate that there's a high "bounce rate"[7] for users of a certain browser or type of device, that's a clue for closer investigation. It may be that your method of delivering content online is not effectively usable for them, and you need to determine how to change your delivery methods to better meet their needs. While this type of feedback can be invaluable, it is only an indicator and cannot substitute for user studies.

When selecting participants for a user study, it is best to first identify your target audiences and select participants carefully to ensure they are representative. This involves considerations of age group, educational level, expertise in navigating digital content and software, knowledge of the topical subject, range of disabilities, primary and secondary languages, and (if evaluating an existing system) familiarity with the system in question. When analyzing the results of your study, you will want to divide responses according to these same criteria to determine the impact they have upon user needs. For examples, those for whom English is a second language may interpret textual instructions, metadata, and indicators differently; those who are unfamiliar with the topical subject may be less able to locate the information they need; and so forth. Carefully assessing potential external impacts upon the results of your study will help you determine the common threads that best reflect the intent of your research.

Before proceeding with any form of user study, be sure to review your organization's protocol and requirements for complying with research involving human subjects. In the United States, this falls under the Food and Drug Administration's regulations for clinical trials.[8] In Europe, a primary resource would be the European Network of Research Ethics Committees.[9] For more information about guidelines and rules in various countries, review the "International Compilation of Human Research Standards."[10] Adequate protection of human subjects needs to guide and inform all user studies.

Interviews

There are two basic types of interviews: those that ask a user to evaluate a specific service and those that investigate how the user performs his or her tasks. The first is most suitable when developing a new service or evaluating its effectiveness, and the second is most useful when you want to find out what you don't know. Thus, the latter form of interview can be exceptionally powerful, as you will discover areas of user need that surprise and inform your choices and future work. This is especially important when seeking how best to provide access to older documents, as the needs of scholars and patrons change throughout time, reflecting new options and discoveries in the digital realm.

Although early digital library systems have been built around what seemed possible based upon the content itself, this siloed approach is not well-suited to discovery on the web, nor does it consider the different methods scholars seek to use materials. "Build it and they will come" is an approach that rarely works for more than a small segment of the desired audience. While initial marketing will raise awareness of the service temporarily, ongoing outreach of some sort is required, and even then, the user is required to navigate to the site, rather than the desired content being pushed to the user at the point of need. This latter approach is the desired method for online delivery, and much marketing today focuses on predicting user need to deliver appropriate options.

Today, Google paves the way in efforts to provide users with what they need upon request. Yet, few of the resources we have in databases are represented in those search engines effectively. Thus, knowing where your designated community or target audience is going to locate materials or how they are seeking them will greatly inform what work you need to do for your resources to reach your audience. Beyond that, you need to know how they want to use the resources so the form in which you provide the materials is suitable to their needs. Whether you are asking users to evaluate a specific service or studying how they perform their tasks "in the wild," you will want to include two types of measurements:

1. "Performance" measures: How well can they do what they're trying to do?
2. "Satisfaction" measures: How happy are they with the methods and interfaces they're using?

To measure satisfaction, most user studies include survey-type questions that ideally are measurable and comparable. These often are based on the Likert scales, in which participants select the degree to which they agree to a statement:

1. strongly disagree
2. disagree
3. neither agree nor disagree
4. agree
5. strongly agree[11]

These types of responses are easy to analyze and compare to earlier and later studies, as you can assign a weight, or value, to each selection. For example, the aforementioned options could be assigned values from one to five, where "strongly agree" is measured as five. Then, an average of rankings for all survey participants will show whether most tend to agree or disagree with the statement.

Common types of measurements used to assess performance include the following:

task success
time on task
errors
efficiency
interface learnability[12]

How these values are defined and measured vary from researcher to researcher, so be sure to clarify in writing the choices and decisions you make when developing your study. The most recent and comprehensive analysis of how best to define and evaluate these measurements for software systems is described in an International Organization for Standardization standard now under development.[13] Even so, the analysis of the results of each of these can be complex. For example, if few of your participants can complete the assigned task, either the task selected is too difficult, the participants selected are not representative of your target audience, or there are serious problems with your interface. Creating, implementing, and analyzing user studies is a painstaking process, and one should never jump to conclusions.

Be sure to document every decision, and include them in your publication about the results so others can repeat your study or learn from your oversights. The extent to which you document decisions is crucial to determining the validity of the study. For instance, how you measure errors determines what an error is. If what you have constituted as an error turns out to be a different approach to locating the same material, perhaps that aspect of the study should be evaluated under "learnability."

Oftentimes participants bring personal qualities to the table that become outliers in the resulting statistics. For example, the author once had a user study participant who was determined to succeed, exhaustively, regardless of the extent of time needed. Her "time on task" and "number of clicks" (our measures of efficiency) were extreme, although she always "succeeded." In another situation, a participant preferred to locate the finding aid first and peruse it extensively to determine what documents would best meet the search requirements. While eventually successful, this also skewed the results in terms of "time on task" and "efficiency." Yet, who is to say this is not an appropriate approach to locating archival content? In a third example, a participant (for whom English is a second language) could not succeed in the tasks, as she had difficulty understanding both the words in the queries posed and the terminology in the interface; we had failed to screen participants for cultural background. During your analysis, take care to understand the perspectives of the study participants, as these may shed light on assumptions and decisions that can completely change your study outcome.

When developing a user study, carefully consider all feasible perspectives on your choices. Consider potential causes of problems encountered, as this will inform changes to your system or offerings. For example, if a participant was unable to find a document using a specific term, check to see if the term is indeed in the metadata but was not properly indexed. Also check to see whether the search engine, database, web server, or network had problems during the time of the test. Lastly, ensure that the interface captured the entered term and sent it to the proper database or search engine. Since web delivery of content is multifaceted, the issues found by users may have been caused by different layers of delivery; finding the cause of the problem is sometimes as difficult as finding the problem itself.

When designing your interviews, you will want to provide structure that will enable you to extract the information you seek, without leading or influencing the participant responses. If, for instance, you are assessing the usability of your current delivery system, you will likely assign questions or tasks for the participant, with or without a request for them to speak aloud about what they see and the decisions they are making. Often in such situations, the participant will turn to the interviewee and ask questions, but any information shared with one participant must be shared with all of them or the results of the study may be impacted. It is well known that experiments are impacted simply by being observed,[14] and participants may try to "please" the interviewee rather than provide uncomfortable feedback that would be far more useful. To avoid undue influence, the person leading the interview must be carefully unbiased, even in inflection and body language,[15] and hence should ideally be someone with no benefit to gain from the outcome of the study. When possible, the study should be videotaped for later review for any information that may have been missed or misconstrued during the interview.

If seeking to understand the user's workflow, arrange for the interview in the participant's normal working environment. The most groundbreaking work of this type involved an anthropologist to help gain a deeper understanding of student needs at the University of Rochester several years ago[16]; however, even without an anthropologist on staff, you can gain a great deal of insight into user needs simply by seeking the opportunity to learn, particularly from the patrons you serve. For this type of interview, the structure you will provide will be minimal to provide for a broad range of variability in response: How else will you find out what you don't know?

For example, the author once studied how experienced researchers use online primary source materials. In this study,[17] participants (in their own offices) were asked to locate as many as three online resources for primary source materials that they particularly liked or didn't like and then asked to walk through the process as to how they use those sites and materials, describing each step. This was a highly qualitative study based on "grounded theory," in which the results of the study evolve from organizing the gathered information.[18] In this case, the author organized the information by the commonality of steps in the process (e.g., browsing, searching, selecting from results, accessing items, and downloading content). The most surprising results from this study were threefold:

1. Researchers didn't know how to find out about newly available content.
2. Researchers were confused by the variations in search interfaces (and didn't trust the algorithms).
3. Researchers are seriously challenged to effectively organize the information they gather for reuse.

There were other interesting findings as well, but none of these are tied to any specific interface. Yet, these findings suggested such interface improvements as the following:

Develop notification systems and publicity outreach for newly available content.
Work with users to ensure the search interface is clear, well documented, and easy to use.
Label download files with a combination of creator, title, and date for easy location later.

Each interface improvement could and should be included in the next interface study.

Interface-specific user studies should cover the basics of interaction with the software or service and should be carefully developed to answer specific questions. If, for example, you want to determine whether users benefit from a new method of description, you will want to perform an A-B type test, in which the same content is hosted in two interfaces: one with the first version of description and one with the second. Participants will need to be asked to perform the same searches in both interfaces (alternating which is first), and the results should then be compared.

Interface-focused user studies should be implemented during and after development of new services and then at least annually, since user needs change and services quickly become outdated. Before initiating a new service, however, another approach is particularly useful: the focus group.

Focus Groups

When you do not yet have something to show participants, the focus group is a preferred approach to gaining user input and guidance. As with the interviews, you will want to obtain a representative sample of your target audience. Unlike the interviews, participants do not perform tasks but instead react and respond to open questions and scenarios that you pose. Multiple participants attend the same session, and it is your job to encourage and facilitate open discussion to gather information. Sometimes you will want to form two or three focus group sessions, each with different participants, especially if you have multiple target audiences (each group should represent a single target audience). You may also want to set up a series of focus groups, to explore potential services or concepts in depth.

The benefits of focus groups include that participants are often encouraged to share more information when they hear of similar perspectives and experiences by others. The drawbacks of focus groups include that participants often seek to "please" the observer (the "observer effect"[19]), the most vocal or charismatic participants may sway the discussions, and there is no confidentiality or anonymity. If you choose to hold focus groups, engage a skilled facilitator to increase the engagement of the quieter participants and control strong personalities in the group. Select your questions and scenarios carefully to avoid leading the group to the conclusions you hope they will choose. For instance, asking a group to consider the potential ways they might want to use a map digitally is an open question, whereas asking them to compare potential methods of access to maps may lead them to believe you prefer one method to the others. In such a case, the results you get will likely be an affirmation of your own preferences, as people are good at reading nonverbal communications and generally seek to please to gain acceptance and approval.[20]

Observations

Although observation is incorporated into every form of assessment, it deserves a discussion point of its own. Multiple levels of observation should be made on a regular basis, and this will require engaging others within your organization. Those who are regularly involved with patrons (whether in the patron's own working environment or in shared spaces like libraries) should be alert for changes in types of equipment, software, storage devices, and methods of using content. Those who are supporting servers need to be aware of the impact changing systems and updated software libraries can have on access capabilities and new options for the content you are providing. Those who are monitoring web services need to be watching for indications in the weblogs that users may be having difficulty using current services. And you, as a primary stakeholder, need to be observing changes in the field and how other leading institutions and businesses are providing access to content similar to what you are offering. Monitor conference website publications and relevant journals on a regular basis, looking for new methods of access and use. To stay abreast of user needs requires a coordinated effort and excellent communication with colleagues who have access to the information you need.

Voluntary Feedback

Most of our target audiences access the content we provide via the web; this means there is no one present to talk to when something goes wrong, when they need something that isn't provided, or when they simply have a question or suggestion. To what extent do your online services offer opportunity for interaction or feedback, and are those access points available at every point of need? Some sites include online chat functionality, at least during certain hours, but all sites should include a link asking for feedback and suggestions that is incorporated into every web page (usually via a common "footer" file). Make sure the link is visible and inviting; test regularly to ensure it works. If the link is to an online form, regularly fill it out and verify that the information is received and acted upon appropriately by the dedicated personnel, who have backups for when they are unavailable. Normally, all submissions should be stored in a database and periodically reviewed for commonality of complaints or issues, as they may indicate something of which you are unaware.

Although most of your voluntary feedback will likely come via online feedback forms or e-mail messages, other feedback can be gathered by "listening in" on listservs where users are discussing various websites and service offerings, in conference conversations, or in class discussions. The latter is information that may be gathered by archivists and librarians performing outreach and instructional training, where the discussions uncover attitudes and anecdotes about your services and offerings. Seek out such feedback, from whatever source available. Negative feedback about your site, services, or content is "gold," as you cannot ask for more useful information upon which to act.

Surveys

Ideally, all survey-collected data will be anonymous. Why? People who are asked directly for information (e.g., by telephone or in person) provide more positive feedback than when asked anonymously via the web.[21] This is called the "social desirability bias,"[22] which, of course, plagues the focus group approach. Privacy and anonymity are empowering and encourage even the meek to voice their opinion without fear of reprisal. In today's world, where new methods are developed constantly to better track what we do online, participants are more and more wary of just how anonymous their responses are.

It is standard practice to incorporate some questions to identify the segment of your target audience who is responding; however, if the questions asked are sufficient to identify participants, they have lost the protection of anonymity, and their responses may reflect their recognition of that loss. Even if you make it voluntary for a participant to provide contact information for later follow-up, be aware that in doing so, there's a good chance their responses may be skewed by knowing that you know who they are and may share that information with others. Remember that the most valuable information you can obtain via surveys will be the most anonymous.

Most web-administered surveys are sent to numerous potential respondents and offer no method of obtaining clarifications to questions. Include a brief introduction to provide context, and have multiple people review your questions to ensure they are clear. A study in 2008 found that response rates for surveys that went to individuals averaged a little better than 50 percent and were not improved by incentives.[23] Since many recipients are oversurveyed, limit the survey to ten minutes or less (twenty at most), and provide access to a list of the survey questions in the invitation, with the expected time expenditure. Explain why participant responses to the survey will be useful in terms that apply to them. For example, state that your methods for providing access to digital content will be modified to better meet the needs of those responding to the survey. That alone will provide incentive to those who use your materials but encounter difficulty in doing so.

Since you may expect many responses to a survey, it's important to use rating scales like Likert, to which you can assign values for comparison and counting. Use the same type of scale throughout the survey to avoid confusion, and be sure to include an opportunity to select "neutral" or "no opinion" so as not to force someone to take a stand on something. Respondents who feel pressured or coerced are unlikely to complete the survey. When providing a multiple-choice list of options from which to select, be sure to offer an "other" response, with a text box for the respondent to fill in. The author finds it useful to include at least one open question in each survey for any other feedback or clarifications. While this is not easy to compile or analyze, it serves as a clear invitation and often provides a rich source of valuable information.

LEVELS OF ACCESS AND "WHAT'S A COLLECTION?"

From a librarian perspective, a "collection" can be any grouping of content that meets patron information needs, which is a broad definition indeed. "Collection development policies" are designed and maintained locally to clarify the scope of content collection, how content is managed and maintained, and the strengths of the local collection. Oftentimes, collections are topic or genre-based groupings, for instance, a "social sciences" collection or a "fiction" collection. The method of locating content within a collection is via search on metadata (catalog records) captured about individual items, or via browse options: online (metadata-based) or by physically navigating the collocated content in the facility. It was for the latter purpose (grouping like content) that such classification methods as the Dewey Decimal System[24] and the Library of Congress Classification System[25] were developed.

In the online environment, however, the classifications need to be included in the item-level description and available for browsing in a user-friendly method. Since both these schemes involve numbers and/or letters and punctuation for reference, the associated labels for the values are most useful for browsing purposes, for the lay person. Similarly, online browse options by subject, topic, genre, geographical location, or type of material can be helpful. Numerous controlled vocabularies are available for these but are only useful within a system to the extent that they are consistently applied, indexed, and available as linked results sets for users, organized in an intuitive fashion. Additionally, since web search engines cannot delve into back-end databases to unearth content for indexing, be sure to provide sitemaps[26] and either descriptive metatags[27] or schema.org[28] encoding in the web display so your resources can be effectively located on the internet.

In the archival realm, the determination of "what is a collection" is normally based on "provenance": where did it come from and who owned it? As defined by the Society of American Archivists, the "principle of provenance, or the 'respect des fonds,' dictates that records of different origins (provenance) be kept separate to preserve their context."[29] In practice, this often means that a collection is named for the person, family, or organization from whence it came, for instance, "The John Raymond Smith Family Papers." From a discovery perspective, this raises the question of how a user is to locate which collections contain what they seek, if the seeker doesn't know who might have owned that information in the past.

Locating information in archives is often heavily dependent upon the resident experts of the collections, and in the web environment access to that curator's knowledge is rarely available. When collections are organized by the source or origin, the finding aids and access points need to be heavily developed and web-searchable to facilitate content location. The now-international Social Networks and Archival Context (SNAC) project is deeply involved in building this type of accessibility into finding aids,[30] leveraging the EAC-CPF (Encoded Archival Context-Corporate Bodies, Persons, and Families) standard,[31] which is used with the EAD (Encoded Archival Description) standard.[32]

If choosing this approach, the author highly recommends careful encoding of relevant information that users may need for locating content within the collection and use of schema.org[33] encodings for the finding aids to assist web search engines in effective indexing and retrieval. Duke University has demonstrated how to encode access to digital content at the intellectual item level into a finding aid,[34] making use of the Digital Archival Object (or "dao" element) of the EAD finding aid schema.[35] They leverage multiple types of viewers and players to provide access to the digitized content online, whether it's streaming video, a slide show, or images, and links are embedded to "request this record" if the file is not readily available online.[36]

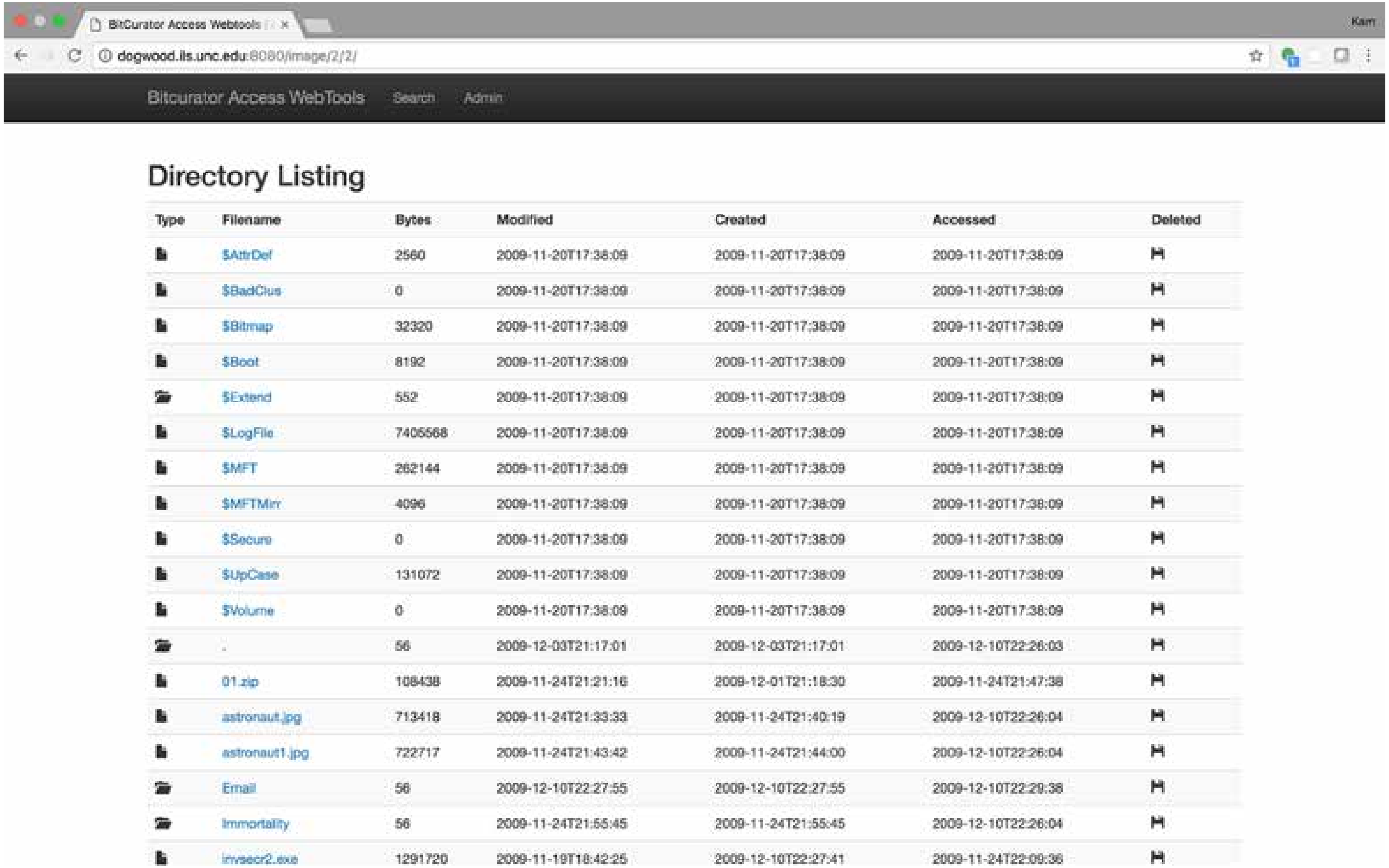
The University of Michigan approach is to link from the finding aid to an external system that hosts the digitized item, noting in the finding aid if access is restricted and until what date.[37] To be feasible on a long-term basis, you must make use of persistent links within the EAD metadata for each link to content, as it's not reasonable to have to go back and modify the finding aid links every time you need to update a system, move content, provide a new version of the resource, or change servers. Instead, changes to the links would then be made (in bulk) at the database level where the redirection from the persistent URL is maintained.

Feasibility and resource limitations often prevent detailed description of all intellectual items within your collections, and you will need to follow the "More Product, Less Process" method[38] most archivists embrace, in which you focus the level of description based on the value of the content itself. In this approach, your priority is to first provide some level of access to all collections and then focus your efforts on deeper and more detailed description of content that is of special interest to your target audience. If, for example, your target audience is K–12 (kindergarten through twelfth grade) students, you would be more likely to provide item-level description to such famous documents as the U.S. Constitution or famous speeches like the *I Have a Dream* speech given by Reverend Martin Luther King Jr. in 1963. Documents that can effectively be grouped for access may be described at the grouping level (or container). This brings us to the discussion of digital content description beyond the intellectual item level.

The form your content takes may determine how you will organize it for access and retrieval. If you have collected a large quantity of hard drives or websites to which you want to provide access, it is likely that you do not have the resources to provide intellectual document-level description for the contents of each one. For archives of web-harvested content, the leading method is to leverage the Internet Archive web harvester by subscribing to their Web Archiving Services. Example collections are freely available online; an example set of interest would be the several "Obama" collections.[39]

If you're working with hard drives, the most suitable method for providing access via the web is to use the open source BitCurator Access Webtools[40] (see figure 8.1), which builds upon the freely available BitCurator Environment for accessing hard drives and curating the content within.[41] A recent exploration into building upon this access to leverage "Emulation as a Service" (EAAS)[42] for older content has uncovered just how crucial it is to gather information about the source hardware and software.[43] Ideally, when a user clicks to access an older disk image, the website would provide the emulator service with sufficient information to spin up a hosted environment, which would provide the user with the original experience for accessing the resources. Be sure to keep an eye on the BitCurator progress[44] with this goal.

Whatever approach you take, remember that the primary goal is to meet the needs of users. Knowing in advance what you want the results to be will help you determine the scope of your work, as well as how to proceed. In 2012, Niu analyzed the usability of existing web archives using a functionality checklist that may be helpful; she considered basic functionality to include searching by URL and keywords, and narrowing the search by dates, domain, and media type.[45] More advanced functionality, for instance, data mining support, would be ideal. An example here would be the UK Web Archive,[46] which offers browse options, a Memento[47] service to locate other captures, and visualization tools (e.g., N-gram search, tag clouds, and link and format analysis). Much of the functionality you will be able to provide for users will be dependent upon the metadata created during development or initial capture. For example, if browsing by subject, date, and media type is desired, you will need to assign subject headings, media types, and dates for each site captured. For keyword searching, you will need to extract full text and (of course) provide an indexing service.



BitCurator Access Webtools
dogwood.ils.unc.edu:8080/image/2/2/
Kam
Bitcurator Access WebTools
Search
Admin

Directory Listing

Type	Filename	Bytes	Modified	Created	Accessed	Deleted
📄	$AttrDef	2560	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$BadClus	0	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$Bitmap	32320	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$Boot	8192	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📂	$Extend	552	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$LogFile	7405568	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$MFT	262144	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$MFTMirr	4096	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$Secure	0	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$UpCase	131072	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📄	$Volume	0	2009-11-20T17:38:09	2009-11-20T17:38:09	2009-11-20T17:38:09	💾
📂	.	56	2009-12-03T21:17:01	2009-12-03T21:17:01	2009-12-10T22:26:03	💾
📄	01.zip	108438	2009-11-24T21:21:16	2009-12-01T21:18:30	2009-11-24T21:47:38	💾
📄	astronaut.jpg	713418	2009-11-24T21:33:33	2009-11-24T21:40:19	2009-12-10T22:26:04	💾
📄	astronaut1.jpg	722717	2009-11-24T21:43:42	2009-11-24T21:44:00	2009-12-10T22:26:04	💾
📂	Email	56	2009-12-10T22:27:55	2009-12-10T22:27:55	2009-12-10T22:29:38	💾
📂	Immortality	56	2009-11-24T21:55:45	2009-11-24T21:55:45	2009-12-10T22:26:04	💾
📄	invsecr2.exe	1291720	2009-11-19T18:42:25	2009-12-10T22:27:41	2009-11-24T22:09:36	💾

Figure 8.1. Disk Image Access for the Web. *"Screenshot (as of June 2017) of BitCurator Access Webtools; software available from https://github.com/bitcurator/bitcurator-access-webtools. Reprinted with permission of the BitCurator Access project team." Note: For a discussion of an earlier prototype of this software, called DIMAC (Disk Image Access for the Web), see Sunitha Misra, Christopher A. Lee and Kam Woods, "A Web Service for File-Level Access to Disk Images," Code4Lib Journal 25 (2014-07-21), available from http://journal.code4lib.org/articles/9773.*

Clearly, providing access to digital content can be as simple as letting users download the complete set of data in its original form or as complex as item level description, emulation of content, data mining, and visualization. The depth and breadth of your access methods should be suited to the needs and capabilities of your target audience.

DERIVATIVE SELECTION AND ASSESSMENT

Online delivery of content is heavily dependent upon the extent to which current browsers can support the file types chosen. For example, the current browser comparison for image files shows that JPEG, GIF, and PNG are by far the most widely supported image formats at the time of this writing.[48] For audio, MP3 and AAC in an MP4 wrapper are most widely supported, and for video, your best options currently are VP8 and Vorbis in WebM, H.264 and MP3 in MP4, or H.264 and AAC in MP4[49] (video contains images and audio—and sometimes text—in a wrapper that explains how the parts work together).

