



DIGITAL PRESERVATION FOR LIBRARIES, ARCHIVES, & MUSEUMS



EDWARD M. CORRADO AND HEATHER LEA MOULAISON

FOR EDUCATIONAL PURPOSES ONLY

Digital Preservation for Libraries, Archives, and Museums

FOR EDUCATIONAL PURPOSES ONLY

FOR EDUCATIONAL PURPOSES ONLY



Digital Preservation for Libraries, Archives, and Museums

Edward M. Corrado and
Heather Lea Moulaison

ROWMAN & LITTLEFIELD

Lanham • Boulder • New York • Toronto • Plymouth, UK

Published by Rowman & Littlefield
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
www.rowman.com

10 Thornbury Road, Plymouth PL6 7PP, United Kingdom

Copyright © 2014 by Edward Corrado and Heather Moulaison

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

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Corrado, Edward M., 1971-

Digital preservation for libraries, archives, and museums / Edward M. Corrado and Heather Lea Moulaison.

pages cm

Includes bibliographical references and index.

ISBN 978-0-8108-8712-1 (pbk. : alk. paper) — ISBN 978-0-8108-8713-8 (ebook) 1.

Digital preservation. 2. Preservation metadata. 3. Electronic information resources—Management. I. Moulaison, Heather Lea. II. Title.

Z701.3.C65C67 2014

025.8'4—dc23

2013034021



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

Figures	xi
Tables	xiii
Foreword by Michael Lesk	xv
Preface	xix
Acknowledgments	xxiii

Part I: Introduction to Digital Preservation

1 What Is Digital Preservation?	3
Digital Preservation Is <i>Not</i> . . .	3
Digital Preservation Is Not Only about Backups and Recovery	3
Digital Preservation Is