When selecting a web-accessible format to serve as a "derivative" or "access" from your master archival file, you will need to review current support among browsers and then consider the size of the file. Faster internet speeds mean faster loading of web pages. But those working via wireless access points are limited by the amount of bandwidth they can get from the closest router point (after all, they are competing with the other wireless devices nearby) and the speed of their wireless modem. If you are lucky enough to be working for an institution that provides high-speed access, remember that most of your target audience will probably be less well-supported. You will need to find a balance between providing the highest quality online representation you can and the download access speed available to most users. Lower resolutions are normal for web access than are acceptable for archiving and preservation. It won't do you any good to offer high-resolution images if your patron's browser hangs up and is unable to open the page as a result. If serving up video or audio, it's best to arrange for a streaming server to provide access, as these are large files. A streaming server will send out a few bits of the file at a time so the user doesn't have to download the entire file before accessing it. Most streaming servers will allow you to set multiple speeds of access so each user can be provided with a feed of content suitable to their download speed. Without this, user access to the files will break up or freeze, or may be lost entirely.

Another consideration when selecting derivatives is the extent to which the chosen file format represents your resource well (or as well as it can, in the given environment). Be sure to at least spot check the results when selecting file format and resolution. Unfortunately, some of the types of content you collect will not have web-accessible formats and will require special services (e.g., web archive delivery software or emulation) to allow users to experience the resource online. This is why so much research is going into how to emulate older computer environments and how to best migrate older file types to new ones. In the meantime, the lowest-cost method of providing access is to simply allow users to download the original and trust them to find their own way to render or use the resource. This, however, does not protect portions that should be redacted (e.g., those containing social security numbers) or should not be available until a certain date (under copyright, for instance). This is also not a user-friendly approach and is not recommended.

As of 2016, most institutions still provided access to bulk digital content (not described at the item or folder level) only within the bounds of their institution walls and at the discretion of their archivists.[50] This again is not user-friendly, as it requires the researcher to physically travel to the institution, locate parking (and perhaps lodging), arrive during open hours, and request access in person (which may even then be denied).

In summary, preserving digital content is not sufficient; we must provide access to it, to the extent legally and ethically possible. Otherwise, there is no point in preserving and curating the resources. Access, however, is an area where much research and development is needed. We have a plethora of older content arriving in archives and special collections every day, in every conceivable format and media type. This content must be evaluated, sifted, and then effectively curated and managed for access. But what type of access do our users need? We have much to learn in this regard, and it will be an ongoing challenge to stay abreast of those needs and find ways to meet them in the digital environment.

KEY POINTS

1. User needs change radically throughout time, as our culture and digital environment evolve; regular assessment of your target audience requirements is a necessity.
2. When selecting participants for a study, ensure that they represent your target audience and carefully control for variables that can impact the results.
3. Semistructured interviews with participants in their normal working environment will provide surprising information about general needs; structured interviews are best for evaluating existing services or interfaces.
4. If you don't yet have an interface or service, hold a focus group to evaluate ideas and options.
5. Invite regular feedback from users via web interface and surveys, but also ask for regular feedback from technical support personnel and those who observe users in public spaces and classrooms.
6. While most digital libraries today are based on intellectual item-level description, other levels of granularity of access become necessary when working with digital content in bulk quantities.
7. The method of web delivery and the formats chosen should provide the best possible representation of the resource possible, within the constraints of browser support, bandwidth, and download speed.
8. Initial research into web delivery of disk images via emulation is promising, and much remains to be developed.

9

How Can I Leverage the Community?

Effective digital preservation and curation requires a community. Although it may seem that you are alone in the face of daunting challenges, and you may certainly be alone in facing a tremendous influx of digital content to be managed, it is simply not possible for anyone to do all of it without community support. Similarly, it is not possible for the digital preservation and curation community to make progress and develop best practices and standards without the input and recommendations of a multitude of curators in the field. Your hands-on experiences will inform others, if you communicate them, just as paying attention to others' experiences will inform your own and allow you to avoid many missteps. In many ways, we are on a shared journey through a wilderness, mapping out the best trails, putting up signposts, sending out scouts, and laying roads in the areas we are certain are the best ones, for now. Since none of us can foresee the future, we can never be certain. Our paths and roads will likely change as we enter new parts of the forest and find waterfalls, swamps, and dangers to avoid, and new vantage points, over new beautiful open fields of opportunity, to guide us further.

Developing best practices and guidelines is often via a trial and error approach, as those in the field test new tools and equipment, develop new software, and discover new possibilities. We all need to help one another and keep updated on the latest findings and options. To stay abreast of what's current, pay attention to what's happening in the field and contribute your own findings via listservs, presentations at conferences, and article publications. To help you get started, the following list of resources is organized by focus area: tools for selection and capture, partnerships for storage and protection, efforts that focus on specific types of content, and options for learning and contributing back to the community.

FOCUS ON SELECTION AND CAPTURE

No matter what type of content you are collecting, you will find the community-developed tool collections and file format registries invaluable. For example, without the U.K. format registry Pronom,[1] you would be unable to verify many types of files against the specifications that make them unique. While currently we depend on the DROID software[2] to leverage this registry, another registry that would extend capabilities is under development by New Zealand, Australia, the National Archives and Records Administration, and the University of Portsmouth.[3]

The following three community efforts engage scholars and practitioners alike to identify, create, and share tools that are invaluable for the tasks of digital curation. Institutional memberships are available for the Open Preservation Foundation and the BitCurator Consortium. DigiPres Commons provides a place where you can share your findings, get questions answered, and submit new tools for use by others.

The Open Preservation Foundation is an international nonprofit organization that leverages member involvement to provide shared solutions for digital preservation. Notable among their current projects is an extremely useful tool: VeraPDF,[4] an open-source validator for PDF/A[5] (the archival version of PDF), developed in collaboration with Dual Lab, the Digital Preservation Coalition, and KEEP Solutions.[6] One of the previous products still maintained and developed is JHOVE,[7] which is widely used for file format identification and validation. The Open Preservation Foundation began as a four-year digital preservation project (Planets) for cultural and scientific assets, cofunded by the European Commission.[8]

Among the products of Planets was Plato, a preservation planning tool to identify the characteristics of digital objects and measure the effectiveness of tools to preserve them.[9] Also included in the Planets Suite is an Interoperability Framework, through which users can discover tools and define and execute preservation workflows.[10] Available training videos cover everything from why and how to preserve to selecting tools and using the Planets software.[11] Openly available publications include case studies, presentations, articles, and technical reports.[12] Planets was so successful that it won an award in 2012, for outstanding contributions for research and innovation.[13] This success led to the establishment of the Open Planets Foundation to continue support, which, in 2014, became the Open Preservation Foundation.[14]

Two types of organizational membership are available: software support (focused primarily on JHOVE maintenance and development)[15] and membership either as a charter or affiliate member, with costs based on operating budget.[16] Affiliate members may pay their dues via work contributions, and at this writing, fees begin at ten days of support or 1250 British pounds (approximately $1,700) per year.[17] Charter members help to set the strategy and prioritize software, but all members have access to interest groups, training materials, and more.[18]

DigiPres Commons, hosted by the Open Preservation Foundation, seeks to provide a gateway to the community-owned, community-oriented resources that are helpful for digital preservation.[19] This site offers helpful information for curators at any stage of their career, from how to save content immediately[20] to a site where you can submit your digital preservation question or your own answers to other peoples' questions.[21] If you're trying to find the right tools for different aspects of digital curation, their Community-Owned Digital Preservation Tool Registry (COPTR) has what you need. At this writing, the registry lists 426 tools,[22] organized by functional category and the type of content they act upon, or helpfully organized by both in the POWRR Tool Grid.[23] The success of this effort depends upon you. Contributions are welcome.[24] Share your experiences of using a tool, suggest new ones, send feedback, or provide updates or enhancements to existing entries.[25]

Related to the Open Preservation Foundation is the BitCurator Consortium, a relatively new group focused on the development and support of digital forensics tools for digital curation.[26] This group is centered on support of the openly available BitCurator software,[27] which provides a prepackaged Linux software environment encompassing several powerful tools to enable you to analyze and assess quantities of incoming digital content, for instance, hard drives. Institutional membership dues (which currently stand at $2,000 annually) go to support the software development and provide the latest in findings, training opportunities, networking, access to a help desk, and more.[28] If you are in one of the many institutions struggling for how best to navigate the influx of digital materials into your special collections and archives, membership in this community should be a priority on your list.

FOCUS ON STORAGE AND PROTECTION

One of the big challenges in storage and protection is how best to ensure that content is replicated across diverse geographical locations and types of hardware to ensure that no one type of disaster will destroy our precious content. Since cloud storage and hosted solutions often overlook some of our needs in the fine print or come at a price we can't afford, you might want to consider consortial, regional, national, or local agreements with other organizations. Another benefit to these networks is the opportunity for you to create an agreement with another institution to take over the curation of your content should your institution fail or lose funding: business continuity. In this time of rapid change, it's an insurance policy for valuable content that researchers need.

The leading content-agnostic freely available software system developed for this is LOCKSS (Lots of Copies Keep Stuff Safe), so this section describes the variety of groups organized to use that software for different purposes. Also mentioned is a research institution–led group called the Digital Preservation Network and a free service for researchers, the Open Science Framework.

LOCKSS is an award-winning open-source networked digital preservation system that enables partners to replicate their files across disparate geographical areas and systems to ensure the bit-level files do not become corrupt.[29] Each hosting partner maintains the hardware ("LOCKSS box") on which the software resides, automatically syncing files with network partners (see figure 9.1). As the first and only system built on the purchase-and-own library model for electronic materials, LOCKSS began as a model for such published content as journals but now is being used for cultural heritage materials as well. The software is freely available[30]; membership in the LOCKSS Alliance includes technical support from Stanford (the annual cost is free for publishers and based on size of institution for others).[31]

There are two basic approaches to make use of this software (and service): Join the Global LOCKSS Network (GLN) or participate in one or more Private LOCKSS Networks (PLN). The Global LOCKSS Network (GLN)[32] is for general content, involving libraries and vendors to preserve open-access titles and subscription e-books and journals. For nonopen-access content, cultural heritage materials, data sets, institutional repository content, government documents, and digitized content, Private LOCKSS Networks are the solution of choice.

CLOCKSS (Controlled LOCKSS), a single example of a Private LOCKSS Network, is the nonprofit venture between leading academic publishers and research libraries that uses LOCKSS to create a dark archive of web-based scholarly publications for the world's benefit.[33] In 2014, this system was evaluated (by the Council for Research Libraries) against the Trustworthy Repository Audit Criteria[34] and "awarded an overall score matching the previous best."[35] Publishers and libraries who wish to participate and support the network are welcome; fees are based on revenues or budgets, respectively.[36]

Some examples of Private LOCKSS Networks[37] of interest include the following:

Digital Federal Depository Library Program (USDocs), a partnership between libraries and the U.S Government Publishing Office to preserve born-digital government publications[38]

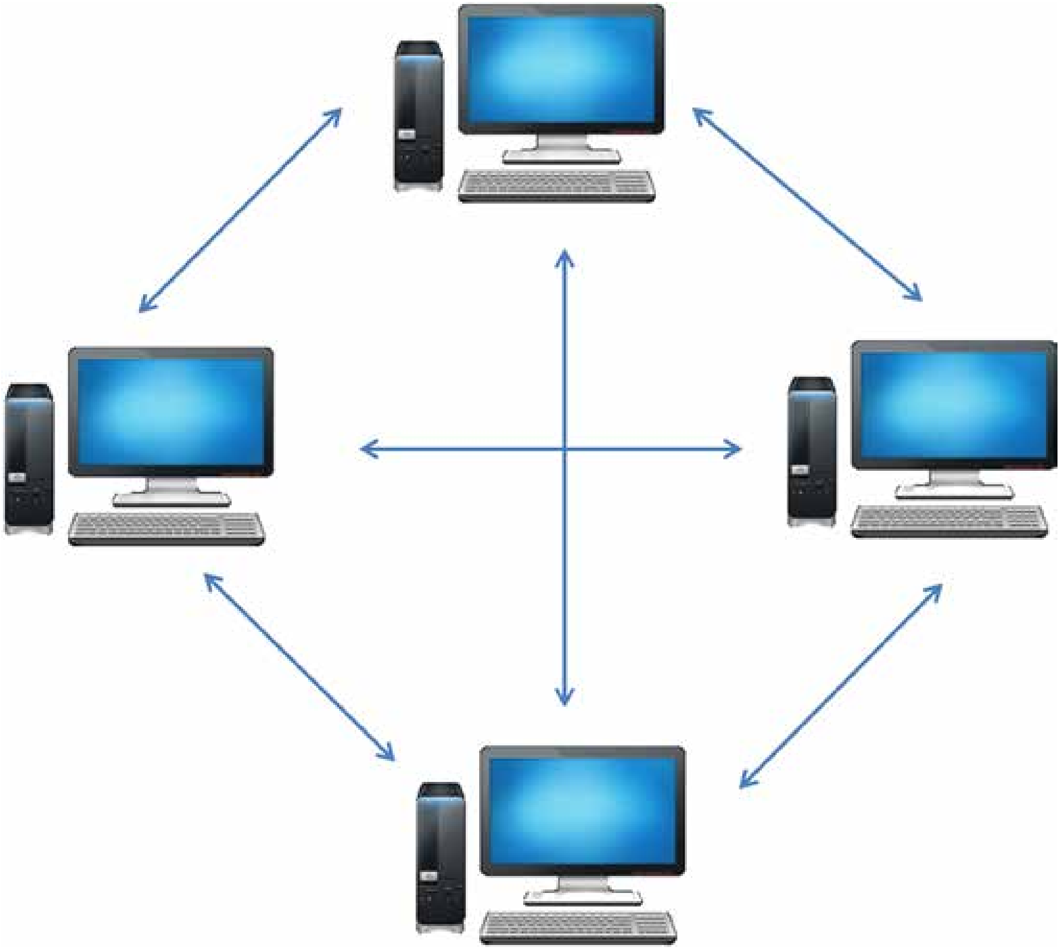


Figure 9.1. Diagram of Peer-to-Peer Replication of Content. ***Copyright Jody DeRidder, 2017.***

Canadian Government Information LOCKSS Network, a similar program in Canada to preserve at-risk Canadian government information[39]

Council of Prairie and Pacific University Libraries (COPPUL), which focuses on freely available digital web content important to British Columbia, Alberta, Saskatchewan, and Manitoba[40]

MetaArchive Cooperative, run by the Educopia Institute, which supports cultural heritage materials and is the largest Private LOCKSS Network[41]

Alabama Digital Preservation Network (for an interesting approach that bears replication), which reduces costs by having members manage the network and has a cost model that allows smaller members to contribute content (and not host content for others) at a reduced fee[42]

Joining or forming a LOCKSS Network is arguably the cheapest and possibly safest method of bit-level preservation. Since long-term storage is one of the most important aspects of digital curation, and since keeping costs down is crucial for most institutions, participating in a LOCKSS Network is an excellent method to leverage the community to help meet your needs.

To ensure continued access to key collections for scholars, the Digital Preservation Network has developed a set of preservation services offered to academic libraries for a fee, with the intent of continuing to preserve submitted content even if an organization must leave the network (although after 20 years, funding must be obtained).[43] Submitting institutions may choose from the varying services offered by the Ingest Nodes; once ingested, the content is replicated at least twice to Replicating Nodes, where it is monitored for change and repaired, if necessary.[44] Initial membership fees are reportedly steep, although they include five new terabytes of content deposited per year for each of six years, and in 2015, additional terabyte storage was priced at $5,500 each.[45]

Researchers and scientists often need a way to track their files, data, and protocols in one secure location, where they can control levels of access and collaboration and be certain their content won't disappear or get lost. The Open Science Framework (OSF)[46] provides this service for free. Maintained and developed by the Center for Open Science, it is supported through grants and has established a preservation fund that will protect the data for more than fifty years if they lose funding.[47]

FOCUS ON TYPES OF CONTENT

Many organizations or alliances focus on specific types or aspects of content. As these groups appear almost continually, the reader is encouraged to perform web searches to locate current groups focused in the areas of interest. What follows is a synopsis of current efforts and leading organizations in the areas of web archiving; software and emulation; audiovisual content; social sciences research; Australian online publications; and clarifying authenticity, appraisal, and significant properties for born-digital documents.

Wikipedia lists numerous web archiving efforts dating back to 1995.[48] How do you do it? If you don't want to "roll your own," Archivethe.Net offers a collaborative service developed by the Internet Memory Foundation, through which organizations can select what to harvest and manage it, without having to build their own infrastructure.[49] Similarly, the Internet Archive offers Archive-It, a subscription web archiving service,[50] with a shared portal to content,[51] multiple shared projects,[52] and an annual partner meeting.[53]

With so many groups collecting web content, how do you avoid duplication? To answer this question, Harvard, UCLA, and the California Digital Library are collaborating to develop the IMLS-funded Cobweb service, which will allow stakeholders to nominate URLs of interest, claim what they plan to harvest, and describe the scope of what each partner has collected.[54] According to a July 2017 presentation at an Archive-It Partner Meeting, the official rollout of the service will be in 2018, at the IIPC Web Archiving Conference.[55] (The International Internet Preservation Consortium [IIPC] organizes web archiving efforts and is open to institutions with significant experience in web archiving.[56]) Before you launch your own web archive, check to see what others have already collected.

Some efforts are national, for instance, the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA)[57] project by the National Library of Australia. This is a collaborative effort to capture, archive, and provide long-term access to significant online publications.[58] Each partner establishes the portion of the Australian collection for which it will be responsible and uses PANDAS (the PANDORA Digital Archiving System) to contribute titles to the central archive.[59] The National Library developed PANDAS to provide the first available integrated, web-based web archiving management system.[60] The National Library of Australia provides persistent identifiers and resolution for titles contributed (and their components) and is committed to digital preservation and continued access for all holdings.[61]

If you are interested in collecting software, consider joining the Software Preservation Network[62] or at least signing up for their listserv.[63] To coordinate software preservation efforts, they have developed working groups focused on legal issues, metadata, research, curation readiness, and technical infrastructure.[64] Since software is needed to experience the files we are saving, efforts like this are part of the community support necessary to form the basis for emulation for access.[65]

Are you focused primarily on curating audio or audiovisual content? PrestoCentre is a community of audiovisual archives with a strong tradition in sharing knowledge about digital preservation.[66] Free bimonthly newsletters[67] include highlighted publications, news, job and event announcements, and calls for papers.[68] Many resources are provided for free, including a list of websites and LinkedIn groups, technical reports, standards and tools archives, a discussion of Hot Topics, and an up-to-date events calendar.[69] Maintained by the Netherlands Institute for Sound and Vision, PrestoCentre represents a worldwide network of professionals working with audiovisual cultural heritage materials.[70]

If you care about the data used to support social science research, consider becoming involved in the Data Preservation Alliance for the Social Sciences (Data-PASS). Begun in 2004 and originally funded by the National Digital Information Infrastructure and Preservation Program (NDIIP), this well-designed sustainable effort is now supported by research grants[71] and partner-contributed computing resources and personnel time; there are no membership fees.[72] Data-PASS partners share access to their collections via a union catalog, promote best practices, and safeguard collections through succession planning and live replication. The shared online catalog increases access to partners' archived collections, where users can search and browse most of the holdings, and download the publicly available studies.[73] Partners' content can be included via the OAI-PMH protocol,[74] by contributing to the Harvard Dataverse,[75] or by creating a virtual archive using the freely available Dataverse software.[76] The Data-PASS project also developed the openly available SafeArchive software, a "self-contained system that can be installed, used, and maintained by institutional staff without technical expertise."[77] SafeArchive is a policy-driven auditing tool for distributed replication of content[78] that supports the geographical duplication of content needed for digital preservation.

Determining the crucial aspects for electronic documents in terms of authenticity, appraisal, and significant properties is a pressing need for us all. How can we ensure that the document provided is the original? How do we determine its value? How do we figure out what needs to be retained for long-term access when migrating from one format to another? For problems like these, it is best to consult with the experts.

The International Research on Permanent Authentic Records Electronic Systems (InterPARES) project is an interdisciplinary, international research effort focused on solving the problems of authenticity, appraisal, preservation, and strategies for digital curation of electronic documents.[79] Now in its fourth phase, this partnership includes more than seventy institutional partners and more than 300 researchers and research assistants.[80] The first phase of the project focused on authenticity, and the results include a template for analysis, requirements for assessment, and a model for selection.[81] Phase two analyzed how to support, maintain, and preserve authentic records within interactive, dynamic information systems and developed policies, metadata, and tools for the functionality needed in these infrastructures.[82] Phase three focused on implementation of their findings in small and medium-sized archival organizations in twelve different countries[83]; details of their methodology and results are freely available.[84] The fourth phase has established InterPARES Trust, a research project to generate frameworks to develop local, national, and international policies, procedures, regulations, standards, and legislation that will ensure public trust via "good governance, a strong digital economy, and a persistent digital memory."[85] Their results include many online publications and presentations[86] that will be useful if you are in an administrative capacity or seeking to influence government decisions on any level.

FOCUS ON LEARNING AND CONNECTING

"Where do I start?" is a common question for those new to the field of digital curation. Joining an organization, reading journals, attending conferences, and signing up for a listserv or two are recommended. Once you start exploring tools and software options, consider getting involved in groups that are developing best practices and guidelines. If you are interested in laying the groundwork for standards, explore participation in NISO (National Information Standards Organization) working groups or ISO (International Organization for Standardization). NISO is nationally focused in the United States, and the standards they develop are freely available.[87] ISO is international, and their standards come at a cost.[88] Additional standards, like the Audio Engineering Society standard for audio metadata (AES57-2011),[89] are published by the National Institute of Standards and Technology (NIST), which is part of the U.S. Department of Commerce.[90]

To assist you in deepening your education, the following includes a brief list of organizations (two in the United States, two in the United Kingdom), journals for recommended reading, conferences that focus on digital curation, and a handful of listserv options that can help you get involved and stay abreast of changes in the field.

The Digital Preservation Outreach and Education (DPOE), an initiative of the Library of Congress, has grown from a national to an international train-the-trainer network.[91] As described in chapter 3, the curriculum is divided into six focused modules ("Identify," "Select," "Store," "Protect," "Manage," and "Provide") intended for practitioners in the field who are new to digital preservation.[92] The low-cost workshops are organized regionally, and new trainers are expected to share their new knowledge within six months following the training.[93] A listserv enables trainers to stay in touch and share their adventures and knowledge, building a network of digital preservation practitioners who can support one another and potentially build collaborations.

In the United Kingdom, the Digital Curation Centre (DCC)[94] leads the way with tremendous digital curation resources,[95] events,[96] trainings,[97] and a series of community development initiatives[98] intended to create a self-sustaining network of data practitioners. To build on their exceptional Curation Lifecycle Model[99] (described in chapter 3), DCC provides how-to guides and checklists for each area of curation,[100] an online reference manual,[101] and recommendations for developing research data platforms.[102] One of their best-known tools is DMPonline,[103] which provides researchers with tailored guidance for developing the data management plans now required by most funders. Although originally intended simply to support institutions in the United Kingdom, the DC now has international partners and an international impact, and it is an excellent resource for digital curators everywhere.

Also located in the United Kingdom, the Digital Preservation Coalition (DPC) focuses on advocacy, capacity building, partnership, and workforce building to support long-term access to digital content. Organizational membership brings voting rights and access to webinars, publications, assistance, and other resources; associate membership comes at a lower cost and has, accordingly, fewer perks.[104] Even nonmembers, however, can benefit from their excellent and openly available resources,[105] especially the technology reports and their online Digital Preservation Handbook. The technology reports include such topics as preserving with PDF/A, preserving social media, preserving transactional data, and personal digital archiving.[106] The Digital Preservation Handbook is a terrific overview of the challenges, strategies, activities, and solutions available for digital preservation.[107] At the time of this writing, the handbook was available both in HTML and PDF formats.

The National Digital Stewardship Alliance (NDSA) is a consortium of 165 organizations of various sorts that are working together to preserve access to U.S. cultural heritage materials.[108] Hosting listservs, an annual meeting and interest groups focused on content, infrastructure, and standards and practices, NDSA supports the open exchange of ideas, services, and software. An outgrowth of the National Digital Information Infrastructure Project (NDIIPP),[109] NDSA is hosted by the Digital Library Federation,[110] and a membership requires no dues or fees.[111]

Professionals in the field regularly report on their explorations and findings in publications. The current leading journal focused on digital curation alone is the openly available International Journal of Digital Curation (published by the University of Edinburgh and the DCC).[112] The DCC also maintains a list of other journals that often carry articles that relate to digital preservation and/or digital curation.[113] One of these, *D-Lib Magazine*, was a highly regarded leading resource for openly available digital curation research from 1995 to 2017, and the archives of this wealth of information are still hosted by the Corporation for National Research Initiatives.[114]

Conferences are an excellent place to learn about new and ongoing efforts, network and develop partnerships, and share your own research findings. Two of the best conference venues for digital curation are the International Digital Curation Conference[115] and Archiving, hosted by the Society for Imaging Science and Technology[116]; however, multiple other conferences regularly include research and presentations on findings and projects in the field. Larger ones focused primarily on research include the annual meeting of the Association for Information Science and Technology (ASIS&T)[117] and the Joint Conference on Digital Libraries (JCDL).[118] Many other conferences (a list of which is maintained by the University of Oregon[119]) also include related presentations. Some conferences, for example, the Coalition for Networked Information (CNI),[120] post presentations or resources on the web after the conference.

By regularly reviewing program pages on conference sites, you can not only get a better sense of the offerings and the community served, but also learn a great deal about new partnerships and findings in the field, without ever having to visit a conference. This is a wealth of resources for the eager learner, and most researchers are happy to respond if you have questions about their work. Even if all you can find is an abstract and the names of the presenters, do not hesitate to reach out to those involved in projects that interest you or may impact your own work. Conferences exist to share information and provide networking. Make use of their resources, even if you don't attend.

Listservs provide an immediate and easy method of both networking and discovering options, tools, and events. Some of the ones that focus primarily on digital curation and digital preservation include the following:

Digital Curation Google Group[121]
Digital Curation Centre Associates Network[122]

Digital-Preservation@jiscmail.ac.uk (for education and research communities in the United Kingdom)[123]
Digipres@lists.ala.org (American Library Association group)[124]

As you learn and explore, you will uncover new opportunities and discover not only what works, but also what doesn't work (e.g., some tools are not effective with some types of materials). Both types of information are valuable to others, as no one has all the answers. Only by communicating our findings to one another can we effectively move forward as a group. If you think about it, the history of today and yesterday depend upon our work; the foundations of future research are built upon what we put in place today. The importance of our success is in our own hands, and you are a key component in the process, so engage in the community and leverage what you find there. Then share your own findings via listserv discussions, conference presentations, articles, and community efforts. Join in the community: We need you.

KEY POINTS

1. Success in digital curation and digital preservation is dependent upon participation in the preservation/curation community.
2. Many projects, organizations, and consortiums are working to further our ability to select, preserve, and maintain access to digital content.
3. By staying in touch with the community through listservs, conferences, and journals, we can leverage the latest in research and development in our own work.
4. Review conference sites for abstracts and names of presenters who are working on projects that interest you; reach out to them and ask for more information.
5. By contributing our own findings to the broader community, we assist the community and our colleagues in developing better solutions for us all.
6. We can do far more together than alone. Participate!

Appendix

Resources

COURSES, TUTORIALS, AND FORMAL EDUCATION

North Carolina Department of Natural and Cultural Resources. “Digital Preservation Education: Tutorials.” Accessed February 18, 2018. http://digitalpreservation.ncdcr.gov/tutorials.html.

Northeast Document Conservation Center. “Preservation Education Curriculum: An Introduction to Preservation.” Accessed February 18, 2018. https://www.nedcc.org/free-resources/preservation-education-curriculum.

School of Information and Library Science, University of North Carolina at Chapel Hill. “Certificate in Digital Curation.” Accessed February 18, 2018. https://sils.unc.edu/programs/certificates/digital_curation.

Society of American Archivists. “Digital Archives Specialist (DAS) Curriculum and Certificate Program.” Accessed February 18, 2018. https://www2.archivists.org/prof-education/das.

Tibbo, Helen R., and Christopher (Cal) Lee. “DigCCurr Professional Institute: Curation Practices for the Digital Object Lifecycle.” Accessed February 18, 2018. https://ils.unc.edu/digccurr/institute.html.

University of North Carolina at Chapel Hill. “Professional Science Master’s: Digital Curation.” Accessed February 18, 2018. https://psm.unc.edu/digital-curation/.

W3C. “XML Tutorial.” Accessed April 14, 2017. https://www.w3schools.com/xml/.

BOOKS AND JOURNALS

General Digital Curation

Anonymous. “National Digital Stewardship Residency Programs.” Accessed February 18, 2018. https://ndsr-program.org/.

Cornell University Library, ICPSR, and MIT Libraries. “Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions. Timeline: Digital Technology and Preservation.” Online tutorial. Last modified 2013. Accessed March 5, 2017. http://www.dpworkshop.org/dpm-eng/timeline/viewall.html.

Data Conservancy. “Data Curation Curriculum Search.” Accessed February 18, 2018. http://cirss.ischool.illinois.edu/DCCourseScan1/.

Digital Preservation Coalition. *Digital Preservation Handbook*, 2nd ed. Accessed July 21, 2017. http://dpconline.org/handbook/.

Harvey, Ross. *Digital Curation: A How-to-Do-It Manual*. Chicago: Neal-Schuman, 2010.

Redwine, Gabriela, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas. *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*. Washington, DC: Council on Library and Information Resources, 2013. Accessed March 5, 2017. https://www.clir.org/pubs/reports/pub159.

University of Edinburgh and Digital Curation Centre. "International Journal of Digital Curation." Accessed September 24, 2017. http://www.ijdc.net/index.php/ijdc/index.

Metadata and Rights

Cullen, Charles T., Peter B. Hirtle, David Levy, Clifford A. Lynch, and Jeff Rothenberg. *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources, 2000. Pub 92. Accessed March 16, 2017. https://www.clir.org/pubs/reports/pub92.

Heath, Tom. "Linked Data—Connect Distributed Data across the Web: Guides and Tutorials." Accessed March 20, 2017. http://linkeddata.org/guides-and-tutorials.

Hirtle, Peter B., Emily Hudson, and Andrew T. Kenyon. *Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives, and Museums*. Ithaca, NY: Cornell University Library, 2009. Accessed March 17, 2017. https://ecommons.cornell.edu/bitstream/handle/1813/14142/Hirtle-Copyright_final_RGB_lowres-cover1.pdf.

World Intellectual Property Organization. "WIPO Intellectual Property Handbook: Policy, Law, and Use." Accessed February 18, 2018. http://www.wipo.int/about-ip/en/iprm/.

Access and Use

Tullis, Tom, and Bill Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Burlington, MA: Morgan Kaufmann, 2008.

STANDARDS, GUIDELINES, AND BEST PRACTICES

General Digital Curation

Consultative Committee for Space Data Systems. "Audit and Certification of Trusted Digital Repositories." Recommended Practice CCSDS 652.0-M-1. Last modified September 2011. Accessed March 5, 2017. https://public.ccsds.org/pubs/652x0m1.pdf.

———. "Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1 Blue Book." Last modified January 2002. Accessed March 5, 2017. https://siarchives.si.edu/sites/default/files/pdfs/650x0b1.PDF.

Cornell University Library, ICPSR, and MIT Libraries. "Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions." Online tutorial. Accessed March 5, 2017. http://www.dpworkshop.org/.

DeRidder, Jody L. "Introduction to Digital Preservation: A Three-Part Webinar Series, Based on the Library of Congress' Digital Preservation Outreach and Education (DPOE) Model." Accessed May 9, 2017. http://www.aserl.org/archive/#INTRODUCTION.

Digital Curation Centre. "DCC: Because Good Research Needs Good Data." Accessed September 24, 2017. http://www.dcc.ac.uk.

Kille, Leighton Walter. "The Growing Problem of Internet 'Link Rot' and Best Practices for Media and Online Publishers." Harvard Kennedy School Shorenstein Center Journalist's Resource. Last modified October 9, 2015. Accessed March 5, 2017. https://journalistsresource.org/studies/society/internet/website-linking-best-practices-media-online-publishers.

LeFurgy, Bill. "Steps in a Digital Preservation Workflow." Filmed March 7, 2012. Association for Library Collections and Technical Services. 58:48. Accessed April 23, 2017. http://www.ala.org/alcts/confevents/upcoming/webinar/030712.

Library of Congress. “Digital Preservation Outreach and Education.” Accessed May 7, 2017. http://www.digitalpreservation.gov/education/.

National Archives and Records Administration. “Strategic Directions: Appraisal Policy.” September 2007. Accessed June 1, 2017. https://www.archives.gov/records-mgmt/initiatives/appraisal.html.

National Digital Stewardship Alliance. “Levels of Digital Preservation.” Accessed May 28, 2017. http://ndsa.org/activities/levels-of-digital-preservation.

Snyder, James. “Preservation Workflows at the Library of Congress.” Filmed 2012. PrestoCentre . 1:09:31. Accessed April 23, 2017. https://www.prestocentre.org/resources/preservation-workflows-library-congress.

File Formats

Library of Congress. “Sustainability of Digital Formats: Planning for Library of Congress Collections.” Last modified March 10, 2017. Accessed April 12, 2017. https://www.loc.gov/preservation/digital/formats/index.html.

Metadata Standards and Encodings

Anonymous. “UTF-8 and Unicode.” Last modified August 28, 2014. Accessed April 12, 2017. http://www.utf-8.com/.

Audio Engineering Society. “AES57-2011: AES Standard for Audio Metadata—Audio Object Structures for Preservation and Restoration.” Accessed April 12, 2017. http://www.aes.org/publications/standards/search.cfm?docID=84.

Bekaert, Jeroen, Patrick Hochstenbach, and Herbert Van de Sompel. “Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library.” *D-Lib Magazine* 9 (November, 2003): 11. Accessed March 26, 2017. http://www.dlib.org/dlib/november03/bekaert/11bekaert.html.

Caplan, Priscilla. “Library of Congress Network Development and MARC Standards Office: Understanding PREMIS.” Library of Congress. February 1, 2009. Accessed April 16, 2017. http://www.loc.gov/standards/premis/understanding-premis.pdf.

DCMI Usage Board, Dublin Core Metadata Initiative. “DCMI Metadata Terms.” Last modified June 14, 2012. Accessed April 12, 2017. http://dublincore.org/documents/dcmi-terms/.

Dublin Core Metadata Initiative. “Dublin Core Metadata Element Set, Version 1.1.” Last modified June 14, 2012. Accessed April 12, 2017. http://dublincore.org/documents/dces/.

ECMA International. “Introducing JSON.” Accessed March 20, 2017. http://www.json.org/.

Federal Geographic Data Committee. “Geospatial Metadata Standards and Guidelines.” Accessed April 12, 2017. https://www.fgdc.gov/metadata/geospatial-metadata-standards.

FileFormat.Info. “UTF-8 Encoding.” Accessed April 14, 2017. http://www.fileformat.info/info/unicode/utf8.htm.

International Press Telecommunications Council. “IPTC Video Metadata Hub—Recommendation 1.0/Properties.” Last modified November 23, 2016. Accessed April 12, 2017. http://www.iptc.org/std/videometadatahub/recommendation/IPTC-VideoMetadataHub-props-Rec_1.0.html.

Knowledge Network for Biocomplexity. “Ecological Metadata Language (EML).” Accessed April 12, 2017. https://knb.ecoinformatics.org/#external//emlparser/docs/index.html.

Library of Congress. “EAD: Encoded Archival Description.” Accessed January 21, 2018. http://www.loc.gov/ead/.

———. “METS Metadata Encoding Transmission Standard.” Last modified August 9, 2016. Accessed March 18, 2017. http://www.loc.gov/standards/mets/.

———. “MIX: NISO Metadata for Images in XML Schema, Technical Metadata for Digital Still Images Standard.” Last modified November 23, 2015. Accessed March 18, 2017. https://www.loc.gov/standards/mix/.

———. "MODS Metadata Object Description Schema." Accessed March 18, 2017. http://www.loc.gov/standards/mods/.

———. "TextMD Technical Metadata for Text." Last modified April 12, 2017. Accessed March 18, 2017. https://www.loc.gov/standards/textMD/.

Open Geospatial Consortium. "Geography Markup Language." Accessed March 18, 2017. http://www.opengeospatial.org/standards/gml.

Planets. "XCL—eXtensible Characterization Language." Accessed April 19, 2017. http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl.html.

Powell, Andy, and Pete Johnston, Dublin Core Metadata Initiative. "Guidelines for Implementing Dublin Core in XML." Last modified April 4, 2003. Accessed March 18, 2017. http://dublincore.org/schemas/xmls/.

PREMIS Editorial Committee. "Conformant Implementation of the PREMIS Data Dictionary." Accessed April 16, 2017. http://www.loc.gov/standards/premis/premis-conformance-20150429.pdf.

———. "PREMIS Data Dictionary for Preservation Metadata, Version 3.0." Last modified November 2015. Accessed April 16, 2017. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

Society of Motion Picture and Television Engineers. "ST 377-1:2011—SMPTE Standard—Material Exchange Format (MXF)—File Format Specification." June 7, 2011. doi: 10.5594/SMPTE.ST377-1.2011.

Staatsbibliothek zu Berlin and Society of American Archivists. "EAC-CPF." Accessed January 21, 2018. http://eac.staatsbibliothek-berlin.de/.

TDWG Biodiversity Information Standards. "Darwin Core." Last modified June 5, 2015. Accessed April 16, 2017. http://rs.tdwg.org/dwc/.

Unicode, Inc. "General Information: What Is Unicode?" Last modified December 1, 2015. Accessed April 16, 2017. http://www.unicode.org/standard/WhatIsUnicode.html.

W3C. "ASCII Codes Table: Standard Characters." Accessed March 26, 2017. http://ascii.cl/.

———. "Date and Time Formats." Last modified September 15, 1997. Accessed April 16, 2017. https://www.w3.org/TR/NOTE-datetime.

———. "Extensible Markup Language (XML) 1.0 (Fifth Edition)." Last modified February 7, 2013. Accessed April 16, 2017. https://www.w3.org/TR/xml/#NT-Name.

———. "Linked Data." Accessed March 20, 2017. https://www.w3.org/standards/semanticweb/data.

———. "Provenance Current Status." Accessed May 8, 2017. https://www.w3.org/standards/techs/provenance.

———. "RDF/XML Syntax Specification (Revised)." Last updated February 10, 2004. https://www.w3.org/TR/REC-rdf-syntax/.

———. "W3C Math Home: What is MathML?" Accessed August 6, 2017. https://www.w3.org/Math/.

W3Schools. "XML Namespaces." Accessed March 26, 2017. https://www.w3schools.com/XML/xml_namespaces.asp.

Unicode, Inc. "General Information: What Is Unicode?" Last modified December 1, 2015. Accessed April 16, 2017. http://www.unicode.org/standard/WhatIsUnicode.html.

Rights

EUREC. "European Network of Research Ethics Committees." Accessed January 14, 2018. Accessed April 16, 2017. http://www.eurecnet.org/index.html.

Lifshitz-Goldberg, Yael. "Orphan Works: Lecture Summary." World Intellectual Property Organization Seminar. May 2010. Accessed March 18, 2017. http://www.wipo.int/edocs/mdocs/sme/en/wipo_smes_ge_10/wipo_smes_ge_10_ref_theme11_02.pdf.

SAA Intellectual Property Working Group, Society of American Archivists. "Guide to Implementing Rights Statements from RightsStatements.org." December 2016. Accessed April 22, 2017. http://www2.archivists.org/sites/all/files/RightsStatements_IPWG%20Guidance.pdf.

Society of American Archivists. “Orphan Works: Statement of Best Practices.” Last modified June 17, 2009. Accessed April 16, 2017. http://www.archivists.org/standards/OWBP-V4.pdf.

U.S. Copyright Office. “Copyright: Public Catalog.” Accessed March 18, 2017. http://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First.

U.S. Department of Health and Human Services. “Clinical Trials and Human Subject Protection.” Last modified September 25, 2017. Accessed April 22, 2017. https://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/default.htm.

U.S. Department of Health and Human Services Office for Human Research Protections. “International Compilation of Human Research Standards.” 2017 edition. Accessed January 14, 2017. https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part1.pdf and https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part2.pdf.

Storage and Protection

Center for Research Libraries. “Digital Preservation Metrics.” Accessed February 18, 2018. https://www.crl.edu/archiving-preservation/digital-archives/metrics.

International Organization for Standardization. “ISO 16363:2012, Space Data, and Information Transfer Systems—Audit and Certification of Trusted Digital Repositories.” Last modified February 2012. Accessed February 18, 2018. https://www.iso.org/standard/56510.html.

Thesauri and Controlled Vocabularies

Getty Research Institute. “Art and Architecture Thesaurus Online.” Accessed March 18, 2017. http://www.getty.edu/research/tools/vocabularies/aat/.

———. “Getty Thesaurus of Geographic Names Online.” Accessed March 18, 2017. http://www.getty.edu/research/tools/vocabularies/tgn/index.html.

Library of Congress. “ISO 639.2: Codes for the Representation of Names of Languages.” Last modified July 25, 2013. Accessed March 18, 2017. https://www.loc.gov/standards/iso639-2/php/code_list.php.

———. “LC Linked Data Service: Authorities and Vocabularies.” Accessed March 18, 2017. http://id.loc.gov/index.html.

———. “Library of Congress Names.” Accessed March 18, 2017. http://id.loc.gov/authorities/names.html.

———. “Library of Congress Subject Headings.” Accessed March 18, 2017. http://id.loc.gov/authorities/subjects.html.

———. “Thesaurus for Graphic Materials.” Accessed March 18, 2017. http://www.loc.gov/pictures/collection/tgm/.

OCLC. “FAST (Faceted Application of Subject Terminology).” Accessed March 18, 2017. http://fast.oclc.org/.

———. “VIAF: The Virtual International Authority File.” Accessed March 18, 2017. https://viaf.org/.

TOOLS AND REGISTRIES

Atlassian JIRA. “Conversion Software Registry.” Accessed May 9, 2017. http://isda.ncsa.uiuc.edu/NARA/CSR/php/search/conversions.php.

bwFLA. “Emulation as a Service.” Accessed August 5, 2017. http://bw-fla.uni-freiburg.de/eaas.html.

Delve, Janet, and David Anderson. “KEEP: Welcome to TOTEM—the Trustworthy Online Technical Environment Metadata Registry.” Accessed August 5, 2017. http://www.keep-totem.co.uk/.

DigiPres Commons. “Community-Owned Digital Preservation Resources.” Accessed September 17, 2017. http://www.digipres.org/.

Harvard University. “File Information Tool Set (FITS).” Accessed April 12, 2017. https://projects.iq.harvard.edu/fits/home.

JSTOR and the President and Fellows of Harvard College. “JHOVE: JHOVE-JSTOR/Harvard Object Validation Environment.” Last modified February 25, 2009. Accessed May 9, 2017. http://jhove.sourceforge.net/.

Library of Congress. “METS Implementation Registry.” Last modified August 29, 2016. Accessed May 9, 2017. http://www.loc.gov/standards/mets/mets-registry.html.

National Archives. “Download DROID: File Format Identification Tool.” Accessed April 12, 2017. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

———. “The Technical Registry PRONOM.” Accessed April 12, 2017. https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

National Library of New Zealand. “Digital Preservation Programme: Digital Preservation Technical Registry.” Accessed October 8, 2017. https://digitalpreservation.natlib.govt.nz/current-projects/technical-registry/.

Olive Archive. “Olive Executable Archive.” Accessed August 5, 2017. https://olivearchive.org/.

Open Preservation Foundation. “Community-Owned Digital Preservation Tool Resource (COPTR).” Accessed March 5, 2017. http://coptr.digipres.org.

University of North Carolina School of Information and Library Science. “BitCurator: About the Project.” Accessed June 1, 2017. https://www.bitcurator.net/bitcurator/.

VeraPDF Consortium. “Industry Supported PDF/A Validation: VeraPDF.” Accessed February 18, 2018. http://verapdf.org/.

CONFERENCES, ORGANIZATIONS, AND COLLABORATIVES

Conferences

National Digital Stewardship Alliance and Digital Library Federation. “Digital Preservation Conferences.” Accessed February 18, 2018. http://ndsa.org/meetings/.

SUB Göttingen. International Conference on Digital Preservation. Accessed February 18, 2018. https://ipres-conference.org/.

Organizations and Collaboratives

Alabama Digital Preservation Network. “The Alabama Digital Preservation Network: Preserving Alabama’s Digital Resources.” Accessed September 17, 2017. http://www.adpn.org/.

Anonymous. “Software Preservation Network.” Accessed April 12, 2017. http://www.softwarepreservationnetwork.org/.

CLOCKSS. “The CLOCKSS Archive: A Trusted Community-Governed Archive.” Accessed September 17, 2017. https://www.clockss.org/clockss/Home.

Council on Library and Information Resources. “National Digital Stewardship Alliance: Overview.” Accessed September 17, 2017. http://coherence.clir.org/almanac/fact-sheet-gallery/national-digital-stewardship-alliance.

Data Preservation Alliance for the Social Sciences. “About Data-PASS.” Accessed September 24, 2017. http://www.data-pass.org.

DataONE.org. “DataONE: Data Observation Network for Earth.” Accessed May 8, 2017. https://www.dataone.org/.

Digital Preservation Network. “Preserving the Historical Record for This and Future Generations.” Accessed September 24, 2017. https://dpn.org/about.

Educopia Institute. "BitCurator Consortium." Accessed September 17, 2017. https://bitcuratorconsortium.org/mission.

———. "MetaArchive Cooperative." Accessed May 26, 2017. http://www.metaarchive.org/.

Internet Archive. "Archive-It: About Us." Accessed October 8, 2017. https://archive-it.org/learn-more/.

Internet Memory Foundation. "Archivethe.Net (AtN): A Shared Web Archiving Platform Operated by Internet Memory." Accessed October 8, 2017. http://internetmemory.org/en/index.php/projects/atn.

InterPARES. "InterPARES Trust." Accessed September 10, 2017. https://interparestrust.org/trust.

Open Archives Initiative. "Open Archives Initiative Protocol for Metadata Harvesting." Accessed September 10, 2017. https://www.openarchives.org/pmh/.

Open Preservation Foundation. "DigiPres Commons: Community-Owned Digital Preservation Resources." Accessed June 1, 2017. http://www.digipres.org/.

Open Science Framework. "Open Science Framework: A Scholarly Commons to Connect the Entire Research Cycle." Accessed September 12, 2017. https://osf.io/.

PrestoCentre. "PrestoCentre Serves the Audiovisual Communities Interested in Digitization and Digital Preservation Worldwide. We Are Open and Free." Accessed September 24, 2017. https://www.prestocentre.org.

Software Preservation Network. "Software Preservation Network." Accessed October 8, 2017. http://www.softwarepreservationnetwork.org/about/.

U.S. National Archives and Records Administration. "Social Networks and Archival Context Cooperative (SNAC)." Accessed January 21, 2018. http://snaccooperative.org/.

W3C. "ODRL Community Group." Accessed March 16, 2017. https://www.w3.org/community/odrl/.

Notes

CHAPTER 1

1. Fact Monster from Information Please, “Life-Changing Science Discoveries,” *Factmonster.com*, last modified 2017, accessed March 6, 2017, http://www.factmonster.com/ipka/A0932440.html.

2. National Cancer Institute, “Milestones in Cancer Research and Discovery,” *National Cancer Institute*, last modified January 1, 2015, accessed March 6, 2017, https://www.cancer.gov/research/progress/250-years-milestones.

3. J. G. Meijer, “Librarianship: A Definition,” University of Illinois Graduate School of Library and Information Science Occasional Papers 155, September 1982, 26, accessed March 6, 2017, http://hdl.handle.net/2142/3979.

4. Matt Rocheleau, “Why Are WhiteHouse.gov Web Pages Disappearing?” *Boston Globe*, January 20, 2017, accessed March 6, 2017, https://www.bostonglobe.com/metro/2017/01/20/why-are-whitehouse-gov-web-pages-disappearing/gd0HEAAU49hrLZMCiOQwuN/story.html.

5. Alex Wellerstein, “Redactions: The Year of the Disappearing Websites,” Restricted Data: The Nuclear Secrecy (blog), December 27, 2013, accessed March 6, 2017, http://blog.nuclearsecrecy.com/2013/12/27/year-disappearing-websites/.

6. Scott Althaus and Kalev Leetaru, “Airbrushing History, American Style,” *Cline Center for Democracy*, November 25, 2008, accessed March 6, 2017, http://www.clinecenter.illinois.edu/research/affiliated/airbrush/.

7. Mike Ashenfelder, “The Average Lifespan of a Web Page,” The Signal (blog), Library of Congress, November 8, 2011, accessed March 6, 2017, https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-web-page/.

8. NPR Staff, “Stopping Link Rot: Aiming to End a Virtual Epidemic,” *National Public Radio*, April 26, 2014, accessed March 7, 2017, http://www.npr.org/sections/alltechconsidered/2014/04/26/307041846/stopping-link-rot-aiming-to-end-a-virtual-epidemic.

9. Palab Ghosh, “Google’s Vint Cerf Warns of ‘Digital Dark Age,’” *BBC News Science and Environment*, February 13, 2015, accessed March 8, 2016, http://www.bbc.com/news/science-environment-31450389.

10. Committee on Future Career Opportunities and Educational Requirements for Digital Curation, National Research Council of the National Academies, *Preparing the Workforce for Digital Curation* (Washington, DC: National Academies Press, 2015), 10, accessed April 16, 2017, https://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation.

11. Grant Burningham, “Father of the Internet Worries Our Digital History Is Disappearing,” *Newsweek Tech & Science*, June 16, 2016, accessed March 6, 2017, http://www.newsweek.com/father-internet-worries-our-digital-history-disappearing-468642.

12. Elizabeth Yakel, “Digital Curation,” *OCLC Systems and Services: International Digital Library Perspectives* 23 (2007): 4, 338, accessed March 6, 2017, doi: 10.1108/10650750710831466; Sarah Higgins, “The DCC Curation Lifecycle Model,” *International Journal of Digital Curation* 1 (2008): 3, 137–38, accessed March 6, 2017, doi: 10.2218/ijdc.v3i1.48.

13. Cornell University Library, ICPSR, and MIT Libraries, “Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions. Time Line: Digital Technology and Preservation,” online tutorial, *Dpworkshop.org*, last modified 2013, accessed March 5, 2017, http://www.dpworkshop.org/dpm-eng/time line/viewall.html.

14. Cornell University Library, ICPSR, and MIT Libraries, “Digital Preservation Management.”

15. Wikipedia, “Time Line of Microsoft Windows,” *Wikipedia*, last modified February 25, 2017, accessed March 5, 2017, https://en.wikipedia.org/wiki/Timeline_of_Microsoft_Windows.

16. RDA Steering Committee, “About RDA,” *Rda-rsc.org*, last modified April 14, 2017, accessed March 5, 2017, http://www.rda-rsc.org/content/about-rda.

17. American Library Association, Canadian Library Association, and the Chartered Institute of Library and Information Professionals, “AACR: Welcome to the Homepage of the Anglo-American Cataloguing Rules,” *AACR2.org*, accessed April 28, 2017, http://www.aacr2.org/.

18. Library of Congress, “EAD: Encoded Archival Description,” *Loc.gov*, accessed January 21, 2018, https://www.loc.gov/ead/.

19. Society of American Archivists, “Describing Archives: A Content Standard (DACS),” *Archivists.org*, accessed April 28, 2017, http://www.archivists.org/governance/standards/dacs.asp.

20. Peter B. Hirtle, “Archival Authenticity in a Digital Age,” in Charles T. Cullen, Peter B. Hirtle, David Levy, Clifford A. Lynch, and Jeff Rothenberg, *Authenticity in a Digital Environment*, Pub 92 (Washington, DC: Council on Library and Information Resources, 2000), accessed March 16, 2017, https://www.clir.org/pubs/reports/pub92/hirtle.html.

21. Jason Curtis, “Museum of Obsolete Media,” *Obsoletemedia.org*, accessed April 30, 2017, http://www.obsoletemedia.org/.

22. “OAIster . . . Find the Pearls,” *OAIster*, accessed March 16, 2017, http://archives.getty.edu:30008/o/oaister/.

23. “OAIster: Find the Pearls,” accessed March 16, 2017, http://oaister.worldcat.org/.

24. Kat Hagedorn, “OAIster: A ‘No Dead Ends’ Digital Object Service,” Library and Information Technology Association (LITA) National Forum, Norfolk, VA, October 3, 2003, accessed March 16, 2017, http://archives.getty.edu:30008/o/oaister/pres/LITA03_Hagedorn.ppt.

CHAPTER 2

1. Library of Congress, “Sustainability of Digital Formats: Planning for Library of Congress Collections. PDF/A-1a, PDF for Long-Term Preservation, Use of PDF 1.4, Level A Conformance,” *Loc.gov*, last modified March 10, 2017, accessed October 22, 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000251.shtml.

2. Library of Congress, “Sustainability of Digital Formats.”

3. Library of Congress, “Personal Archiving: Preserving Your Digital Memories,” *Loc.gov*, accessed October 15, 2017, http://www.digitalpreservation.gov//personalarchiving/records.html.

4. Gabriela Redwine, “Personal Digital Archiving: DPC Technology Watch Report 15-01 December 2015,” *DPConline.org*, accessed October 22, 2017, https://www.dpconline.org/docs/technology-watch-reports/1460-twr15-01/file.

5. Jody DeRidder, University of Alabama, “Recommendations for Authors and Creators,” *Lib.ua.edu*, accessed October 22, 2017, http://www.lib.ua.edu/wiki/digcoll/index.php/Recommendations_for_Authors_and_Creators.

6. North Carolina State Archives, Natural and Cultural Resources, “Digital Records Policies and Guidelines,” *NCDCR.gov*, accessed October 22, 2017, http://archives.ncdcr.gov/For-Government/Digital-Records/Digital-Records-Policies-and-Guidelines.

7. Creative Commons, “About the Licenses: What the Licenses Do,” *Creativecommons.org*, accessed October 15, 2017, https://creativecommons.org/licenses/.

8. Internet Corporation for Assigned Names and Numbers, “ICANN WHOIS,” *Whois.icann.org*, accessed December 10, 2017, https://whois.icann.org/en/lookup.

9. Code4Lib Journal, “Mission,” *Code4lib.org*, accessed December 10, 2017, http://journal.code4lib.org/mission.

10. Digital Library Federation, “What’s the DLF?” *Diglib.org*, accessed December 10, 2017, https://www.diglib.org/.

11. Society of American Archivists, “SAA Annual Meeting,” *Archivists.org*, accessed December 10, 2017, https://www2.archivists.org/conference.

12. Digital Library Federation, “DLF Groups,” *Diglib.org*, accessed December 10, 2017, https://www.diglib.org/groups/.

13. National Information Standards Organization, “NISO,” *NISO.org*, accessed December 10, 2017, http://www.niso.org/.

14. Indiana University, “Media Digitization and Preservation Initiative,” *Mdpi.iu.edu*, accessed October 15, 2017, https://mdpi.iu.edu/.

15. Alabama Digital Preservation Network, “About ADPNet,” *ADPN.org*, accessed December 10, 2017, http://www.adpn.org/.

16. Council on Library and Information Resources, “CLIR,” *CLIR.org*, accessed December 10, 2017, https://www.clir.org/.

17. National Endowment for the Humanities, “National Endowment for the Humanities,” *NEH.gov*, accessed December 10, 2017, https://www.neh.gov/.

18. U.S. National Archives and Records Administration, “National Historical Publications and Records Commission,” *Archives.gov*, accessed December 10, 2017, https://www.archives.gov/nhprc.

19. Andrew W. Mellon Foundation, “AWMF,” *Mellon.org*, accessed December 10, 2017, https://mellon.org/.

20. Institute of Museum and Library Services, “Institute of Museum and Library Services,” *IMLS.gov*, accessed December 10, 2017, https://www.imls.gov/.

21. Coalition for Networked Information, “CNI: Coalition for Networked Information,” *CNI.org*, accessed December 10, 2017, https://www.cni.org/.

22. Big Ten Academic Alliance, “BIG Academic Alliance,” *BTAA.org*, accessed October 15, 2017, https://www.btaa.org/home.

23. Orbis Cascade Alliance, “Push Boundaries, Change the Landscape, and Inspire the Profession: Alliance Strategic Agenda,” *Orbiscascade.org*, accessed October 15, 2017, https://www.orbiscascade.org/.

24. Digital POWRR, “Preserving (Digital) Objects with Restricted Resources,” accessed October 15, 2017, http://digitalpowrr.niu.edu/.

25. Society of American Archivists, “SAA: Society of American Archivists,” *Archivists.org*, accessed October 22, 2017, https://www2.archivists.org/.

26. Digital Library Federation, “DLF: Digital Library Federation,” *Diglib.org*, accessed October 22, 2017, https://www.diglib.org/.

27. National Digital Stewardship Alliance and Digital Library Federation, “National Digital Stewardship Alliance,” *NDSA.org*, accessed October 22, 2017, http://ndsa.org/.

28. Best Practices Exchange, “Best Practices Exchange: Information about the Best Practices Exchange Conferences,” *Bpexchange.wordpress.com*, accessed October 22, 2017, https://bpexchange.wordpress.com/.

29. Digital Curation Centre, “Community of Digital Curators,” *DCC.ac.uk*, accessed October 22, 2017, http://www.dcc.ac.uk/community.

30. National Information Standards Organization, “Welcome to NISO,” *NISO.org*, accessed October 22, 2017, http://www.niso.org.

31. U.S. Department of Commerce, "NIST: National Institute of Standards and Technology," *NIST.gov*, accessed October 22, 2017, https://www.nist.gov/.

32. International Organization for Standardization, "ISO: International Organization for Standardization," *ISO.org*, accessed October 22, 2017, https://www.iso.org/home.html.

33. Internet Archive, "Frequently Asked Questions: Uploading Content," *Archive.org*, accessed December 10, 2017, https://archive.org/about/faqs.php#Uploading_Content.

CHAPTER 3

1. Digital Curation Centre, "DCC Curation Lifecycle Model," *DCC.ac.uk*, accessed May 3, 2017, http://www.dcc.ac.uk/resources/curation-lifecycle-model.

2. Digital Curation Centre, "DCC Curation Lifecycle Model."

3. Digital Curation Centre, "Lifecycle Model FAQ," *DCC.ac.uk*, accessed May 3, 2017, http://www.dcc.ac.uk/resources/curation-lifecycle-model/lifecycle-model-faqs.

4. Digital Curation Centre, "Lifecycle Model FAQ."

5. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1 Blue Book," *Si.edu*, last modified January 2002, https://siarchives.si.edu/sites/default/files/pdfs/650x0b1.PDF.

6. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System," 65.

7. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System," 70–71.

8. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System," 89.

9. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System," 148.

10. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System," 148.

11. Library of Congress, "Digital Preservation Outreach and Education," *Digitalpreservation.gov*, accessed May 7, 2017, http://www.digitalpreservation.gov/education/.

12. Library of Congress, "Digital Preservation Outreach and Education: DPOE Curriculum," *Digitalpreservation.gov*, accessed May 7, 2017, http://www.digitalpreservation.gov/education/curriculum.html.

13. Jody L. DeRidder, "An Introduction to Digital Preservation: A Three-Part Webinar Series, Based on the Library of Congress' Digital Preservation Outreach and Education (DPOE) Model," *ASERL.org*, accessed May 9, 2017, http://www.aserl.org/archive/#INTRODUCTION.

CHAPTER 4

1. Chris Hoffman, "Beginner Geek: How to Create and Use Virtual Machines," *Howtogeek.com*, September 8, 2014, accessed August 7, 2017, https://www.howtogeek.com/196060/beginner-geek-how-to-create-and-use-virtual-machines/.

2. Jack Wallen, "The Best Linux Distribution for New Users," *Linux.com*, June 6, 2014, accessed August 7, 2017, https://www.linux.com/news/best-linux-distribution-new-users.

3. GNOME Project, "Evince," *Wiki.gnome.org*, accessed August 7, 2017, https://wiki.gnome.org/Apps/Evince.

4. Audacity Team, "Audacity," *Audacityteam.org*, last modified March 17, 2017, accessed August 7, 2017, http://www.audacityteam.org/.

5. VideoLAN, "VLC Media Player," *Videolan.org*, accessed August 7, 2017, http://www.videolan.org/vlc/.

6. SourceForge, "LAME," *LAME.SourceForge.net*, accessed August 7, 2017, http://lame.SourceForge.net/.

7. GIMP Team, "GIMP: GNU Image Manipulation Program," *GIMP.org*, accessed August 7, 2017, https://www.gimp.org/.

8. Anonymous, "FFmpeg," *FFmpeg.org*, accessed August 7, 2017, http://www.ffmpeg.org/.

9. Bram Moolenaar, "Vim: The Ubiquitous Text Editor," *vim.org*, accessed August 7, 2017, http://www.vim.org/.

10. Apache Software Foundation, "Apache OpenOffice: The Free and Open Productivity Suite," *OpenOffice.org*, accessed August 7, 2017, http://www.openoffice.org/.

11. SAS Institute, "Byte Ordering on Big-Endian and Little-Endian Platforms," *V8doc.sas.com*, accessed August 7, 2017, https://v8doc.sas.com/sashtml/lgref/z1270373.htm.

12. Jeffrey van der Hoeven, "Dioscuri: Emulation for Digital Preservation," paper presented at the annual meeting of Planets, CASPAR, and DPE, Lisbon, Portugal, September 5–6, 2007, 5, accessed August 4, 2017, http://www.planets-project.eu/docs/presentations/2007-09-05_emulation_wepreserve_portugal_jrvanderhoeven.ppt.

13. Sofie Laier Henriksen, Wiel Seuskens, and Gaby Wijers, "Digitizing Contemporary Art: D6.1 Guidelines for a Long-Term Preservation Strategy for Digital Reproductions and Metadata," Rev. 1.0, *DCA-project.eu*, February 13, 2012, accessed August 7, 2017, http://www.dca-project.eu/images/uploads/varia/DCA_D61_Guidelines_Long_Term_Preservation_Strategy_20120213_V1.pdf.

14. Henriksen, Seuskens, and Wijers, "Digitizing Contemporary Art," 8.

15. Digital Curation Centre, "Dioscuri," *DCC.ac.uk*, accessed August 5, 2017, http://www.dcc.ac.uk/resources/external/Dioscuri.

16. Koninklijke Bibliotheek, National Archief, Planets, and KEEP, "Dioscuri—the Modular Emulator," *SourceForge.net*, accessed August 5, 2017, http://dioscuri.SourceForge.net/.

17. Bram Lohman, Jeffrey van der Hoeven and Edo Noordermeer, "KEEP Emulation Framework System User Guide Release 2.0.0 (February 2012)," accessed July 4, 2017, available from http://emuframework.sourceforge.net/docs/System-User-Guide_2.0.pdf.

18. Digital Preservation Coalition, "DPA 2012: DPC Award for Research and Innovation—Finalists," *DPConline.org*, last modified August 12, 2016, accessed September 24, 2017, http://www.dpconline.org/events/dpa-2012-dpc-award-for-research-and-innovation-finalists.

19. SourceForge, "KEEP Emulation Framework (EF)," accessed August 5, 2017, http://emuframework.SourceForge.net/.

20. Digital Curation Centre, "KEEP Emulation Framework," *DCC.ac.uk*, accessed August 5, 2017, http://www.dcc.ac.uk/resources/external/keep-emulation-framework.

21. KEEP (Keeping Emulation Environments Portable), "Emulation Framework: Architectural Design Document, Release 2.0.0 (February 2012)," *EMUframework.SourceForge.net*, accessed August 5, 2017, http://emuframework.SourceForge.net/docs/Architectural-Design-Document_2.0.pdf.

22. David S. H. Rosenthal, "Emulation and Virtualization as Preservation Strategies," Andrew Mellon Foundation, 2015, 9, *Mellon.org*, accessed August 5, 2017, https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf.

23. Olive Archive, "What Is Olive?" *Olivearchive.org*, accessed August 5, 2017, https://olivearchive.org/about/.

24. Olive Archive, "What Is Olive?" 7.

25. bwFLA, "Emulation as a Service—Demo Page," *Demo.eaas.uni-freiburg.de*, accessed August 8, 2017, http://demo.eaas.uni-freiburg.de/.

26. bwFLA, "Emulation as a Service."

27. bwFLA, "Emulation as a Service."

28. Olive Archive, "Olive Executable Archive," *Olivearchive.org*, accessed August 5, 2017, https://olivearchive.org/.

29. Olive Archive, "Virtual Machines in Our Collection," *Olivearchive.org*, accessed August 5, 2017, https://olivearchive.org/docs/collection/.

30. Janet Delve and David Anderson, "KEEP: Welcome to TOTEM—the Trustworthy Online Technical Environment Metadata Registry," *Keep-totem.co.uk*, accessed August 5, 2017, http://www.keep-totem.co.uk/.

31. Janet Delve, Leo Konstantelos, and Antonio Ciuffreda, "TOTEM: Trusted Online Technical Environment Metadata—A Long-Term Solution for a Relational Database/RDF Ontologies," paper presented at the annual meeting of iPRES, Singapore, November 1–4, 2011, 2, accessed August 6, 2017, https://fedora.phaidra.univie.ac.at/fedora/objects/o:294265/methods/bdef:Content/get.

32. Delve, Konstantelos, and Ciuffreda, "TOTEM," 2.

33. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," *Loc.gov*, last modified March 10, 2017, accessed October 22, 2017, https://www.loc.gov/preservation/digital/formats/index.html.

34. Gareth Knight, "Same as It Ever Was? Significant Properties and the Preservation of Meaning over Time," paper presented at *Decoding the Digital: A Common Language for Preservation*, London, July 27, 2010, accessed August 7, 2017, http://www.dpconline.org/docs/miscellaneous/events/486-decodingknight-pdf/file.

35. W3C, "W3C Math Home: What Is MathML?" *W3.org*, accessed August 6, 2017, https://www.w3.org/Math/.

36. SourceForge, "Apache OpenOffice," *SourceForge.net*, accessed August 6, 2017, https://Source-forge.net/projects/openofficeorg.mirror/.

37. SUB Göttingen, "International Conference on Digital Preservation," accessed August 6, 2017, https://ipres-conference.org/.

38. Society for Imaging Science and Technology, "Archiving 2017," *Imaging.org*, accessed August 6, 2017, http://www.imaging.org/site/IST/Conferences/Archiving/IST/Conferences/Archiving/Archiving_Home.aspx.

39. Digital Curation Centre, International Digital Curation Conference (IDCC), accessed August 6, 2017, http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc.

40. Rosenthal, "Emulation and Virtualization as Preservation Strategies," 13.

CHAPTER 5

1. Jody L. DeRidder, "An Introduction to Digital Preservation: Steps to Identify and Select Content," 20, *ASERL.org*, February 7, 2011, accessed June 1, 2017, http://www.aserl.org/wp-content/uploads/2012/02/Intro_Dig_Pres_1.pdf.

2. National Archives and Records Administration, "Strategic Directions: Appraisal Policy," *Archives.gov*, September 2007, accessed June 1, 2017, https://www.archives.gov/records-mgmt/initiatives/appraisal.html.

3. DeRidder, "An Introduction to Digital Preservation," 14.

4. Library of Congress, "Digital Preservation Outreach and Education: DPOE Curriculum," *Digitalpreservation.gov*, accessed May 7, 2017, http://www.digitalpreservation.gov/education/curriculum.html.

5. DeRidder, "An Introduction to Digital Preservation," 21.

6. DeRidder, "An Introduction to Digital Preservation," 21.

7. Jody L. DeRidder and Alissa Matheny Helms, "Intake of Digital Content: Survey Results from the Field," *D-Lib Magazine* 22, no. 11/12 (November/December 2016), accessed June 1, 2017, doi:10.1045/november2016-deridder.

8. DeRidder and Helms, "Intake of Digital Content."

9. University of North Carolina School of Information and Library Science, "BitCurator: About the Project," *Bitcurator.net*, accessed June 1, 2017, https://www.bitcurator.net/bitcurator/.

10. Open Preservation Foundation, "DigiPres Commons: Community-Owned Digital Preservation Resources," *Digipres.org*, accessed June 1, 2017, http://www.digipres.org/.

11. Open Preservation Foundation, "Community-Owned Digital Preservation Tool Registry (COPTR)," *Digipres.org*, accessed June 28, 2017, http://coptr.digipres.org/.

12. Niels Brügger, "Archiving Websites: General Considerations and Strategies," *Centre for Internet Research*, Århus, Denmark January 2005, 10–11, accessed July 17, 2017, http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf.

13. Jinfang Niu, "An Overview of Web Archiving," *D-Lib Magazine* 18, no. 3/4 (March/April 2012), accessed June 1, 2017, doi:10.1045/march2012-niu1.

14. Jeremy Floyd, "Saving Digital Ephemera: Strategies for Preserving University Websites, Blogs, and Social Media," in *A Practical Approach to Web Archiving: Applying Archival Theory to Websites and Social Media Pages*, paper presented at the annual meeting of the Society of Southwest Archivists, May 23, 2013, 22, accessed June 1, 2017, https://societyofsouthwestarchivists.wildapricot.org/Resources/Documents/SSA2013Presentations/Weddle-How%20do%20you%20Archive%20the%20Web.pdf.

15. Floyd, "Saving Digital Ephemera," 34.

16. International Organization for Standardization, "ISO 28500:2009 Information and Documentation—WARC File Format," *ISO.org*, accessed July 19, 2017, https://www.iso.org/standard/44717.html.

17. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—WARC, Web ARChive File Format," *Loc.gov*, accessed July 19, 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.

18. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—WARC, Web ARChive File Format."

19. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—WARC, Web ARChive File Format."

20. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—ARC_IA, Internet Archive ARC File Format," *Loc.gov*, accessed July 19, 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml.

21. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—WARC, Web ARChive File Format."

22. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections—Websites and Pages—Quality and Functionality Factors," *Loc.gov*, accessed July 19, 2017, https://www.loc.gov/preservation/digital/formats/content/webarch_quality.shtml.

23. Brad Jones, "Internet Archive Will Ignore Robots.txt Files to Keep Historical Record Accurate," *DigitalTrends.com*, April 24, 2017, accessed July 17, 2017, https://www.digitaltrends.com/computing/internet-archive-robots-txt/.

24. Internet Archive, "About the Internet Archive," *Archive.org*, accessed July 17, 2017, https://archive.org/about/.

25. Molly Bragg, Kristine Hanna, Lori Donovan, Graham Hukill, and Anna Peterson, "The Web Archiving Life Cycle Model," *Archive.org*, March 2013, accessed July 19, 2017, http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf.

26. Digital Curation Centre, "DCC Curation Lifecycle Model," accessed July 19, 2017, http://www.dcc.ac.uk/resources/curation-lifecycle-model.

27. Niu, "An Overview of Web Archiving."

28. Library of Congress, "Web Archiving at the Library of Congress," *Loc.gov*, accessed July 19, 2017, https://www.loc.gov/webarchiving/technical.html.

29. International Internet Preservation Consortium, "OpenWayback," *Netpreserve.org*, accessed July 19, 2017, http://netpreserve.org/web-archiving/openwayback/.

30. Sara Day Thomson, "Preserving Social Media," DPC Technology Watch Report 16-01, February 2016, *DPCconline.org*, accessed July 20, 2017, http://www.dpconline.org/docman/technology-watch-reports/1486-twr16-01/file.

31. Thomson, "Preserving Social Media," 8.

32. Thomson, "Preserving Social Media," 20–21.

33. National Archives and Records Administration, "White Paper on Best Practices for the Capture of Social Media Records," *Archives.gov*, May 2013, accessed July 19, 2017, https://www.archives.gov/files/records-mgmt/resources/socialmediacapture.pdf.

34. Thomson, "Preserving Social Media," 18–19.

35. Scott McLemee, "The Archive Is Closed," *Insidehighered.com*, June 3, 2015, accessed July 19, 2017, https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research.

36. Digital Curation Centre, "Web Archiving," *DCC.ac.uk*, accessed July 17, 2017, http://www.dcc.ac.uk/resources/external/category/web-archiving.

37. Internet Archive, "Archive-It," accessed July 17, 2017, https://archive-it.org/.

38. Heidi Dowling, "Information for Donors," *Indiana University Digital Preservation Born Digital Preservation Lab*, last modified January 6, 2017, accessed July 21, 2017, https://wiki.dlib.indiana.edu/display/DIGIPRES/Information+for+Donors.

39. University of Wisconsin–Madison Libraries, "UW Archives and Records Management: Web and Born Digital Collections Policy and Procedures," *Wisc.edu*, accessed July 21, 2017, https://www.library.wisc.edu/archives/archives/our-collections-2/online-collections/web-and-born-digital-policy-and-procedures/.

40. Gabriela Redwine, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas, *Born Digital: Guidance for Donors, Dealers, and Archival Repositories*, Council on Library and Information Resources Publication 159 (Washington, DC: Council on Library and Information Resources, 2013), accessed March 5, 2017, https://www.clir.org/pubs/reports/pub159.

41. Creative Commons, "Share Your Work," *Creativecommons.org*, accessed July 21, 2017, https://creativecommons.org/share-your-work/.

42. Creative Commons, "Share Your Work."

CHAPTER 6

1. Alissa Matheny Helms and Jody L. DeRidder, "Workflow Guidelines for Digital Content Preservation: A Snapshot of Current Practice," *Journal of Digital Media Management* 6, no. 2 (Winter 2017/2018): 143.

2. Peter B. Hirtle, Emily Hudson, and Andrew T. Kenyon, *Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives, and Museums* (Ithaca, NY: Cornell University Library, 2009), accessed March 17, 2017, https://ecommons.cornell.edu/bitstream/handle/1813/14142/Hirtle-Copyright_final_RGB_lowres-cover1.pdf.

3. 105th Congress, United States of America, "Public Law 105-304: Digital Millennium Copyright Act (DMCA), Oct. 1998," *GPO.gov*, accessed May 14, 2017, https://www.gpo.gov/fdsys/pkg/PLAW-105publ304/pdf/PLAW-105publ304.pdf.

4. U.S. Copyright Office, "Copyright: Public Catalog," *Loc.gov*, accessed March 18, 2017, http://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First.

5. Marybeth Peters, "The 'Orphan Works' Problem and Proposed Legislation," *Copyright.gov*, March 13, 2008, accessed March 18, 2017, https://www.copyright.gov/docs/regstat031308.html.

6. European Union Intellectual Property Office Observatory, "Orphan Works Database," *Europa.eu*, accessed March 18, 2017, https://euipo.europa.eu/ohimportal/en/web/observatory/orphan-works-database.

7. Yael Lifshitz-Goldberg, "Orphan Works: Lecture Summary," World Intellectual Property Organization Seminar, May 2010, accessed March 18, 2017, http://www.wipo.int/edocs/mdocs/sme/en/wipo_smes_ge_10/wipo_smes_ge_10_ref_theme11_02.pdf.

8. Society of American Archivists, "Orphan Works: Statement of Best Practices," *Archivists.org*, last modified June 17, 2009, accessed March 18, 2017, http://www.archivists.org/standards/OWBP-V4.pdf.

9. Internet Archive, "Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy," *Archive.org*, last modified December 31, 2014, accessed March 18, 2017, https://archive.org/about/terms.php.

10. Brewster Kahle, "Providing Universal Access to Modern Materials—and Living to Tell the Tale," YouTube video, 1:04:53, from a presentation at the Coalition for Networked Information Spring Membership Meeting, April 13, 2015, posted by CNI, April 28, 2015, https://www.cni.org/events/membership-meetings/past-meetings/spring-2015/s15-plenary-sessions#opening.

11. Europeana and the Digital Public Library of America, "Rights Statements: About RightsStatements.org," *RightsStatements.org*, accessed March 16, 2017, http://rightsstatements.org/en/about.html.

12. Europeana and the Digital Public Library of America, "Rights Statements."

13. SAA Intellectual Property Working Group, Society of American Archivists, "Guide to Implementing Rights Statements from RightsStatements.org," *Archivists.org*, December 2, 2016, accessed April 22, 2017, http://www2.archivists.org/sites/all/files/RightsStatements_IPWG%20Guidance.pdf.

14. W3C, "ODRL Community Group," *W3.org*, accessed March 16, 2017, https://www.w3.org/community/odrl/.

15. Library of Congress, "Library of Congress Subject Headings," *Loc.gov*, accessed March 18, 2017, http://id.loc.gov/authorities/subjects.html.

16. Online Computer Library Center, "FAST (Faceted Application of Subject Terminology)," *OCLC.org*, accessed March 18, 2017, http://fast.oclc.org/.

17. Getty Research Institute, "Art and Architecture Thesaurus Online," *Getty.edu*, accessed March 18, 2017, http://www.getty.edu/research/tools/vocabularies/aat/.

18. Library of Congress, "Thesaurus for Graphic Materials," *Loc.gov*, accessed March 18, 2017, http://www.loc.gov/pictures/collection/tgm/.

19. Library of Congress, "Library of Congress Names," *Loc.gov*, accessed March 18, 2017, http://id.loc.gov/authorities/names.html.

20. Online Computer Library Center, "VIAF: The Virtual International Authority File," *VIAF.org*, accessed March 18, 2017, https://viaf.org/.

21. Indiana University, "Knowledge Base: Archived: What Is a URL?" *Kb.iu.edu*, last modified December 18, 2014, accessed March 18, 2017, https://kb.iu.edu/d/adnz.

22. Getty Research Institute, "Getty Thesaurus of Geographic Names Online," *Getty.edu*, accessed March 18, 2017, http://www.getty.edu/research/tools/vocabularies/tgn/index.html.

23. Open Geospatial Consortium, "Geography Markup Language," *Opengeospatial.org*, accessed March 18, 2017, http://www.opengeospatial.org/standards/gml.

24. DDI Alliance, "Document, Discover, and Interoperate," *DDIalliance.org*, accessed April 12, 2017, http://www.ddialliance.org/.

25. The Knowledge Network for Biocomplexity, "Ecological Metadata Language (EML)," *Ecoinformatics.org*, accessed April 12, 2017, https://knb.ecoinformatics.org/#external//emlparser/docs/index.html.

26. TDWG Biodiversity Information Standards, "Darwin Core," *TDWG.org*, last modified June 5, 2015, accessed April 12, 2017, http://rs.tdwg.org/dwc/.

27. Federal Geographic Data Committee, "Geospatial Metadata Standards and Guidelines," *FGDC.gov*, accessed April 12, 2017, https://www.fgdc.gov/metadata/geospatial-metadata-standards.

28. Bert Bos, W3C, "W3C Math Home: What Is MathML?" *W3.org*, last modified February 3, 2017, accessed April 12, 2017, https://www.w3.org/Math/.

29. Unicode, Inc., "General Information: What Is Unicode?" *Unicode.org*, last modified December 1, 2015, accessed April 12, 2017, http://www.unicode.org/standard/WhatIsUnicode.html.

30. Anonymous, "UTF-8 and Unicode," *UTF-8.com*, last modified August 28, 2014, accessed April 12, 2017, http://www.utf-8.com/.

31. Trusted Software Aps, "Opening CSV Files," *File.org*, accessed March 26, 2017, http://file.org/extension/csv.

32. Trusted Software Aps, "Opening TSV Files."

33. W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)," *W3.org*, last modified February 7, 2013, https://www.w3.org/TR/xml/#NT-Name.

34. W3C, "Extensible Markup Language."

35. W3Schools, "XML Tutorial," *W3schools.com*, accessed April 14, 2017, https://www.w3schools.com/xml/.

36. Library of Congress, "MODS Metadata Object Description Schema," *Loc.gov*, accessed March 18, 2017, http://www.loc.gov/standards/mods/.

37. FileFormat.Info, "UTF-8 Encoding," *Fileformat.info*, accessed April 14, 2017, http://www.fileformat.info/info/unicode/utf8.htm.

38. Library of Congress, "ISO 639.2: Codes for the Representation of Names of Languages," *Loc.gov*, last modified July 25, 2013, accessed April 14, 2017, https://www.loc.gov/standards/iso639-2/php/code_list.php.

39. W3C, "Schema," *W3.org*, accessed April 14, 2017, https://www.w3.org/standards/xml/schema.

40. W3C, "XSL Transformations (XSLT) Version 1.0," *W3.org*, last modified November 16, 1999, accessed April 14, 2017, https://www.w3.org/TR/xslt.

41. Anonymous, "ASCII Codes Table: Standard Characters," *ASCII.cl*, accessed March 26, 2017, http://ascii.cl/.

42. Acronis International GmbH, "Knowledge Base: 39790: Illegal Characters on Various Operating Systems," *Acronis.com*, accessed March 19, 2017, https://kb.acronis.com/content/39790.

43. Victor Laurie, "Naming Windows Files," *Vlaurie.com*, accessed March 19, 2017, http://vlaurie.com/computers2/Articles/filenames.htm.

44. U.S. Geographical Survey, "USGS Data Management: Persistent Identifiers," *USGS.gov*, last modified February 7, 2017, accessed March 19, 2017, https://www2.usgs.gov/datamanagement/preserve/persistentIDs.php.

45. International DOI Foundation, "DOI: The DOI System," *DOI.org*, accessed March 19, 2017, https://www.doi.org/.

46. Perry Willett and John Kunze, "ARK (Archival Resource Key) Identifiers," *UCOP.edu*, last modified October 24, 2016, accessed March 19, 2017, https://wiki.ucop.edu/display/Curation/ARK.

47. Internet Archive, "PURL Administration," *Archive.org*, accessed March 19, 2017, https://archive.org/services/purl/.

48. Corporation for National Research Initiatives, "Handle.Net Registry," *Handle.net*, last modified March 2, 2016, accessed March 19, 2017, http://www.handle.net/.

49. Leighton Walter Kille, "The Growing Problem of Internet 'Link Rot' and Best Practices for Media and Online Publishers," *Journalistsresource.org*, last modified October 9, 2015, accessed March 19, 2017, https://journalistsresource.org/studies/society/internet/website-linking-best-practices-media-online-publishers.

50. Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set, Version 1.1," *Dublincore.org*, last modified June 14, 2012, accessed March 19, 2017, http://dublincore.org/documents/dces/.

51. DCMI Usage Board, Dublin Core Metadata Initiative, "DCMI Metadata Terms," *Dublincore.org*, last modified June 14, 2012, accessed March 19, 2017, http://dublincore.org/documents/dcmiterms/.

52. Library of Congress, "Understanding MARC Bibliographic: Machine-Readable Cataloging," *Loc.gov*, last modified September 9, 2013, accessed March 19, 2017, http://www.loc.gov/marc/umb/.

53. W3C, "Linked Data," *W3.org*, accessed March 20, 2017, https://www.w3.org/standards/semanticweb/data.

54. W3C, "RDF/XML Syntax Specification (Revised)," *W3.org*, last modified February 10, 2004, accessed March 20, 2017, https://www.w3.org/TR/REC-rdf-syntax/.

55. ECMA International, "Introducing JSON," *JSON.org*, accessed March 20, 2017, http://www.json.org/.

56. Tom Heath, "Linked Data—Connect Distributed Data across the Web: Guides and Tutorials," *Linkeddata.org*, accessed March 20, 2017, http://linkeddata.org/guides-and-tutorials.

57. Tom Heath, "Linked Data—Connect Distributed Data Across the Web," *Linkeddata.org*, accessed April 12, 2017, http://linkeddata.org/faq.

58. W3C, "Date and Time Formats," *W3.org*, last modified September 15, 1997, accessed April 12, 2017, https://www.w3.org/TR/NOTE-datetime.

59. W3Schools, "XML Namespaces," *W3schools.com*, accessed March 26, 2017, https://www.w3schools.com/XML/xml_namespaces.asp.

60. Andy Powell and Pete Johnston, Dublin Core Metadata Initiative, "Guidelines for Implementing Dublin Core in XML: Example Mixing DC and ODRL Metadata," *Dublincore.org*, last modified April 4, 2003, accessed March 26, 2017, http://dublincore.org/documents/dc-xml-guidelines/#6.2.

61. Mark A. Green and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *American Archivist* 68 (Fall/Winter 2005): 208–63, accessed March 20, 2017, http://www.archivists.org/prof-education/pre-readings/IMPLP/AA68.2.MeissnerGreene.pdf.

62. Green and Meissner, "More Product, Less Process," 208.

63. Jody L. DeRidder and Alissa Matheny Helms, "Intake of Digital Content: Survey Results from the Field," *D-Lib Magazine* 22, no. 11/12 (November/December 2016), accessed March 20, 2017, doi:10.1045/november2016-deridder.

64. Jody L. DeRidder and Alissa Matheny Helms, "Digital Content Intake 20161102," in Jody L. DeRidder and Alissa Matheny Helms, *Incoming Digital Content Management*, Open Science Framework, February 14, 2017, accessed March 26, 2017, https://drive.google.com/file/d/0B7FbQWYX-ggJYnExMFlXOU5NeFk/view.

65. Library of Congress, "EAD: Encoded Archival Description," accessed March 20, 2017, https://www.loc.gov/ead/.

66. Andy Powell and Pete Johnston, Dublin Core Metadata Initiative, "Guidelines for Implementing Dublin Core in XML," *Dublincore.org*, last modified April 4, 2003, accessed March 26, 2017, http://dublincore.org/schemas/xmls/.

67. Library of Congress, "METS Metadata Encoding Transmission Standard," *Loc.gov*, last modified August 9, 2016, accessed March 26, 2017, http://www.loc.gov/standards/mets/.

68. Jeroen Bekaert, Patrick Hochstenbach, and Herbert Van de Sompel, "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine* 9, no. 11 (November 2003), accessed March 26, 2017, http://www.dlib.org/dlib/november03/bekaert/11bekaert.html.

69. Society of Motion Picture and Television Engineers, "ST 377-1:2011—SMPTE Standard—Material Exchange Format (MXF)—File Format Specification," June 7, 2011, accessed March 26, 2017, doi:10.5594/SMPTE.ST377-1.2011.

70. Library of Congress, "METS Implementation Registry," *Loc.gov*, last modified August 29, 2016, accessed March 26, 2017, http://www.loc.gov/standards/mets/mets-registry.html.

71. Jody L. DeRidder, "From Confusion and Chaos to Clarity and Hope," in *Digitization in the Real World: Lessons Learned from Small to Medium-Sized Digitization Projects*, ed. Kwong Bor Ng and Jason Kucsma, 333–54 (New York: Metropolitan New York Library Council, 2010), 340, 346, accessed March 26, 2017, http://metroblogs.typepad.com/files/ditrw_21.pdf.

72. University of Alabama Libraries Digital Services, "File Naming Schemes," *UA.edu*, last modified January 3, 2014, accessed March 26, 2017, http://www.lib.ua.edu/wiki/digcoll/index.php/File_naming_schemes.

73. John A. Kunze, Martin Haye, Erik Hetzner, Mark Reyes, and Cory Snavely, "Pairtrees for Collection Storage (V0.1)," *UCOP.edu*, last modified December 12, 2008, accessed March 26, 2017, https://confluence.ucop.edu/display/Curation/PairTree?preview=/14254128/16973838/PairtreeSpec.pdf.

74. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections: PDF/A-1a, PDF for Long-Term Preservation, Use of PDF 1.4, Level A Conformance," *Loc.gov*, last modified February 2, 2017, accessed March 26, 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000251.shtml.

75. John Edward Silva, "An Overview of Cryptographic Hash Functions and Their Uses," SANS Institute GIAC Security Essentials Practical Version 1.4b Option 1, *SANS.org*, January 15, 2003, accessed

March 26, 2017, https://www.sans.org/reading-room/whitepapers/vpns/overview-cryptographic-hash-functions-879.

76. David S. H. Rosenthal, "Keeping Bits Safe: How Hard Can It Be?" LOCKSS Program, Stanford University Library, 2010, *Lockss.org*, accessed March 27, 2017, https://lockss.org/locksswiki/files/ACM2010.pdf.

77. Wikipedia, "Comparison of File Verification Software," *Wikipedia*, last modified February 28, 2017, accessed March 27, 2017, https://en.wikipedia.org/wiki/Comparison_of_file_verification_software.

78. University of Alabama Libraries Digital Services, "Known FITS Errors," *UA.edu*, last modified December 2, 2014, accessed March 27, 2017, http://www.lib.ua.edu/wiki/digcoll/index.php/Known_FITS_errors.

79. JSTOR and the president and fellows of Harvard College, "JHOVE: JSTOR/Harvard Object Validation Environment," *SourceForge.net*, last modified February 25, 2009, accessed March 27, 2017, http://jhove.sourceforge.net/.

80. Gary McGath, "JHOVE Usage Notes," *Garymcgath.com*, last modified October 19, 2012, accessed March 27, 2017, http://www.garymcgath.com/jhovenote.html.

81. BitCurator Consortium, "BitCurator," *Bitcurator.net*, last modified March 18, 2017, accessed March 27, 2017, https://wiki.bitcurator.net/index.php?title=Main_Page.

82. Artefactual Systems, Inc., "Archivematica: Preserving Memory since 2009," accessed April 12, 2017, https://www.archivematica.org/en/.

83. National Archives, "Download DROID: File Format Identification Tool," *Nationalarchives.gov*, accessed April 12, 2017, https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

84. National Archives, "The Technical Registry PRONOM," *Nationalarchives.gov*, accessed April 12, 2017, https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

85. Ned Freed and Murray Kucherawy, Internet Assigned Numbers Authority, "Media Types," *IANA.org*, last modified April 11, 2017, accessed April 12, 2017, https://www.iana.org/assignments/media-types/media-types.xhtml.

86. JSTOR and the president and fellows of Harvard College, "JHOVE."

87. Library of Congress, "MIX: NISO Metadata for Images in XML Schema, Technical Metadata for Digital Still Images Standard," *Loc.gov*, last modified November 23, 2015, accessed April 12, 2017, https://www.loc.gov/standards/mix/.

88. Audio Engineering Society, "AES57-2011: AES Standard for Audio Metadata—Audio Object Structures for Preservation and Restoration," *AES.org*, accessed April 12, 2017, http://www.aes.org/publications/standards/search.cfm?docID=84.

89. Library of Congress, "TextMD Technical Metadata for Text," *Loc.gov*, last modified April 12, 2017, accessed April 12, 2017, https://www.loc.gov/standards/textMD/.

90. European Broadcasting Union, "Metadata Specifications: EBUCore," *Tech.ebu.ch*, accessed April 12, 2017, https://tech.ebu.ch/MetadataEbuCore.

91. Carol Chou and Andrea Goethals, "Document Metadata: Document Technical Metadata for Digital Preservation," *FCLA.edu*, last modified November 30, 2012, accessed April 12, 2017, https://share.fcla.edu/FDAPublic/DAITSS/documentMD.pdf.

92. Chou and Goethals, "Document Metadata."

93. Harvard University, "File Information Tool Set (FITS)," accessed April 12, 2017, https://projects.iq.harvard.edu/fits/home.

94. Open Preservation Foundation and PDF Association, "VeraPDF (PDF/A Validation)," *Openpreservation.org*, accessed April 12, 2017, http://openpreservation.org/about/projects/verapdf/.

95. Anonymous, "Software Preservation Network," *Softwarepreservationnetwork.org*, accessed April 12, 2017, http://www.softwarepreservationnetwork.org/.

96. International Press Telecommunications Council, "IPTC Video Metadata Hub—Recommendation 1.0/Properties," *IPTC.org*, last modified November 23, 2016, accessed April 12, 2017, http://www.iptc.org/std/videometadatahub/recommendation/IPTC-VideoMetadataHub-props-Rec_1.0.html.

97. Adobe Systems, Inc., "Acrobat for Legal Professionals: Using Save As to to [*sic*] Conform to PDF/A," *Adobe.com*, accessed April 14, 2017, http://blogs.adobe.com/acrolaw/2011/05/using-save-as-to-to-conform-to-pdfa/.

98. W3C, "Provenance Current Status," *W3.org*, accessed May 8, 2017, https://www.w3.org/standards/techs/provenance.

99. DataONE.org, "DataONE: Data Observation Network for Earth," *Dataone.org*, accessed May 8, 2017, https://www.dataone.org/.

100. DataONE.org, "DataONE: Data Observation Network for Earth, Cyberinfrastructure," *Dataone.org*, accessed May 8, 2017, https://www.dataone.org/working_groups/cyberinfrastructure.

101. DataONE Cyberinfrastructure Working Group, "ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance, Draft 01 May 2016," *Dataone.org*, accessed May 8, 2017, http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html.

102. Priscilla Caplan, "Library of Congress Network Development and MARC Standards Office: Understanding PREMIS," *Loc.gov*, February 1, 2009, 4, accessed April 16, 2017, http://www.loc.gov/standards/premis/understanding-premis.pdf.

103. PREMIS Editorial Committee, "Conformant Implementation of the PREMIS Data Dictionary," *Loc.gov*, 4, accessed April 16, 2017, http://www.loc.gov/standards/premis/premis-conformance-20150429.pdf.

104. PREMIS Editorial Committee and METS Editorial Board, "Guidelines for PREMIS with METS for Exchange," *Loc.gov*, last modified January 2017, accessed April 16, 2017, https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf.

105. PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata, Version 3.0," *Loc.gov*, last modified November 2015, 44, accessed April 16, 2017, http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

106. PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata," 14.

107. PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata," 14.

108. University of Michigan, "Deep Blue Preservation and Format Support Policy," *Umich.edu*, accessed April 16, 2017, https://deepblue.lib.umich.edu/static/about/deepbluepreservation.html.

109. Margaret Hedstrom and Christopher A. Lee, "Significant Properties of Digital Objects: Definitions, Applications, Implications," *Proceedings of the DLM-Forum 2002*, 218, accessed April 16, 2017, https://ils.unc.edu/callee/sigprops_dlm2002.pdf.

110. Wikipedia, "Microsoft Word," *Wikipedia*, last modified April 18, 2017, accessed April 16, 2017, https://en.wikipedia.org/wiki/Microsoft_Word.

111. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," *Loc.gov*, last modified March 10, 2017, accessed April 16, 2017, https://www.loc.gov/preservation/digital/formats/index.shtml.

112. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections. Sustainability Factors," *Loc.gov*, last modified March 10, 2017, accessed April 16, 2017, https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml.

113. DeRidder and Helms, "Intake of Digital Content."

114. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections. ESRI Shapefile," *Loc.gov*, last modified February 28, 2017, accessed April 16, 2017, https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml.

115. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections. Sustainability Factors."

116. Stanford Libraries, "ePADD," *Stanford.edu*, accessed April 19, 2017, https://library.stanford.edu/projects/epadd.

117. Community-Owned Digital Preservation Tool Registry, "Category: File Format Migration," *Digipres.org*, last modified October 29, 2014, accessed April 19, 2017, http://coptr.digipres.org/Category:File_Format_Migration.

118. Atlassian JIRA, "Conversion Software Registry," *NCSA.UIUC.edu*, accessed May 9, 2017, http://isda.ncsa.uiuc.edu/NARA/CSR/php/search/conversions.php.

119. Planets, "XCL—eXtensible Characterization Language," *Planetarium.hki.uni-koeln.de*, accessed April 19, 2017, http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl.html.

120. Jody DeRidder, "simpleFits.pl," last modified April 21, 2017, http://jodyderidder.com/scripts/simpleFits.txt.

121. Perl.org, "Perl—Download," *Perl.org*, accessed April 21, 2017, https://www.perl.org/get.html.

122. Harvard University, "File Information Tool Set (FITS)," *Harvard.edu*, accessed April 21, 2017, https://projects.iq.harvard.edu/fits/downloads.

123. Don Ho, "Notepad++," *Notepad-plus-plus.org*, accessed April 21, 2017, https://notepad-plus-plus.org/.

124. Bram Moolenaar, "Vim: The editor," *vim.org*, accessed April 21, 2017, http://www.vim.org/.

125. Perl.org, "The Perl Programming Language," *Perl.org*, accessed April 21, 2017, https://www.perl.org/.

126. Python Software Foundation, "Python," *Python.org*, accessed April 21, 2017, https://www.python.org/.

127. Steve Parker, "Shell Scripting Tutorial," *Shellscript.sh*, last modified December 20, 2016, accessed April 21, 2017, https://www.shellscript.sh/.

128. PHP Group, "What Is PHP?" *PHP.net*, accessed April 21, 2017, http://php.net/manual/en/intro-whatis.php.

129. Members of the Ruby Community, "Ruby: A Programmer's Best Friend," *Ruby-lang.org*, last modified April 1, 2017, accessed April 21, 2017, https://www.ruby-lang.org/en/.

130. W3C, "XSL Transformations (XSLT) Version 1.0."

131. University of Virginia Library Digital Production Group, "Born Digital Collections: An Inter-Institutional Model for Stewardship," *Virginia.edu*, accessed April 22, 2017, http://dcs.library.virginia.edu/aims/project-team-and-overview/.

132. AIMS Work Group, "AIMS Born Digital Collections: An Inter-Institutional Model for Stewardship," *Virginia.edu*, January 2012, accessed April 22, 2017, http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf.

133. BitCurator Consortium, "BitCurator: BitCurator Environment," *Bitcurator.net*, last modified March 16, 2017, accessed April 22, 2017, https://wiki.bitcurator.net/index.php?title=BitCurator_Environment.

134. BitCurator Consortium, "BitCurator: Screencast Tutorials," *Bitcurator.net*, last modified December 13, 2016, accessed April 22, 2017, http://wiki.bitcurator.net/index.php?title=Screencast_Tutorials.

135. BitCurator Consortium, "BitCurator Users," *Bitcurator.net*, accessed April 23, 2017, https://groups.google.com/forum/#!forum/bitcurator-users.

136. Society of American Archivists, "Digital Forensics for Archivists: Advanced," *Archivists.org*, accessed April 23, 2017, http://www2.archivists.org/prof-education/course-catalog/digital-forensics-for-archivists-advanced.

137. Artefactual Systems, "Archivematica."

CHAPTER 7

1. David S. H. Rosenthal, "Bit Preservation: A Solved Problem?" paper presented at *IPres 2008: The Fifth International Conference on Preservation of Digital Objects*, London, England, September 29–30, 2008, accessed May 12, 2017, https://www.bl.uk/ipres2008/presentations_day2/43_Rosenthal.pdf.

2. David S. H. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito, "Requirements for Digital Preservation Systems: A Bottom-Up Approach," *D-Lib Magazine* 11, no. 11 (November 2005), accessed May 18, 2017, http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html.

3. Library of Congress, "CD-ROM Longevity Research," *Loc.gov*, accessed May 19, 2017, https://www.loc.gov/preservation/scientists/projects/cd_longevity.html.

4. Wikipedia, "List of Microsoft Windows Versions," *Wikipedia*, accessed March 19, 2017, https://en.wikipedia.org/wiki/List_of_Microsoft_Windows_versions.

5. Wikipedia, "MacOS Version History," *Wikipedia*, accessed May 19, 2017, https://en.wikipedia.org/wiki/MacOS_version_history.

6. Research Libraries Group and Online Computer Library Center, "Trusted Digital Repositories: Attributes and Responsibilities," *OCLC.org*, May 2002, accessed May 21, 2017, http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

7. Center for Research Libraries, "Ten Principles," accessed May 21, 2017, http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re.

8. Online Computer Library Center and Center for Research Libraries, "Trustworthy Repositories Audit and Certification: Criteria and Checklist," Version 1.0, *CRL.edu*, February 2007, accessed May 21, 2017, http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

9. Consultative Committee for Space Data Systems, "Audit and Certification of Trusted Digital Repositories," Recommended Practice CCSDS 652.0-M-1, *CCSDS.org*, last modified September 2011, accessed May 21, 2017, https://public.ccsds.org/pubs/652x0m1.pdf.

10. International Organization for Standardization, "ISO 16363:2012, Space Data and Information Transfer Systems—Audit and Certification of Trusted Digital Repositories," *ISO.org*, last modified February 2012, accessed May 21, 2017, https://www.iso.org/standard/56510.html.

11. Center for Research Libraries, "Certification and Assessment of Digital Repositories," *CRL.edu*, accessed May 26, 2017, https://www.crl.edu/archiving-preservation/digital-archives/certification-assessment.

12. National Digital Stewardship Alliance, "Levels of Digital Preservation," *NDSA.org*, accessed May 28, 2017, http://ndsa.org/activities/levels-of-digital-preservation/.

13. Stanford University, "LOCKSS: Lots of Copies Keep Stuff Safe," *LOCKSS.org*, accessed May 26, 2017, https://www.lockss.org/.

14. Network of Alabama Academic Libraries, "The Alabama Digital Preservation Network: Preserving Alabama's Digital Resources," *ADPN.org*, accessed May 26, 2017, http://www.adpn.org/.

15. Educopia Institute, "MetaArchive Cooperative," *Metaarchive.org*, accessed May 26, 2017, http://www.metaarchive.org/.

16. Amazon Web Services, "Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region," *Amazon.com*, accessed May 26, 2017, https://aws.amazon.com/message/41926/.

17. Register, "AWS's S3 Outage Was So Bad Amazon Couldn't Get into Its Own Dashboard to Warn the World: Websites, Apps, Security Cams, IoT Gear Knackered," *Theregister.co.uk*, March 1, 2017, accessed May 26, 2017, https://www.theregister.co.uk/2017/03/01/aws_s3_outage/.

18. National Archives, "The Technical Registry PRONOM," *Nationalarchives.gov*, accessed March 16, 2017, https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

19. National Archives, "Download DROID: File Format Identification Tool," *Nationalarchives.gov*, accessed March 16, 2017, http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

20. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," last modified March 10, 2017, accessed April 16, 2017, https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml.

21. Atlassian JIRA, "Conversion Software Registry," *NCSA.UIUC.edu*, accessed May 9, 2017, http://isda.ncsa.uiuc.edu/NARA/CSR/php/search/conversions.php.

CHAPTER 8

1. Computer History Museum, "Timeline of Computer History," *Computerhistory.org*, accessed December 17, 2017, http://www.computerhistory.org/timeline/computers/.

2. Apple, Inc., “Apple Watch: Series 3,” *Apple.com*, accessed December 17, 2017, https://www.apple.com/watch/.

3. Wareable, “Alexa Gets in Your Face with LET Labs’ New Smartglasses,” *Wareable.com*, December 14, 2017, https://www.wareable.com/smartglasses/alexa-let-labs-glasses-291.

4. Will Greenwald, “The Best VR (Virtual Reality) Headsets of 2018,” *PCMAG.com*, December 5, 2017, https://www.pcmag.com/article/342537/the-best-virtual-reality-vr-headsets.

5. Apache, “Log Files,” *Apache.org*, accessed December 17, 2017, https://httpd.apache.org/docs/1.3/logs.html.

6. Google, “Google Analytics Solutions,” *Google.com*, accessed December 17, 2017, https://analytics.google.com/analytics/web/.

7. “Bounce rate” is the percentage of visitors who navigate away from a website after viewing only a single page.

8. U.S. Department of Health and Human Services, “Clinical Trials and Human Subject Protection,” *FDA.gov*, last modified September 25, 2017, accessed March 18, 2017, https://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/default.htm.

9. EUREC, “European Network of Research Ethics Committees,” *EURECnet.org*, accessed January 14, 2018, http://www.eurecnet.org/index.html.

10. U.S. Department of Health and Human Services Office for Human Research Protections, “International Compilation of Human Research Standards,” 2017 edition, *HHS.gov*, accessed January 14, 2017, https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part1.pdf and https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part2.pdf.

11. Tom Tullis and Bill Albert, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics* (Burlington, MA: Morgan Kauffmann, 2008), 124.

12. Tullis and Albert, *Measuring the User Experience*, 63–97.

13. International Organization for Standardization, “ISO/IEC CD 25020: Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Quality Measurement,” Edition 2, *ISO.org*, accessed January 21, 2018, https://www.iso.org/standard/72117.html.

14. Wikipedia, “Hawthorne Effect,” *Wikipedia*, accessed January 7, 2018, https://en.wikipedia.org/wiki/Hawthorne_effect.

15. Wikipedia, “Observer-Expectancy Effect,” *Wikipedia*, accessed January 8, 2018, https://en.wikipedia.org/wiki/Observer-expectancy_effect.

16. Nancy Fried Foster and Susan Gibbons, eds., *Studying Students: The Undergraduate Research Project at the University of Rochester* (Chicago: Association of College and Research Libraries 2007), accessed January 8, 2018, http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/Foster-Gibbons_cmpd.pdf.

17. Jody L. DeRidder and Kathryn G. Matheny, “What Do Researchers Need? Feedback on Use of Online Primary Source Materials,” *D-Lib Magazine* 20, no. 7/8 (July/August 2014), accessed January 8, 2018, doi:10.1045/july2014-deridder.

18. Wikipedia, “Grounded Theory,” Wikipedia, accessed January 8, 2018, https://en.wikipedia.org/wiki/Grounded_theory.

19. Wikipedia, “Observer-Expectancy Effect.”

20. Clive Nancarrow, Ian Brace, and Len Tiu, “‘Tell Me Lies, Tell Me Sweet Little Lies’: Dealing with Socially Desirable Responses in Market Research,” *Marketing Review* 2, no. 1 (Spring 2001): 55–69, accessed January 8, 2018, doi:10.1362/1469347012569427.

21. Don A. Dillman, Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L. Messer, “Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet,” *Social Science Research* 38, no. 1 (2009): 1–18, accessed January 8, 2018, doi:10.1016/j.ssresearch.2008.03.007.

22. Nancarrow, Brace, and Tiu, “‘Tell Me Lies, Tell Me Sweet Little Lies.’”

23. Yehuda Baruch and Brooks C. Holtom, “Survey Response Rate Levels and Trends in Organizational Research,” *Human Relations* 61, no, 8 (August 1, 2008): 1,139–60, accessed January 8, 2018, doi:10.1177/0018726708094863.

24. Online Computer Learning Center, “Dewey Decimal Classification Summaries: A Brief Introduction to the Dewey Decimal Classification System,” *OCLC.org*, accessed January 21, 2018, https://www.oclc.org/en/dewey/features/summaries.html.

25. Library of Congress, “Library of Congress Classification System Outline,” *Loc.gov*, accessed January 21, 2018, https://www.loc.gov/catdir/cpso/lcco/.

26. Google, Yahoo, and Microsoft, “What Are Sitemaps?” *Sitemaps.org*, accessed January 21, 2018, https://www.sitemaps.org/index.html.

27. W3Schools.com, “HTML <meta> Tag,” *W3Schools.com*, accessed January 21, 2018, https://www.w3schools.com/tags/tag_meta.asp.

28. Google, Microsoft, Yahoo, and Yandex, “Schema.org,” *Schema.org*, accessed January 21, 2018, http://schema.org/.

29. Society of American Archivists, “A Glossary of Archival and Records Terminology: Provenance,” *Archivists.org*, accessed January 21, 2018, https://www2.archivists.org/glossary/terms/p/provenance.

30. U.S. National Archives and Records Administration, “Social Networks and Archival Context Cooperative (SNAC),” *SNACcooperative.org*, accessed January 21, 2018, http://snaccooperative.org/.

31. Staatsbibliothek zu Berlin and Society of American Archivists, “EAC-CPF,” *EAC.staatsbibliothek-berlin.de*, accessed January 21, 2018, http://eac.staatsbibliothek-berlin.de/.

32. Library of Congress, “EAD: Encoded Archival Description,” *Loc.gov*, accessed January 21, 2018, http://www.loc.gov/ead/.

33. Google, Microsoft, Yahoo, and Yandex, “Schema.org.”

34. Noah Huffman, “The Tao of the DAO: Embedding Digital Objects in Finding Aids,” *Duke.edu*, July 5, 2015, accessed January 21, 2018, https://blogs.library.duke.edu/bitstreams/2015/07/06/the-tao-of-the-dao-embedding-digital-objects-in-finding-aids/.

35. Library of Congress, “EAD: Encoded Archival Description: EAD Elements. <dao> Digital Archival Object,” *Loc.gov*, accessed January 28, 2018, http://www.loc.gov/ead/tglib/elements/dao.html.

36. Duke University Libraries, “Guide to the Stephanie Strickland Papers, 1955–2016,” *Duke.edu*, accessed January 28, 2018, https://library.duke.edu/rubenstein/findingaids/stricklandstephanie/.

37. Bentley Historical Library, University of Michigan, “James J. Duderstadt Papers: 963–2016 (Bulk 1970–1996): University of Michigan Presidency 1986–1997,” *Umich.edu*, accessed January 28, 2018, https://quod.lib.umich.edu/b/bhlead/umich-bhl-9811?byte=88055554;focusrgn=C02;subview=standard;view=reslist.

38. Mark A. Green and Dennis Meissner, “More Product, Less Process: Revamping Traditional Archival Processing,” *American Archivist* 68 (Fall/Winter 2005): 208–63, accessed January 21, 2018, http://www.archivists.org/prof-education/pre- readings/IMPLP/AA68.2.MeissnerGreene.pdf.

39. Archive-It, “Explore All Archives: ‘Obama,’” *Archive-it.org*, accessed January28, 2018, https://archive-it.org/explore?q=Obama&show=Collections.

40. BitCurator, “BitCurator Access Webtools,” *Bitcurator.net*, accessed January 28, 2018, https://wiki.bitcurator.net/index.php?title=BitCurator_Access_Webtools.

41. BitCurator, “BitCurator Environment,” *Bitcurator.net*, accessed January 28, 2018, https://wiki.bitcurator.net/index.php?title=BitCurator_Environment.

42. University of Freiburg, “bwFLA: Emulation as a Service,” *Uni-freiburg.de*, accessed January 28, 2018, http://eaas.uni-freiburg.de/.

43. Kam Woods, Christopher A. Lee, Oleg Stobbe, Thomas Liebetraut, and Klaus Rechert, “Functional Access to Forensic Disk Images in a Web Service,” *Proceedings of the 12th International Conference on Digital Preservation*, Chapel Hill, North Carolina, November 2–6, 2015, 191–95, accessed January 28, 2018, handle: 11353/10.429627.

44. BitCurator, “BitCurator Access,” *Bitcurator.net*, accessed January 28, 2018, https://wiki.bitcurator.net/index.php?title=BitCurator_Access.

45. Jinfang Niu, “Functionalities of Web Archives,” *D-Lib Magazine* 18, no. 3/4 (March/April 2012), accessed January 8, 2018, doi:10.1045/march2012-niu2.

46. British Library, “UK Web Archive: Preserving UK Websites,” *Webarchive.org.uk*, accessed June 17, 2017, https://www.webarchive.org.uk/ukwa/.

47. Research Library of the Los Alamos National Laboratory and the Computer Science Department of Old Dominion University, "Memento Guide: Introduction to Memento," *Mementoweb.org*, last modified January 19, 2015, accessed June 17, 2017, http://www.mementoweb.org/guide/quick-intro/.

48. Wikipedia, "Comparison of Web Browsers: Image Formats," *Wikipedia*, accessed January 28, 2018, https://en.wikipedia.org/wiki/Comparison_of_web_browsers#Image_format_support.

49. Mozilla MDN Web Docs, "Media Formats for HTML Audio and Video," *Mozilla.org*, last modified November 15, 2017, accessed January 28, 2018, https://developer.mozilla.org/en-US/docs/Web/HTML/Supported_media_formats.

50. Jody L. DeRidder and Alissa Matheny Helms, "Intake of Digital Content: Survey Results from the Field," *D-Lib Magazine* 22:11/12 (November/December 2016), accessed January 28, 2018, DOI:10.1045/november2016-deridder.

CHAPTER 9

1. National Archives, "The Technical Registry PRONOM," *Nationalarchives.gov*, accessed October 8, 2017, https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

2. National Archives, "DROID (Digital Record and Object Identification)," *Digital-preservation.github.io*, accessed October 8, 2017, https://digital-preservation.github.io/droid/.

3. National Library of New Zealand, "Digital Preservation Programme: Digital Preservation Technical Registry," *Natlib.govt.nz*, accessed October 8, 2017, https://digitalpreservation.natlib.govt.nz/current-projects/technical-registry/.

4. VeraPDF Consortium, "VeraPDF," *Verapdf.org*, accessed September 17, 2017, http://verapdf.org/.

5. PDF Association, "PDF/A FAQ," *PDFA.org*, accessed September 17, 2017, https://www.pdfa.org/pdfa-faq/.

6. Open Preservation Foundation, "VeraPDF (PDF/A Validation)," *Openpreservation.org*, accessed September 17, 2017, http://openpreservation.org/about/projects/verapdf/.

7. Open Preservation Foundation, "JHOVE," *Openpreservation.org*, accessed September 17, 2017, http://openpreservation.org/technology/products/jhove/.

8. Open Preservation Foundation, "Planets (Preservation and Long-Term Access through NETworked Services)," *Openpreservation.org*, accessed September 17, 2017, http://openpreservation.org/about/projects/planets/.

9. Planets, "Components," *SourceForge.net*, accessed September 17, 2017, http://planets-suite.sourceforge.net/components/.

10. Planets, "Components."

11. Planets, "Planets Training Materials," *Planets-project.eu*, accessed September 17, 2017, http://planets-project.eu/training-materials/.

12. Planets, "Publications," *Planets-project.eu*, accessed September 17, 2017, http://planets-project.eu/publications/.

13. Open Preservation Foundation, "Planets (Preservation and Long-term Access through NETworked Services)."

14. Open Preservation Foundation, "Planets (Preservation and Long-term Access through NETworked Services)."

15. Open Preservation Foundation, "Software Supporters," *Openpreservation.org*, accessed September 17, 2017, http://openpreservation.org/about/join/software-supporters/.

16. Open Preservation Foundation, "Membership," *Openpreservation.org*, accessed September 17, 2017, http://openpreservation.org/about/join/membership/.

17. Open Preservation Foundation, "Membership."

18. Open Preservation Foundation, "Membership."

19. DigiPres Commons, "Community-Owned Digital Preservation Resources," *Digipres.org*, accessed September 17, 2017, http://www.digipres.org/.

20. DigiPres Commons, "Save Digital Stuff Right Now," *Digipres.org*, accessed September 17, 2017, http://www.digipres.org/#save-digital-stuff-right-now.

21. DigiPres Commons, "Digital Preservation Q&A," *Digipres.org*, accessed September 17, 2017, http://qanda.digipres.org/.

22. DigiPres Commons, "Community-Owned Digital Preservation Tool Registry (COPTR)," *Digipres.org*, accessed September 17, 2017, http://coptr.digipres.org/Main_Page.

23. DigiPres Commons, "POWRR Tool Grid," *Digipres.org*, accessed September 17, 2017, http://www.digipres.org/tools/.

24. DigiPres Commons, "Community-Owned Digital Preservation Tool Registry (COPTR)."

25. DigiPres Commons, "COPTR Needs YOU!" *Digipres.org*, accessed September 17, 2017, http://coptr.digipres.org/COPTR_needs_YOU!

26. Educopia Institute, "BitCurator Consortium," *Bitcuratorconsortium.org*, accessed September 17, 2017, https://bitcuratorconsortium.org/mission.

27. UNC School of Information and Library Science, "Bitcurator," *Bitcurator.net*, accessed September 17, 2017, https://www.bitcurator.net/bitcurator/.

28. BitCurator Consortium, "Why Join the BitCurator Consortium (BCC)?" *Bitcuratorconsortium.org*, accessed September 17, 2017, https://bitcuratorconsortium.org/join.

29. Stanford University, "LOCKSS: What Is LOCKSS?" *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/about/what-is-lockss/.

30. SourceForge, "LOCKSS (Lots of Copies Keep Stuff Safe)," *SourceForge.net*, accessed September 17, 2017, https://sourceforge.net/projects/lockss/.

31. LOCKSS, "How to Join," *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/join/.

32. LOCKSS, "Publishers and Titles (GLN)," *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/community/publishers-titles-gln/.

33. CLOCKSS, "The CLOCKSS Archive: A Trusted Community-Governed Archive," *CLOCKSS.org*, accessed September 17, 2017, https://www.clockss.org/clockss/Home.

34. Online Computer Learning Center and Center for Research Libraries, "Trustworthy Repositories Audit and Certification: Criteria and Checklist," Version 1.0, *CRL.edu*, February 2007, accessed September 17, 2017, http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

35. LOCKSS, "How LOCKSS Works," *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/about/how-it-works/.

36. CLOCKSS, "Contribute to CLOCKSS," *CLOCKSS.org*, accessed September 17, 2017, https://www.clockss.org/clockss/Contribute_to_CLOCKSS.

37. LOCKSS, "Global and Private LOCKSS Networks," *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/community/networks/.

38. LOCKSS, "Digital Federal Depository Library Program," *LOCKSS.org*, accessed September 17, 2017, https://www.lockss.org/community/networks/digital-federal-depository-library-program/.

39. PLNWIKI, "CGI Network," *LOCKSS.org*, accessed September 17, 2017, https://plnwiki.lockss.org/index.php?title=CGI_network.

40. COPPUL, "Council of Prairie and Pacific University Libraries," *COPPUL.ca*, accessed September 17, 2017, http://www.coppul.ca/.

41. Educopia Institute, "MetaArchive," *Metaarchive.org*, accessed September 17, 2017, https://metaarchive.org/.

42. Alabama Digital Preservation Network, "The Alabama Digital Preservation Network: Preserving Alabama's Digital Resources," *ADPN.org*, accessed September 17, 2017, http://www.adpn.org/.

43. DPN FAQ Google Groups, "How Long Does My Content Stay in DPN?" *Google.com*, accessed September 12, 2017, https://groups.google.com/forum/#!topic/dpn-faq/djm89-YJejg.

44. Digital Preservation Network, "Vision," *DPN.org*, accessed September 12, 2017, https://dpn.org/about.

45. Digital Preservation Network, "Services, Technical and Administrative: Frequently Asked Questions," *DPN.org*, June 2, 2015, accessed September 12, 2017, http://dpn.org/wp-content/uploads/2015/06/DPN_FAQ.6.2.15.pdf.

46. Open Science Framework, "Open Science Framework: A Scholarly Commons to Connect the Entire Research Cycle," *OSF.io*, accessed September 12, 2017, https://osf.io/.

47. Open Science Framework, "OSF Guides," *OSF.io*, accessed September 12, 2017, http://help.osf.io/m/faqs/l/726460-faqs.

48. Wikipedia, "List of Web Archiving Initiatives," *Wikipedia*, accessed October 8, 2017, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

49. Internet Memory Foundation, "Archivethe.Net (AtN): A Shared Web Archiving Platform Operated by Internet Memory," *Internetmemory.org*, accessed October 8, 2017, http://internetmemory.org/en/index.php/projects/atn.

50. Internet Archive, "Archive-It: About Us," *Archive-it.org*, accessed October 8, 2017, https://archive-it.org/learn-more/.

51. Archive-It, "Explore All Archives," *Archive-it.org*, accessed October 8, 2017, https://archive-it.org/explore.

52. Archive-It, "Archive-It Projects and Programs," *Archive-it.org*, accessed October 8, 2017, https://archive-it.org/blog/projects/.

53. Archive-It, "Community Calendar," *Archive-it.org*, accessed October 8, 2017, https://archive-it.org/blog/conferences/.

54. Andrea Goethals, Harvard Library, "Cobweb: Collaborative Collection Development for Web Archives," *Harvard.edu*, November 2, 2016, accessed October 8, 2017, http://library.harvard.edu/10312016-1311/cobweb-collaborative-collection-development-web-archives.323w2.

55. Kathryn Stine, "Community-Based, Collaborative, Collection-Development Cobweb: github.com/CobwebOrg/cobweb," presentation at the Archive-It Partner Meeting, Portland, Oregon, July 25, 2017, 9, accessed October 8, 2017, https://support.archive-it.org/hc/article_attachments/115013792503/2017-AITPM_04_Stine.pdf.

56. International Internet Preservation Consortium, "Join the IIPC," *Netpreserve.org*, accessed October 8, 2017, http://netpreserve.org/join-iipc/.

57. National Library of Australia and Partners, "PANDORA: Australia's Web Archive," *Pandora.nla.gov.au*, accessed September 10, 2017, http://pandora.nla.gov.au/.

58. Kyong-Ho Lee, Oliver Slattery, Richang Lu, Xiao Tang, and Victor McCrary, "The State of the Art and Practice of Digital Preservation," *Journal of Research of National Institute of Standards Technology* 107, no. 1 (January/February 2002): 93–106, accessed September 10, 2017, doi:10.6028/jres.107.010.

59. National Library of Australia and Partners, "PANDORA."

60. National Library of Australia and Partners, "PANDORA Digital Archiving System: PANDAS," *Pandora.nla.gov.au*, accessed September 10, 2017, http://pandora.nla.gov.au/pandas.html.

61. National Library of Australia, "Digital Preservation Policy, 4th Edition (2013)," *Nla.gov.au*, accessed September 10, 2017, http://www.nla.gov.au/policy-and-planning/digital-preservation-policy.

62. Software Preservation Network, "About," *Softwarepreservationnetwork.org*, accessed October 8, 2017, http://www.softwarepreservationnetwork.org/about/.

63. Software Preservation Network, "Join/Subscribe," *Softwarepreservationnetwork.org*, accessed October 8, 2017, http://www.softwarepreservationnetwork.org/contact/.

64. Software Preservation Network, "Working Groups," *Softwarepreservationnetwork.org*, accessed October 8, 2017, http://www.softwarepreservationnetwork.org/working-groups/.

65. Jessica Meyerson, Zach Vowell, Wendy Hagenmaier, Aliza Leventhal, Fernando Rios, Elizabeth Russey Roke, and Tim Walsh, "The Software Preservation Network (SPN): A Community Effort to Ensure Long-Term Access to Digital Cultural Heritage," *D-Lib Magazine* 23, no. 5/6 (May/June 2017), accessed October 8, 2017, https://doi.org/10.1045/may2017-meyerson.

66. PrestoCentre, "About Us," *Prestocentre.org*, accessed September 24, 2017, https://www.prestocentre.org/about-us.

67. PrestoCentre, "Sign Up for the Bimonthly PrestoCentre Newsletter," accessed September 24, 2017, *Prestocentre.org*, https://www.prestocentre.org/resources/newsletter.

68. PrestoCentre, “Newsletter Archive,” *Prestocentre.org*, accessed September 24, 2017, https://www.prestocentre.org/resources/newsletter-archive.

69. PrestoCentre, “Resources Archive,” *Prestocentre.org*, accessed September 24, 2017, https://www.prestocentre.org/resources.

70. PrestoCentre, “About Us.”

71. Data Preservation Alliance for the Social Sciences, “About Data-PASS,” *Data-pass.org*, accessed September 10, 2017, http://www.data-pass.org/about.jsp.

72. Data Preservation Alliance for the Social Sciences, “Joining the Partnership,” *Data-pass.org*, accessed September 10, 2017, http://www.data-pass.org/join.jsp.

73. Data Preservation Alliance for the Social Sciences, “Shared Catalog,” *Data-pass.org*, accessed September 10, 2017, http://www.data-pass.org/call.jsp.

74. Open Archives Initiative, “Open Archives Initiative Protocol for Metadata Harvesting,” *Openarchives.org*, accessed September 10, 2017, https://www.openarchives.org/pmh/.

75. Dataverse, “Harvard Dataverse,” *Harvard.edu*, accessed September 10, 2017, https://dataverse.harvard.edu/.

76. Dataverse, “Institutions,” *Harvard.edu*, accessed September 10, 2017, https://dataverse.org/institutions.

77. Data Preservation Alliance for the Social Sciences, “SafeArchive,” *Data-pass.org*, accessed September 10, 2017, http://www.data-pass.org/syndicated-storage.jsp.

78. SafeArchive, “Why SafeArchive?” *Safearchive.org*, accessed September 10, 2017, http://www.safearchive.org/.

79. Lee et al., “The State of the Art and Practice of Digital Preservation.”

80. InterPARES, “InterPARES Trust,” *Interparestrust.org*, accessed September 10, 2017, https://interparestrust.org/trust.

81. InterPARES, “The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project,” *Interpares.org*, accessed September 10, 2017, http://www.interpares.org/book/index.cfm.

82. InterPARES, “InterPARES 2 Project: Objectives,” *Interpares.org*, accessed September 10, 2017, http://www.interpares.org/ip2/ip2_objectives.cfm; InterPARES, “International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive, and Dynamic Records,” *Interpares.org*, accessed September 10, 2017, http://www.interpares.org/ip2/book.cfm.

83. InterPARES, “InterPARES 3 Project: Project Summary,” *Interpares.org*, accessed September 10, 2017, http://www.interpares.org/ip3/ip3_index.cfm.

84. InterPARES, “InterPARES 3 Project Sitemap,” *Interpares.org*, accessed September 10, 2017, http://www.interpares.org/sitemap.cfm?proj=ip3.

85. InterPARES, “InterPARES Trust.”

86. InterPARES Trust, “Dissemination,” *Interparestrust.org*, accessed September 10, 2017, https://interparestrust.org/trust/research_dissemination.

87. National Information Standards Organization, “Welcome to NISO,” *NISO.org*, accessed October 8, 2017, http://www.niso.org/home/.

88. International Organization for Standardization “International Organization for Standardization: Great Things Happen When the World Agrees,” *ISO.org*, accessed October 8, 2017, https://www.iso.org/home.html.

89. Audio Engineering Society, “AES57-2011 (r2017): AES Standard for Audio Metadata—Audio Object Structures for Preservation and Restoration,” *AES.org*, accessed October 8, 2017, http://www.aes.org/publications/standards/search.cfm?docID=84.

90. U.S. Department of Commerce, “NIST: National Institute of Standards and Technology,” *NIST.gov*, accessed October 8, 2017, https://www.nist.gov/.

91. Library of Congress, “Digital Preservation Outreach and Education: About DPOE,” accessed September 17, 2017, *Digitalpreservation.gov*, http://www.digitalpreservation.gov/education/index.html.

92. Library of Congress, "Digital Preservation Outreach and Education: DPOE Curriculum," *Digitalpreservation.gov*, accessed September 17, 2017, http://www.digitalpreservation.gov/education/curriculum.html.

93. Library of Congress, "Digital Preservation Outreach and Education: DPOE Train-the-Trainer Workshops," *Digitalpreservation.gov*, accessed September 17, 2017, http://www.digitalpreservation.gov/education/ttt.html.

94. Digital Curation Centre, "DCC: Because Good Research Needs Good Data," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/.

95. Digital Curation Centre, "Resources for Digital Curators," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/resources.

96. Digital Curation Centre, "Digital Curation Events," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/events.

97. Digital Curation Centre, "Digital Curation Training," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/training.

98. Digital Curation Centre, "Community of Digital Curators," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/community.

99. Digital Curation Centre, "DCC Curation Lifecycle Model," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/resources/curation-lifecycle-model.

100. Digital Curation Centre, "How-to Guides and Checklists," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/resources/how-guides.

101. Digital Curation Centre, "Curation Reference Manual," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/resources/curation-reference-manual.

102. Digital Curation Centre, "Developing RDM Services," *DCC.ac.uk*, accessed September 17, 2017, http://www.dcc.ac.uk/resources/developing-rdm-services.

103. Digital Curation Centre, "DMPonline," *DCC.ac.uk*, accessed September 17, 2017, https://dmponline.dcc.ac.uk/.

104. Digital Preservation Coalition, "Join Us," *DPConline.org*, accessed September 24, 2017, http://www.dpconline.org/about/join-us.

105. Digital Preservation Coalition, "Search Our Knowledge Base," *DPConline.org*, accessed September 24, 2017, http://www.dpconline.org/knowledge-base.

106. Digital Preservation Coalition, "Technology Watch Reports," *DPConline.org*, accessed September 24, 2017, http://www.dpconline.org/knowledge-base/tech-watch-reports.

107. Digital Preservation Coalition, "Digital Preservation Handbook," *DPConline.org*, accessed September 24, 2017, http://dpconline.org/handbook.

108. Digital Library Federation, "National Digital Stewardship Alliance: advancing the capacity to preserve our nation's digital resources for the benefit of present and future generations," accessed September 17, 2017, http://ndsa.org/.

109. Digital Library Federation, "About the NDSA," *NDSA.org*, accessed September 17, 2017, http://ndsa.org/about/.

110. Digital Library Federation, "Frequently Asked Questions," *NDSA.org*, accessed September 17, 2017, http://ndsa.org/faq/.

111. Council on Library and Information Resources, "National Digital Stewardship Alliance: Overview," *CLIR.org*, accessed September 17, 2017, http://coherence.clir.org/almanac/fact-sheet-gallery/national-digital-stewardship-alliance.

112. University of Edinburgh and Digital Curation Centre, "International Journal of Digital Curation," *IJDC.net*, accessed September 24, 2017, http://www.ijdc.net/index.php/ijdc/index.

113. Digital Curation Centre, "Curation Journals," *DCC.ac.uk*, accessed September 24, 2017, http://www.dcc.ac.uk/resources/curation-journals.

114. Corporation for National Research Initiatives, "*D-Lib Magazine*," *Dlib.org*, accessed September 24, 2017, http://www.dlib.org/back.html.

115. Digital Curation Centre, “International Digital Curation Conference,” *DCC.ac.uk*, accessed September 24, 2017, http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc.

116. Society for Imaging Science and Technology, “Archiving,” *Imaging.org*, accessed September 24, 2017, http://www.imaging.org/site/IST/Conferences/Archiving/IST/Conferences/Archiving/Archiving_Home.aspx.

117. Association for Information Science and Technology, “Annual Meetings,” *ASIST.org*, accessed September 24, 2017, https://www.asist.org/events/annual-meeting/.

118. Joint Conference on Digital Libraries, “Welcome to JCDL,” *JCDL.org*, accessed September 24, 2017, http://www.jcdl.org/.

119. University of Oregon, “Data Curation Events List and Calendar: U.S. and International Conferences and Workshops on Research Data Curation,” *UOregon.edu*, accessed September 24, 2017, https://datacure.uoregon.edu/.

120. Coalition for Networked Information, “Browse by Topic,” *CNI.org*, accessed September 24, 2017, https://www.cni.org/topics.

121. Google Groups, “Digital Curation,” *Google.com*, accessed September 24, 2017, https://groups.google.com/forum/#!forum/digital-curation.

122. Digital Curation Centre, “DCC Associates Network,” *DCC.ac.uk*, accessed September 24, 2017, http://www.dcc.ac.uk/community/dcc-associates.

123. JISC, “Digital-Preservation Home Page,” *JISCmail.ac.uk*, accessed September 24, 2017, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=DIGITAL-PRESERVATION.

124. American Library Association, “Digipres—Digital Preservation,” *ALA-org*, accessed September 24, 2017, http://lists.ala.org/sympa/info/digipres.

Bibliography

105th Congress, United States of America. "Public Law 105-304: Digital Millennium Copyright Act (DMCA), Oct. 1998." *GPO.gov*. Accessed May 14, 2017. https://www.gpo.gov/fdsys/pkg/PLAW-105publ304/pdf/PLAW-105publ304.pdf.

Acronis International GmbH. "Knowledge Base: 39790: Illegal Characters on Various Operating Systems." *Acronis.com*. Accessed March 19, 2017. https://kb.acronis.com/content/39790.

AIMS Work Group. "AIMS Born Digital Collections: An Inter-Institutional Model for Stewardship." *Virginia.edu*. January 2012. Accessed April 22, 2017. http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf.

Alabama Digital Preservation Network. "The Alabama Digital Preservation Network: Preserving Alabama's Digital Resources." *ADPN.org*. Accessed September 17, 2017. http://www.adpn.org/.

Amazon Web Services. "Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST–1) Region." *Amazon.com*. Accessed May 26, 2017. https://aws.amazon.com/message/41926/.

American Library Association, Canadian Library Association, and the Chartered Institute of Library and Information Professionals. "AACR: Welcome to the Homepage of the Anglo-American Cataloguing Rules." *AACR2.org*. Accessed April 28, 2017. http://www.aacr2.org/.

Anonymous. "ASCII Codes Table: Standard Characters." *ASCII.cl*. Accessed March 26, 2017. http://ascii.cl/.

———. "Software Preservation Network." *Softwarepreservationnetwork.org*. Accessed April 12, 2017. http://www.softwarepreservationnetwork.org/.

———. "UTF-8 and Unicode." *UTF-8.com*. Last modified August 28, 2014. Accessed April 12, 2017. http://www.utf–8.com/.

Apache. "Log Files." *Apache.org*. Accessed December 17, 2017. https://httpd.apache.org/docs/1.3/logs.html.

Artefactual Systems, Inc. "Archivematica: Preserving Memory since 2009." Archivematica.org. Accessed April 12, 2017. https://www.archivematica.org/en/.

Atlassian JIRA. "Conversion Software Registry." *NCSA.UIUC.edu*. Accessed May 9, 2017. http://isda.ncsa.uiuc.edu/NARA/CSR/php/search/conversions.php.

Audio Engineering Society. "AES57-2011: AES Standard for Audio Metadata—Audio Object Structures for Preservation and Restoration." *AES.org*. Accessed April 12, 2017. http://www.aes.org/publications/standards/search.cfm?docID=84.

Baruch, Yehuda, and Brooks C. Holtom. "Survey Response Rate Levels and Trends in Organizational Research." *Human Relations* 61, no. 8 (August 1, 2008): 1,139–60. Accessed January 8, 2018. doi:10.1177/0018726708094863.

Bekaert, Jeroen, Patrick Hochstenbach, and Herbert Van de Sompel. "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library." *D-Lib Magazine* 9, no. 11 (November 2003). Accessed March 26, 2017. http://www.dlib.org/dlib/november03/bekaert/11bekaert.html.

BitCurator Consortium. "BitCurator." *Bitcurator.net*. Last modified March 18, 2017. Accessed March 27, 2017. https://wiki.bitcurator.net/index.php?title=Main_Page.

Bos, Bert, W3C. "W3C Math Home: What Is MathML?" *W3.org*. Last modified February 3, 2017. Accessed April 12, 2017. https://www.w3.org/Math/.

bwFLA. "Emulation as a Service—Demo Page." *Demo.eaas.uni-freiburg.de*. Accessed August 8, 2017. http://demo.eaas.uni-freiburg.de/.

Caplan, Priscilla. "Library of Congress Network Development and MARC Standards Office: Understanding PREMIS." *Loc.gov*. February 1, 2009. Accessed April 16, 2017. http://www.loc.gov/standards/premis/understanding-premis.pdf.

Center for Research Libraries. "Ten Principles." Accessed May 21, 2017. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re.

Chou, Carol, and Andrea Goethals. "Document Metadata: Document Technical Metadata for Digital Preservation." *FCLA.edu*. Last modified November 30, 2012. Accessed April 12, 2017. https://share.fcla.edu/FDAPublic/DAITSS/documentMD.pdf.

CLOCKSS. "The CLOCKSS Archive: A Trusted Community-Governed Archive." *CLOCKSS.org*. Accessed September 17, 2017. https://www.clockss.org/clockss/Home.

Committee on Future Career Opportunities and Educational Requirements for Digital Curation, National Research Council of the National Academies. *Preparing the Workforce for Digital Curation*. Washington, DC: National Academies Press, 2015. Accessed April 16, 2017. https://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation.

Community-Owned Digital Preservation Tool Registry. "Category: File Format Migration." *Digipres.org*. Last modified October 29, 2014. Accessed April 19, 2017. http://coptr.digipres.org/Category:File_Format_Migration.

Consultative Committee for Space Data Systems. "Audit and Certification of Trusted Digital Repositories." Recommended Practice CCSDS 652.0-M-1. *CCSDS.org*. Last modified September 2011. Accessed May 21, 2017. https://public.ccsds.org/pubs/652x0m1.pdf.

———. "Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1 Blue Book." *Si.edu*. Last modified January 2002. Accessed April 28, 2017. https://siarchives.si.edu/sites/default/files/pdfs/650x0b1.PDF.

Cornell University Library, ICPSR, and MIT Libraries. "Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions. Timeline: Digital Technology and Preservation." Online tutorial. *Dpworkshop.org*. Last modified 2013. Accessed March 5, 2017. http://www.dpworkshop.org/dpm-eng/timeline/viewall.html.

Corporation for National Research Initiatives. "*D-Lib Magazine*." *Dlib.org*. Accessed September 24, 2017. http://www.dlib.org/back.html.

———. "Handle.Net Registry." *Handle.net*. Last modified March 2, 2016. Accessed March 19, 2017. http://www.handle.net/.

Council on Library and Information Resources. "National Digital Stewardship Alliance: Overview." *CLIR.org*. Accessed September 17, 2017. http://coherence.clir.org/almanac/fact-sheet-gallery/national-digital-stewardship-alliance.

Creative Commons. "About the Licenses: What the Licenses Do." *Creativecommons.org*. Accessed October 15, 2017. https://creativecommons.org/licenses/.

Cullen, Charles T., Peter B. Hirtle, David Levy, Clifford A. Lynch, and Jeff Rothenberg. *Authenticity in a Digital Environment*, Pub 92. Washington, DC: Council on Library and Information Resources, 2000. Accessed March 16, 2017. https://www.clir.org/pubs/reports/pub92.

Curtis, Jason. "Museum of Obsolete Media." *Obsoletemedia.org*. Accessed April 30, 2017. http://www.obsoletemedia.org.

Data Preservation Alliance for the Social Sciences. "About Data-PASS." *Data-pass.org*. Accessed September 24, 2017. http://www.data-pass.org/about.jsp.

DataONE Cyberinfrastructure Working Group. "ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance, Draft 01 May 2016." *Dataone.org*. Accessed May 8, 2017. http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html.

DataONE.org. "DataONE: Data Observation Network for Earth." Accessed May 8, 2017. https://www.dataone.org/.

Dataverse. "Harvard Dataverse." *Harvard.edu*. Accessed September 10, 2017. https://dataverse.harvard.edu/.

DCMI Usage Board, Dublin Core Metadata Initiative. "DCMI Metadata Terms." *Dublincore.org*. Last modified June 14, 2012. Accessed March 19, 2017. http://dublincore.org/documents/dcmi-terms/.

DDI Alliance. "Document, Discover, and Interoperate." *DDIalliance.org*. Accessed April 12, 2017. http://www.ddialliance.org/.

Delve, Janet, and David Anderson. "KEEP: Welcome to TOTEM—the Trustworthy Online Technical Environment Metadata Registry." Accessed August 5, 2017. http://www.keep-totem.co.uk/.

Delve, Janet, Leo Konstantelos, and Antonio Ciuffreda. "TOTEM: Trusted Online Technical Environment Metadata—A Long-Term Solution for a Relational Database/RDF Ontologies." Paper presented at the annual meeting of iPRES, Singapore, November 1–4, 2011. Accessed August 6, 2017. https://fedora.phaidra.univie.ac.at/fedora/objects/o:294265/methods/bdef:Content/get.

DeRidder, Jody L. "From Confusion and Chaos to Clarity and Hope." In *Digitization in the Real World: Lessons Learned from Small to Medium-Sized Digitization Projects*. Ed. Kwong Bor Ng and Jason Kucsma, 333–54. New York: Metropolitan New York Library Council, 2010. Accessed March 26, 2017. http://metroblogs.typepad.com/files/ditrw_21.pdf.

———. "An Introduction to Digital Preservation: A Three-Part Webinar Series, Based on the Library of Congress' Digital Preservation Outreach and Education (DPOE) Model." *ASERL.org*. Accessed May 9, 2017. http://www.aserl.org/archive/#INTRODUCTION.

DeRidder, Jody L., and Alissa Matheny Helms. "Digital Content Intake 20161102," in DeRidder, Jody, and Alissa M Helms. "Incoming Digital Content Management." Open Science Framework, February 14, 2017. osf.io/qprn4. Accessed March 26, 2017. https://drive.google.com/file/d/0B7FbQWYX-ggJYnExMFlXOU5NeFk/view.

———. "Intake of Digital Content: Survey Results from the Field." *D-Lib Magazine* 22, no. 11/12 (November/December 2016). Accessed March 20, 2017. doi:10.1045/november2016-deridder.

DeRidder, Jody L., and Kathryn G. Matheny. "What Do Researchers Need? Feedback on Use of Online Primary Source Materials." *D-Lib Magazine* 20, no. 7/8 (July/August 2014). Accessed January 8, 2018. doi:10.1045/july2014-deridder.

DigiPres Commons. "Community-Owned Digital Preservation Resources." *Digipres.org*. Accessed September 17, 2017. http://www.digipres.org/.

Digital Curation Centre. "DCC: Because Good Research Needs Good Data." *DCC.ac.uk*. Accessed September 17, 2017. http://www.dcc.ac.uk.

———. "DCC Curation Lifecycle Model." *DCC.ac.uk*. Accessed September 17, 2017. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

———. "Dioscuri." *DCC.ac.uk*. Accessed August 5, 2017, http://www.dcc.ac.uk/resources/external/Dioscuri.

———. "KEEP Emulation Framework." *DCC.ac.uk*. Accessed August 5, 2017. http://www.dcc.ac.uk/resources/external/keep-emulation-framework.

Digital Library Federation. "About the NDSA." *NDSA.org*. Accessed September 17, 2017. http://ndsa.org/about/.

———. "DLF Groups." *Diglib.org*. Accessed December 10, 2017. https://www.diglib.org/groups/.

Digital Preservation Coalition. "DPA 2012: DPC Award for Research and Innovation—Finalists." *DPConline.org*. Last modified August 12, 2016. Accessed September 24, 2017. http://www.dpconline.org/events/dpa–2012-dpc-award-for-research-and-innovation-finalists.

———. “DPC: Digital Preservation Coalition.” *DPConline.org*. Accessed September 24, 2017. http://dpconline.org/handbook.

Digital Preservation Network. “Preserving the Historical Record for This and Future Generations.” *DPN.org*. Accessed September 24, 2017. https://dpn.org/about.

Dublin Core Metadata Initiative. “Dublin Core Metadata Element Set, Version 1.1.” *Dublincore.org*. Last modified June 14, 2012. Accessed March 19, 2017. http://dublincore.org/documents/dces/.

ECMA International. “Introducing JSON.” *JSON.org*. Accessed March 20, 2017. http://www.json.org/.

Educopia Institute. “BitCurator Consortium.” *Bitcuratorconsortium.org*. Accessed September 17, 2017. https://bitcuratorconsortium.org/mission.

———. “MetaArchive Cooperative.” *Metaarchive.org*. Accessed May 26, 2017. http://www.metaarchive.org/.

EUREC. “European Network of Research Ethics Committees.” *EURECnet.org*. Accessed January 14, 2018. http://www.eurecnet.org/index.html.

European Broadcasting Union. “Metadata Specifications: EBUCore.” *Tech.ebu.ch*. Accessed April 12, 2017. https://tech.ebu.ch/MetadataEbuCore.

Europeana and the Digital Public Library of America. “Rights Statements.” *Rightsstatements.org*. Accessed March 16, 2017. http://rightsstatements.org/ page/1.0/?language=en.

Federal Geographic Data Committee. “Geospatial Metadata Standards and Guidelines.” *FGDC.gov*. Accessed April 12, 2017. https://www.fgdc.gov/metadata/geospatial-metadata-standards.

FileFormat.Info. “UTF-8 Encoding.” *Fileformat.info*. Accessed April 14, 2017. http://www.fileformat.info/info/unicode/utf8.htm.

Freed, Ned, and Murray Kucherawy, Internet Assigned Numbers Authority. “Media Types.” *IANA.org*. Last modified April 11, 2017. Accessed April 12, 2017. https://www.iana.org/assignments/media-types/media-types.xhtml.

Getty Research Institute. “Art and Architecture Thesaurus Online.” *Getty.edu*. Accessed March 18, 2017. http://www.getty.edu/research/tools/vocabularies/aat/.

———. “Getty Thesaurus of Geographic Names Online.” *Getty.edu*. Accessed March 18, 2017. http://www.getty.edu/research/tools/vocabularies/tgn/index.html.

Google. “Google Analytics Solutions.” *Google.com*. Accessed December 17, 2017. https://analytics.google.com/analytics/web/.

Google, Microsoft, Yahoo, and Yandex. “Schema.org.” *Schema.org*. Accessed January 21, 2018. http://schema.org/.

Google, Yahoo, and Microsoft. “What Are Sitemaps?” *Sitemaps.org*. Accessed January 21, 2018. https://www.sitemaps.org/index.html.

Green, Mark A., and Dennis Meissner. “More Product, Less Process: Revamping Traditional Archival Processing.” *American Archivist* 68 (Fall/Winter 2005): 208–63. Accessed March 20, 2017. http://www.archivists.org/prof-education/pre-readings/IMPLP/AA68.2.MeissnerGreene.pdf.

Greenwald, Will. “The Best VR (Virtual Reality) Headsets of 2018.” *PCMAG.com*. December 5, 2017. https://www.pcmag.com/article/342537/the-best-virtual-reality-vr-headsets.

Hagedorn, Kat. “OAIster: A ‘No Dead Ends’ Digital Object Service.” Library and Information Technology Association (LITA) National Forum, Norfolk, VA. October 3, 2003. Accessed March 16, 2017. http://archives.getty.edu:30008/o/oaister/pres/LITA03_Hagedorn.ppt.

Harvard University. “File Information Tool Set (FITS).” *Harvard.edu*. Accessed April 21, 2017. https://projects.iq.harvard.edu/fits/home.

Heath, Tom. “Linked Data—Connect Distributed Data across the Web.” Accessed March 20, 2017. http://linkeddata.org/.

Hedstrom, Margaret, and Christopher A. Lee. “Significant Properties of Digital Objects: Definitions, Applications, Implications.” *Proceedings of the DLM-Forum 2002*, 218–23. Accessed April 16, 2017. https://ils.unc.edu/callee/sigprops_dlm2002.pdf.

Helms, Alissa Matheny, and Jody L. DeRidder. “Workflow Guidelines for Digital Content Preservation: A Snapshot of Current Practice.” *Journal of Digital Media Management* 6, no. 2 (Winter 2017/2018): 140–52.

Henriksen, Sofie Laier, Wiel Seuskens, and Gaby Wijers. “Digitizing Contemporary Art: D6.1 Guidelines for a Long-Term Preservation Strategy for Digital Reproductions and Metadata.” Rev. 1.0. *DCA-project.eu*. February 13, 2012. Accessed August 7, 2017. http://www.dca-project.eu/images/uploads/varia/DCA_D61_Guidelines_Long_Term_Preservation_Strategy_20120213_V1.pdf.

Higgins, Sarah. “The DCC Curation Lifecycle Model.” *International Journal of Digital Curation* 1 (2008): 3, 134–40. Accessed March 6, 2017. doi:10.2218/ijdc.v3i1.48.

Hirtle, Peter B., Emily Hudson, and Andrew T. Kenyon. *Copyright and Cultural Institutions: Guidelines for Digitization for U. S. Libraries, Archives, and Museums*. Ithaca, NY: Cornell University Library, 2009. Accessed March 17, 2017. https://ecommons.cornell.edu/bitstream/handle/1813/14142/Hirtle-Copyright_final_RGB_lowres-cover1.pdf.

Huffman, Noah. “The Tao of the DAO: Embedding Digital Objects in Finding Aids.” *Duke.edu.* July 5, 2015. Accessed January 21, 2018. https://blogs.library.duke.edu/bitstreams/2015/07/06/the-tao-of-the-dao-embedding-digital-objects-in-finding-aids/.

Indiana University. “Media Digitization and Preservation Initiative.” *IU.edu*. Accessed October 15, 2017. https://mdpi.iu.edu/.

International DOI Foundation. “DOI: The DOI System.” *DOI.org*. Accessed March 19, 2017. https://www.doi.org/.

International Organization for Standardization. “ISO 16363:2012, Space Data and Information Transfer Systems—Audit and Certification of Trusted Digital Repositories.” *ISO.org*. Last modified February 2012. Accessed May 21, 2017. https://www.iso.org/standard/56510.html.

International Press Telecommunications Council. “IPTC Video Metadata Hub—Recommendation 1.0/Properties.” *IPTC.org*. Last modified November 23, 2016. Accessed April 12, 2017. http://www.iptc.org/std/videometadatahub/recommendation/IPTC-VideoMetadataHub-props-Rec_1.0.html.

Internet Archive. “Archive-It: About Us.” *Archive-it.org*. Accessed October 8, 2017. https://archive-it.org/learn-more/.

———. “Internet Archive’s Terms of Use, Privacy Policy, and Copyright Policy.” *Archive.org*. Last modified December 31, 2014. Accessed March 18, 2017. https://archive.org/about/terms.php.

———. “PURL Administration.” *Archive.org*. Accessed March 19, 2017. https://archive.org/services/purl/.

Internet Memory Foundation. “Archivethe.Net (AtN): A Shared Web Archiving Platform Operated by Internet Memory.” *Internetmemory.org*. Accessed October 8, 2017. http://internetmemory.org/en/index.php/projects/atn.

InterPARES. “InterPARES Trust.” *Interparestrust.org*. Accessed September 10, 2017. https://interparestrust.org/trust.

JSTOR and the president and fellows of Harvard College. “JHOVE: JSTOR/Harvard Object Validation Environment.” *SourceForge.net*. Last modified February 25, 2009. Accessed March 27, 2017. http://jhove.sourceforge.net/.

Kahle, Brewster. “Providing Universal Access to Modern Materials—and Living to Tell the Tale.” YouTube video, 1:04:53, from a presentation at the Coalition for Networked Information Spring Membership Meeting, April 13, 2015. Posted by CNI, April 28, 2015. https://www.cni.org/events/membership-meetings/past-meetings/spring–2015/s15-plenary-sessions#opening.

Knight, Gareth. “Same as It Ever Was? Significant Properties and the Preservation of Meaning over Time.” Paper presented at *Decoding the Digital: A Common Language for Preservation*, London, July 27, 2010. Accessed August 7, 2017. http://www.dpconline.org/docs/miscellaneous/events/486-decodingknight-pdf/file.

Knowledge Network for Biocomplexity. “Ecological Metadata Language (EML).” *Ecoinformatics.org*. Accessed April 12, 2017. https://knb.ecoinformatics.org/#external//emlparser/docs/index.html.

Koninklijke Bibliotheek, National Archief, Planets, and KEEP. “Dioscuri—the Modular Emulator.” *SourceForge.net*. Accessed August 5, 2017. http://dioscuri.SourceForge.net/.

Kunze, John A., Martin Haye, Erik Hetzner, Mark Reyes, and Cory Snavely. “Pairtrees for Collection Storage (V0.1).” *UCOP.edu*. Last modified December 12, 2008. Accessed March 26, 2017. https://confluence.ucop.edu/display/Curation/PairTree?preview=/14254128/16973838/PairtreeSpec.pdf.

Laurie, Victor. “Naming Windows Files.” *Vlaurie.com*. Accessed March 19, 2017. http://vlaurie.com/computers2/Articles/filenames.htm.

Lee, Kyong-Ho, Oliver Slattery, Richang Lu, Xiao Tang, and Victor McCrary. “The State of the Art and Practice of Digital Preservation.” *Journal of Research of National Institute of Standards Technology* 107, no. 1 (January/February 2002): 93–106. Accessed May 19, 2017. doi:10.6028/jres.107.010.

Library of Congress. “CD-ROM Longevity Research.” *Loc.gov*. Accessed May 19, 2017. https://www.loc.gov/preservation/scientists/projects/cd_longevity.html.

———. “Digital Preservation Outreach and Education.” *Loc.gov*. Accessed May 7, 2017. http://www.digitalpreservation.gov/education/.

———. “EAD: Encoded Archival Description.” *Loc.gov*. Accessed January 21, 2018. http://www.loc.gov/ead/.

———. “ISO 639.2: Codes for the Representation of Names of Languages.” *Loc.gov*. Last modified July 25, 2013. Accessed April 14, 2017. https://www.loc.gov/standards/iso639–2/php/code_list.php.

———. “Library of Congress Names.” *Loc.gov*. Accessed March 18, 2017. http://id.loc.gov/authorities/names.html.

———. “Library of Congress Subject Headings.” *Loc.gov*. Accessed March 18, 2017. http://id.loc.gov/authorities/subjects.html.

———. “METS Implementation Registry.” *Loc.gov*. Last modified August 29, 2016. Accessed March 26, 2017. http://www.loc.gov/standards/mets/mets-registry.html.

———. “METS Metadata Encoding Transmission Standard.” *Loc.gov*. Last modified August 9, 2016. Accessed March 26, 2017. http://www.loc.gov/standards/mets/.

———. “MIX: NISO Metadata for Images in XML Schema, Technical Metadata for Digital Still Images Standard.” *Loc.gov*. Last modified November 23, 2015. Accessed April 12, 2017. https://www.loc.gov/standards/mix/.

———. “MODS Metadata Object Description Schema.” *Loc.gov*. Accessed March 18, 2017. http://www.loc.gov/standards/mods/.

———. “Personal Archiving: Preserving Your Digital Memories.” *Loc.gov*. Accessed October 15, 2017. http://www.digitalpreservation.gov//personalarchiving/records.html.

———. “Sustainability of Digital Formats: Planning for Library of Congress Collections.” *Loc.gov*. Last modified March 10, 2017. Accessed April 16, 2017. https://www.loc.gov/preservation/digital/formats/index.html.

———. “TextMD Technical Metadata for Text.” *Loc.gov*. Last modified April 12, 2017. Accessed April 12, 2017. https://www.loc.gov/standards/textMD/.

———. “Thesaurus for Graphic Materials.” *Loc.gov*. Accessed March 18, 2017. http://www.loc.gov/pictures/collection/tgm/.

———. “Understanding MARC Bibliographic: Machine-Readable Cataloging.” *Loc.gov*. Last modified September 9, 2013. Accessed March 19, 2017. http://www.loc.gov/marc/umb/.

Lifshitz-Goldberg, Yael. “Orphan Works: Lecture Summary.” World Intellectual Property Organization Seminar. May 2010. Accessed March 18, 2017. http://www.wipo.int/edocs/mdocs/sme/en/wipo_smes_ge_10/wipo_smes_ge_10_ref_theme11_02.pdf.

Lohman, Bram, Jeffrey van der Hoeven and Edo Noordermeer. “KEEP Emulation Framework System User Guide Release 2.0.0 (February 2012).” Accessed July 4, 2017. http://emuframework.sourceforge.net/docs/System-User-Guide_2.0.pdf.

Meijer, J. G. “Librarianship: A Definition.” University of Illinois Graduate School of Library and Information Science Occasional Papers 155, September 1982, 26. Accessed March 6, 2017. http://hdl.handle.net/2142/3979.

Meyerson, Jessica, Zach Vowell, Wendy Hagenmaier, Aliza Leventhal, Fernando Rios, Elizabeth Russey Roke, and Tim Walsh. “The Software Preservation Network (SPN): A Community Effort to Ensure Long-Term Access to Digital Cultural Heritage.” *D-Lib Magazine* 23, no. 5/6 (May/June 2017). Accessed January 28, 2018. https://doi.org/10.1045/may2017-meyerson.

Mozilla MDN Web Docs. "Media Formats for HTML Audio and Video." *Mozilla.org*. Last modified November 15, 2017. Accessed January 28, 2018. https://developer.mozilla.org/en-US/docs/Web/HTML/Supported_media_formats.

Nancarrow, Clive, Ian Brace, and Len Tiu. "'Tell Me Lies, Tell Me Sweet Little Lies': Dealing with Socially Desirable Responses in Market Research." *Marketing Review* 2, no. 1 (Spring 2001): 55–69. Accessed January 8, 2018. doi:10.1362/1469347012569427.

National Archives. "Download DROID: File Format Identification Tool." *Nationalarchives.gov*. Accessed April 12, 2017. https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.

———. "The Technical Registry PRONOM." *Nationalarchives.gov*. Accessed April 12, 2017. https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

National Archives and Records Administration. "Strategic Directions: Appraisal Policy." September 2007. Accessed June 1, 2017. https://www.archives.gov/records-mgmt/initiatives/appraisal.html.

National Digital Stewardship Alliance. "Levels of Digital Preservation." *NDSA.org*. Accessed May 28, 2017. http://ndsa.org/activities/levels-of-digital-preservation/.

National Library of Australia. "Digital Preservation Policy, 4th Edition (2013)." *Nla.gov.au*. Accessed September 10, 2017. http://www.nla.gov.au/policy-and-planning/digital-preservation-policy.

National Library of Australia and Partners. "PANDORA: Australia's Web Archive." *Pandora.nla.gov.au*. Accessed September 10, 2017. http://pandora.nla.gov.au/.

National Library of New Zealand. "Digital Preservation Programme: Digital Preservation Technical Registry." *Natlib.govt.nz*. Accessed October 8, 2017. https://digitalpreservation.natlib.govt.nz/current-projects/technical-registry/.

Network of Alabama Academic Libraries. "The Alabama Digital Preservation Network: Preserving Alabama's Digital Resources." *ADPN.org*. Accessed May 26, 2017. http://www.adpn.org/.

Niu, Jinfang. "Functionalities of Web Archives." *D-Lib Magazine* 18, no. 3/4 (March/April 2012). Accessed May 26, 2017. doi:10.1045/march2012-niu2.

Olive Archive. "Olive Executable Archive." *Olivearchive.org*. Accessed August 5, 2017. https://olivearchive.org/.

Online Computer Library Center. "FAST (Faceted Application of Subject Terminology)." *OCLC.org*. Accessed March 18, 2017. http://fast.oclc.org/.

———. "VIAF: The Virtual International Authority File." *VIAF.org*. Accessed March 18, 2017. https://viaf.org/.

Online Computer Library Center and Center for Research Libraries. "Trustworthy Repositories Audit and Certification: Criteria and Checklist." Version 1.0. *CRL.edu*. February 2007. Accessed May 21, 2017. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

Open Archives Initiative. "Open Archives Initiative Protocol for Metadata Harvesting." *Openarchives.org*. Accessed September 10, 2017. https://www.openarchives.org/pmh/.

Open Geospatial Consortium. "Geography Markup Language." *Opengeospatial.org*. Accessed March 18, 2017. http://www.opengeospatial.org/standards/gml.

Open Preservation Foundation. "The Open Preservation Foundation." *Openpreservation.org*. Accessed April 12, 2017. http://openpreservation.org.

Open Science Framework. "Open Science Framework: A Scholarly Commons to Connect the Entire Research Cycle." *OSF.io*. Accessed September 12, 2017. https://osf.io/.

Perl.org. "Perl—Download." *Perl.org*. Accessed April 21, 2017. https://www.perl.org/get.html.

Peters, Marybeth. "The 'Orphan Works' Problem and Proposed Legislation." *Copyright.gov*. March 13, 2008. Accessed March 18, 2017. https://www.copyright.gov/docs/regstat031308.html.

Planets. "XCL—eXtensible Characterization Language." *Planetarium.hki.uni-koeln.de*. Accessed April 19, 2017. http://planetarium.hki.uni-koeln.de/planets_cms/about-xcl.html.

PLNWIKI. "CGI Network." *LOCKSS.org*. Accessed September 17, 2017. https://plnwiki.lockss.org/index.php?title=CGI_network.

Powell, Andy, and Pete Johnston, Dublin Core Metadata Initiative. “Guidelines for Implementing Dublin Core in XML.” *Dublincore.org*. Last modified April 4, 2003. Accessed March 26, 2017. http://dublincore.org/schemas/xmls/.

PREMIS Editorial Committee. “Conformant Implementation of the PREMIS Data Dictionary.” *Loc.gov*. Accessed April 16, 2017. http://www.loc.gov/standards/premis/premis-conformance–20150429.pdf.

———. “PREMIS Data Dictionary for Preservation Metadata, Version 3.0.” *Loc.gov*. Last modified November 2015. Accessed April 16, 2017. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

PREMIS Editorial Committee and METS Editorial Board. “Guidelines for PREMIS with METS for Exchange.” *Loc.gov*. Last modified January 2017. Accessed April 16, 2017. https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf.

Preservica. “Preservica Digital Preservation.” *Preservica.com*. Accessed April 12, 2017. http://preservica.com/.

RDA Steering Committee. “About RDA.” *Rda-rsc.org*. Last modified April 14, 2017. Accessed March 5, 2017. http://www.rda-rsc.org/content/about-rda.

Redwine, Gabriela, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas. *Born Digital: Guidance for Donors, Dealers, and Archival Repositories.* Council on Library and Information Resources Publication 159. Washington, DC: Council on Library and Information Resources, 2013. Accessed March 5, 2017. https://www.clir.org/pubs/reports/pub159.

Register. “AWS’s S3 Outage Was So Bad Amazon Couldn’t Get into Its Own Dashboard to Warn the World: Websites, Apps, Security Cams, IoT Gear Knackered.” *Theregister.co.uk*. March 1, 2017. Accessed May 26, 2017. https://www.theregister.co.uk/2017/03/01/aws_s3_outage/.

Research Libraries Group and Online Computer Library Center. “Trusted Digital Repositories: Attributes and Responsibilities.” *OCLC.org*. May 2002. Accessed May 21, 2017. http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

Rosenthal, David S. H. “Bit Preservation: A Solved Problem?” Paper presented at *IPres 2008: The Fifth International Conference on Preservation of Digital Objects*, London, England, September 29–30, 2008. Accessed May 12, 2017. https://www.bl.uk/ipres2008/presentations_day2/43_Rosenthal.pdf.

———. “Emulation and Virtualization as Preservation Strategies.” Andrew Mellon Foundation, 2015. *Mellon.org*. Accessed August 5, 2017. https://mellon.org/media/filer_public/0c/3e/0c3eee7d–4166–4ba6-a767–6b42e6a1c2a7/rosenthal-emulation-2015.pdf.

Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito. “Requirements for Digital Preservation Systems: A Bottom-Up Approach.” *D-Lib Magazine* 11, no. 11 (November 2005). Accessed May 18, 2017. http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html.

SAA Intellectual Property Working Group, Society of American Archivists. “Guide to Implementing Rights Statements from RightsStatements.org.” *Archivists.org*. December 2, 2016. Accessed April 22, 2017. http://www2.archivists.org/sites/all/files/RightsStatements_IPWG%20Guidance.pdf.

SafeArchive. “Why SafeArchive?” *Safearchive.org*. Accessed September 10, 2017. http://www.safearchive.org/.

SAS Institute. “Byte Ordering on Big Endian and Little Endian Platforms.” *V8doc.sas.com*. Accessed August 7, 2017. https://v8doc.sas.com/sashtml/lgref/z1270373.htm.

Society of American Archivists. “Describing Archives: A Content Standard (DACS).” *Archivists.org*. Accessed April 28, 2017. http://www.archivists.org/governance/standards/dacs.asp.

———. “A Glossary of Archival and Records Terminology: Provenance.” *Archivists.org*. Accessed January 21, 2018. https://www2.archivists.org/glossary/terms/p/provenance.

———. “Orphan Works: Statement of Best Practices.” *Archivists.org*. Last modified June 17, 2009. Accessed March 18, 2017. http://www.archivists.org/standards/OWBP-V4.pdf.

Society of Motion Picture and Television Engineers. "ST 377-1:2011—SMPTE Standard—Material Exchange Format (MXF)—File Format Specification." June 7, 2011. Accessed March 26, 2017. doi:10.5594/SMPTE.ST377–1.2011.

Software Preservation Network. "Software Preservation Network." *Softwarepreservationnetwork.org*. Accessed October 8, 2017. http://www.softwarepreservationnetwork.org/about/.

SourceForge. "KEEP Emulation Framework (EF)." *SourceForge.net*. Accessed August 5, 2017. http://emuframework.SourceForge.net/.

Staatsbibliothek zu Berlin and Society of American Archivists. "EAC-CPF." *EAC.staatsbibliothek-berlin.de*. Accessed January 21, 2018. http://eac.staatsbibliothek-berlin.de/.

Stanford Libraries. "ePADD." *Stanford.edu*. Accessed April 19, 2017. https://library.stanford.edu/projects/epadd.

Stanford University. "LOCKSS: Lots of Copies Keep Stuff Safe." *LOCKSS.org*. Accessed May 26, 2017. https://www.lockss.org/.

Stine, Kathryn. "Community-Based, Collaborative, Collection-Development Cobweb: github.com/CobwebOrg/cobweb." Presentation at the Archive-It Partner Meeting, Portland, Oregon, July 25, 2017. Accessed October 8, 2017. https://support.archive-it.org/hc/article_attachments/115013792503/2017-AITPM_04_Stine.pdf.

TDWG Biodiversity Information Standards. "Darwin Core." *TDWG.org*. Last modified June 5, 2015. Accessed April 12, 2017. http://rs.tdwg.org/dwc/.

Tullis, Tom, and Bill Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Burlington, MA: Morgan Kauffmann, 2008.

Unicode, Inc. "General Information: What Is Unicode?" *Unicode.org*. Last modified December 1, 2015. Accessed April 12, 2017. http://www.unicode.org/standard/WhatIsUnicode.html.

University of Edinburgh and Digital Curation Centre. "International Journal of Digital Curation." *IJDC.net*. Accessed September 24, 2017. http://www.ijdc.net/index.php/ijdc/index.

University of Freiburg. "bwFLA: Emulation as a Service." *Uni-freiburg.de*. Accessed January 28, 2018. http://eaas.uni-freiburg.de/.

University of Michigan. "Deep Blue Preservation and Format Support Policy." *Umich.edu*. Accessed April 16, 2017. https://deepblue.lib.umich.edu/static/about/deepbluepreservation.html.

University of North Carolina School of Information and Library Science. "BitCurator: About the Project." *Bitcurator.net*. Accessed June 1, 2017. https://www.bitcurator.net/bitcurator/.

U.S. Copyright Office. "Copyright: Public Catalog." *Loc.gov*. Accessed March 18, 2017. http://cocatalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First.

U.S. Department of Health and Human Services. "Clinical Trials and Human Subject Protection." *FDA.gov*. Last modified September 25, 2017. Accessed March 18, 2017. https://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/default.htm.

U.S. Department of Health and Human Services Office for Human Research Protections. "International Compilation of Human Research Standards," 2017 edition. *HHS.gov*. Accessed January 14, 2017. https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part1.pdf and https://www.hhs.gov/ohrp/sites/default/files/internationalcomp2017-part2.pdf.

U.S. Geographical Survey. "USGS Data Management: Persistent Identifiers." *USGS.gov*. Last modified February 7, 2017. Accessed March 18, 2017. https://www2.usgs.gov/datamanagement/preserve/persistentIDs.php.

U.S. National Archives and Records Administration. "Social Networks and Archival Context Cooperative (SNAC)." *SNACcooperative.org*. Accessed January 21, 2018. http://snaccooperative.org/.

van der Hoeven, Jeffrey. "Dioscuri: Emulation for Digital Preservation." Paper presented at the annual meeting of Planets, CASPAR, and DPE, Lisbon, Portugal, September 5–6, 2007. Accessed August 4, 2017. http://www.planets-project.eu/docs/presentations/2007–09–05_emulation_wepreserve_portugal_jrvanderhoeven.ppt.

W3C. "Date and Time Formats." *W3.org*. Last modified September 15, 1997. Accessed April 3, 2017. https://www.w3.org/TR/NOTE-datetime.

———. “Extensible Markup Language (XML) 1.0 (Fifth Edition).” *W3.org*. Last modified February 7, 2013. Accessed August 6, 2017. https://www.w3.org/TR/xml/#NT-Name.

———. “Linked Data.” *W3.org*. Accessed March 20, 2017. https://www.w3.org/standards/semanticweb/data.

———. “ODRL Community Group.” *W3.org*. Accessed March 16, 2017. https://www.w3.org/community/odrl/.

———. “Provenance Current Status.” *W3.org*. Accessed May 8, 2017. https://www.w3.org/standards/techs/provenance.

———. “RDF/XML Syntax Specification (Revised).” *W3.org*. Last modified February 10, 2004. Accessed March 20, 2017. https://www.w3.org/TR/REC-rdf-syntax/.

———. “Schema.” *W3.org*. Accessed April 14, 2017. https://www.w3.org/standards/xml/schema.

———. “W3C Math Home: What Is MathML?” *W3.org*. Accessed August 6, 2017. https://www.w3.org/Math/.

———. “XSL Transformations (XSLT) Version 1.0.” *W3.org*. Last modified November 16, 1999. Accessed April 14, 2017. https://www.w3.org/TR/xslt.

W3Schools. “XML Namespaces.” *W3schools.com*. Accessed March 26, 2017. https://www.w3schools.com/XML/xml_namespaces.asp.

———. “XML Tutorial.” *W3schools.com*. Accessed April 14, 2017. https://www.w3schools.com/xml/.

Wareable. “Alexa Gets in Your Face with LET Labs’ New Smartglasses.” *Wareable.com*. December 14, 2017. https://www.wareable.com/smartglasses/alexa-let-labs-glasses–291.

Wikipedia. “Comparison of Web Browsers: Image Formats.” *Wikipedia*. Accessed January 28, 2018. https://en.wikipedia.org/wiki/Comparison_of_web_browsers#Image_format_support.

———. “Grounded Theory.” *Wikipedia*. Accessed January 8, 2018. https://en.wikipedia.org/wiki/Grounded_theory.

———. “List of Microsoft Windows Versions.” *Wikipedia*. Accessed March 19, 2017. https://en.wikipedia.org/wiki/List_of_Microsoft_Windows_versions.

———. “List of Web Archiving Initiatives.” *Wikipedia*. Accessed October 8, 2017. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

———. “MacOS Version History.” *Wikipedia*. Accessed May 19, 2017. https://en.wikipedia.org/wiki/MacOS_version_history.

———. “Microsoft Word.” *Wikipedia*. Last modified April 18, 2017. Accessed April 16, 2017. https://en.wikipedia.org/wiki/Microsoft_Word.

———. “Observer-Expectancy Effect.” *Wikipedia*. Accessed January 8, 2018. https://en.wikipedia.org/wiki/Observer-expectancy_effect.

———. “Time Line of Microsoft Windows.” *Wikipedia*. Last modified February 25, 2017. Accessed March 5, 2017. https://en.wikipedia.org/wiki/Timeline_of_Microsoft_Windows.

Willett, Perry, and John Kunze. “ARK (Archival Resource Key) Identifiers.” *UCOP.edu*. Last modified October 24, 2016. Accessed March 19, 2017. https://wiki.ucop.edu/display/Curation/ARK.

Woods, Kam, Christopher A. Lee, Oleg Stobbe, Thomas Liebetraut, and Klaus Rechert. “Functional Access to Forensic Disk Images in a Web Service.” *Proceedings of the 12th International Conference on Digital Preservation*. Chapel Hill, North Carolina, November 2–6, 2015, 191–95. Accessed January 28, 2018. Handle: 11353/10.429627.

Yakel, Elizabeth. “Digital Curation.” *OCLC Systems and Services: International Digital Library Perspectives* 23 (2007): 4, 335–40. Accessed March 6, 2017. doi:10.1108/10650750710831466.

Index

AAT. *See* Art & Architecture Thesaurus
A-B type test, 91. *See also* user studies
Access, 1–8, *3,* 49–72, 87–98; levels of, 27, 94–95; long-term, 3–4, 7–8, 52, 104–5
administrative metadata. *See* metadata, administrative
administrators, roles for, 17–18
AIMS (An Inter-Institutional Model for Stewardship), 71–72
archival formats, *7*, 66–68, 84–86
Archival Resource Keys, 56
Archivematica, 61, 72
archivists, roles for, 12–13, 15–17, 19–20
ARK. *See* Archival Resource Keys
Art & Architecture Thesaurus, 52. *See also* controlled vocabularies
authenticity, 5–6, 75, 78–79*,* 83–84

BitCurator, 42, 61, 72, 95–96; Consortium, 100
bitstream, 64. *See also* PREMIS
business continuity, 18, 27, 78, 83, 101
bwFLA, 34. *See also* emulation

checksum, 13, 60, 64, 77, 78; verification of, 23, 41, 83–84
CLOCKSS, 79, 101. *See also* LOCKSS
Cobweb, 103
collection, defining a, 94–95
collection development, 37, 71, 94,
community: leveraging the, 99–107; roles for, 19–20
computer museum approach, 32, 36
conferences, 20, 35, 105–6, 114
content standardization, 4, 7, 49–67
context, 3, 7, 41–42, 52, 71
controlled vocabularies, 52, 57–58, 68, 94, 113
copyright, 45–47, 50–52, 65, 110
Creative Commons, 13, *14*, 47, 51. *See also* metadata, rights issues
creators: guidelines for, 12–13, 16, 29, *47;* working with, 11–13, 21, 46–47
curation, 3–8, 21–23

data, 3, 12, 53–56, 105; visualization of, 34, 97
Data-PASS, 104, 114
date format, 8, 56
DC. *See* Dublin Core
DCC. *See* Digital Curation Centre
derivatives, 24; selection, 97; creation, 32
descriptive metadata. *See* metadata, descriptive
designated community, *22*, 27, 38
DigiPres Commons, 42, 100
Digital Curation Centre (DCC), 20–21, 44–45, 105–6; Curation Lifecycle Model, 21, *22,* 23, 28–29, 105
digital dark age, 2–3
Digital Object Identifier (DOI), 56
Digital Preservation Coalition (DPC), 45, 100, 105
Digital Preservation Network (DPN), 101, 103
Digital Preservation Outreach and Education (DPOE), 25–29, 105
Dioscuri project, 33. *See also* emulation
disaster recovery, 27, 76–77, 79–82, 86, 101
DOI, 56
donations, 45–46, 51–52
DPOE. *See* Digital Preservation Outreach and Education (DPOE)

DROID, 61, 63, 86, 99. *See also* file testing
Dublin Core, 55–57. *See also* metadata, descriptive

EAC-CPF, 94. *See also* Encoded Archival Description
EAD (Encoded Archival Description), 4, 57, 94. *See also* metadata, descriptive; EAC-CPF
efficiency, measures of, 89–90. *See also* user studies
embedded metadata. *See* metadata, embedded
emulation, 7, 31–34, 36, 49, 78, 95
Emulation As A Service (EAAS), 34, 95, 113
Emulation of Multimedia-objects in Libraries (EMiL), 34
encapsulation, 7, 31–33, 36
Encoded Archival Description. *See* EAD
endianness, 32, 42
eXtensible Markup Language. *See* XML

FAST (Faceted Application of Subject Terminology), 52, 113. *See also* controlled vocabularies
father of the internet. *See* Vint Cerf
feedback, 40, 82, 87–93. *See also* user studies
file: fixity, 23, 24, 60, 79; naming, 47, 53, 55, *59*; testing, 60–63, 68–70, 72, 81, 84. *See also* technical metadata
FITS (File Information Tool Set), 52–54, 68–70. *See also* file testing
fixity. *See* file, fixity. *See also* checksum, verification of
focus groups, 88, 91–92. *See also* user studies
format registry, 61, 99, 113–14
formats, 3–4, 32, 34, 60–64, 111; archival, 84, 86; and creators, 12, 47; derivative, 97; and digitization, 15; and donors, 46; and migration, 7–8, 66–68; testing, 61–63

Geography Markup Language (GML), 53. *See also* metadata standards
Global LOCKSS Network (GLN), 101. *See also* LOCKSS
grounded theory, 91. *See also* user studies

handles, 56. *See also* persistent identifiers
hard drive analysis, 41–42, 45–47, 62, 64; tools for, *41,* 62, 72, 95, 100
identification, 26*,* 37, 39–40, 55; tools for, 41, 61–62, 100
in-house content, 45–46
institutions, roles for, 19
Intellectual Entity, 64. *See also* PREMIS
Intellectual Property, 45–46, 50–51, 110. *See also* rights
Internet Archive, 20, 43–44, 50, 83, 95; Web Archiving Service, 45, 95, 103; Web Archiving Lifecycle Model, 44
InterPARES Trust, 105
interviews, 88–92. *See also* user studies
ISO 8601 date format, 56. *See also* controlled vocabularies

JavaScript Object Notation (JSON), 56
JHOVE, 61–63, 69, 100. *See also* file testing
journals, 79, 92, 101, 105–7, 109–10
JSON, 56

KEEP Project, 33–34. *See also* emulation

LCSH. *See* Library of Congress Subject Headings
levels: of access. *See* access, levels of; of preservation, 65–66, *80*; of rights, *14*
librarians, roles for, 11–20
Library of Congress Subject Headings, 52, 113. *See also* controlled vocabularies
Likert Scales, 89. *See also* user studies
linked data, 53, 56–57, 113. *See also* metadata standards
listservs, 20, 93, 104–6
LOCKSS, 79, 83, 101–3
long-term access. *See* access, long term

metadata, *7,* 15, 49–66; administrative, 8, 64–66, 72; descriptive, 15, 43, 52–58, 64; embedded, 12–13, 15, 47, 63, 67; rights, 8, 12–16, 51, 57, 63–64, 110; standards, 53, 56, 63, 111–12; structural, 7, 58–60; technical, 34, 58, 60–72; *See also* Creative Commons; RightsStatements.org; Open Digital Rights Language
METS (Metadata Encoding Transmission Standard), 58, 64. *See also*: metadata, structural; metadata, standards
migration, 7, 16, 34–36, 72; definition of, 31, 34; testing, 67–68, 84–86
MODS (Metadata Object Description Schema), 54–57. *See also* metadata, descriptive
monitoring content, 27, 75–86, 103. *See also* checksums, verification of; metadata, administrative
More Product, Less Process (MPLP), 57, 95

NAF (Name Authority File), 52. *See also* controlled vocabularies

National Digital Stewardship Alliance, 20, 79, 106; levels of preservation, *80*
NDSA. *See* National Digital Stewardship Alliance
normalization, 7, 67, 71

OAIS model, 23–25, 27–29
observer effect, 92. *See also* user studies
Obsolete Floppy Disks, *5*
OCR (optical character recognition), 35, 55
ODRL. *See* Open Digital Rights Language
Olive, 34. *See also* emulation
Open Archival Information System (OAIS), 23–25, 27–29
Open Digital Rights Language (ODRL), 51, 57. *See also* metadata, rights
Open Preservation Foundation (OPF), 100
Open Science Framework (OSF), 101, 103
organizations, roles for. *See* institutions, roles for
Orphan Works Act of 2008, 50

PANDAS (PANDORA Digital Archiving System), 103–4
PANDORA, 103
persistent identifiers, 55, 104. *See also* unique identifiers
PREMIS (PREservation Metadata Implementation Strategies), 64–66. *See also* metadata, standards
preservation metadata, 64. *See also under* metadata
PrestoCentre, 104
Private LOCKSS Network (PLN), 79, 101–2. *See also* LOCKSS
provenance, 12, 64; and authenticity, 4–6, 23–24, 78–79, 86, 94
PURL. *See* Persistent URL

RDF. *See* Resource Description Framework
representation: information, 23, in PREMIS, 64, *65*; selection, 12, 15, 42, 45, 97. *See also* PREMIS
resolution, image, 97
Resource Description Framework (RDF), 56. *See also* metadata standards
rights, 8, 50–52, 63–64, 71–72, 110; guidance for donors, 46; impact in selection, 39–40; levels of, 14; obtaining, 13, 15; "opt-in" versus "opt-out" model for managing, 44–45; sample document for donors, 51. *See also* metadata, rights
RightsStatements.org, 51. *See also* metadata, rights
roles, 11–20

SafeArchive software, 104
safety, 75–76
Sample Script, 68–70
schema, 54, 56, 57–58, 62. *See also* XML
scripting languages, 70
security, 16, 27, 32, 75–81, 86
selection, 13, 15, 26–29, 37–46
significant properties, 35–36, 66, 71, 83–84, 103–4. *See also* PREMIS; migration
social desirability bias, 93. *See also* user studies
social media, 3, 42, 45–46*,* 103
software registry, 86
standards, 51–58, 61–68, 105–6, 110–12; and standardization, 4, 7–8, 25. *See also* metadata standards
storage, 24–25, 27–29, 75–83, 101–3; bit loss, 60; devices, 46–47; of files, 59–60
streaming server, 97
structural metadata. *See* metadata, structural
surveys, 88–89, 93–94. *See also* user studies

technical metadata. *See* metadata, technical
technical registry, 42, 68, 99–100, 113–14
TGM. *See* Thesaurus for Graphic Materials
TGN. *See* Thesaurus of Geographic Names
Thesaurus for Graphic Materials, 52. *See also* controlled vocabularies
Thesaurus of Geographic Names, 52. *See also* controlled vocabularies
tools registry, 42, 68, 86, 100, 113–14
Trustworthy Online Technical Environment Metadata Database (TOTEM), 34. *See also* KEEP Project

unique identifiers, 53, 55, 56, 58, 72. *See also* persistent identifiers
user: needs, 57, 87–88, 90–92; observations, 88, 92; performance, 89; satisfaction, 89; studies, 88–91. *See also*: interviews, focus groups, observations, feedback, surveys
validation. *See* file testing

VeraPDF, 63, 100. *See also* file testing
VIAF, 52, 57
Vint Cerf, 2–3
Virtual International Authority File (VIAF), 52, 57. *See also* controlled vocabularies
virtual machine, 32, 72

W3C, 51, 64
web archiving, 28, 42–45, 67, 95, 103–4
web logs, 88, 92
World Wide Web Consortium (W3C), 51, 64
XML, 53, 54–55
XSLT, 54–55, 70

About the Author

Jody L. DeRidder is director of metadata frameworks in the Metadata Strategy and Operations Division of OCLC, Inc., where she is facilitating the development of methods to improve the effectiveness and efficiency of managing various kinds of metadata and information. During the prior nine years, she was head of metadata and digital services at the University of Alabama, where she developed practical policies, procedures, and infrastructure for digital preservation, digitization, metadata workflows, and web delivery. There, she developed a tiered and targeted approach to digital preservation for constantly expanding holdings, focusing on practical solutions. In 2011, DeRidder was one of the first contingent of invited participants in the Digital Preservation and Outreach Train-the-Trainer program, and the webinars she subsequently provided through the Association of Southeastern Research Libraries broke all records for attendance and continue to enjoy hundreds of downloads monthly. In 2014, she was honored to be a keynote speaker for the Best Practices Exchange Conference, which focuses on digital preservation. She writes and presents about preservation and digital libraries in both the library and archival fields. With a master's degree in computer science, as well as one in information sciences, DeRidder combines technical expertise with theoretical knowledge to master the difficult aspects of digital curation and communicate this knowledge effectively to a broad audience.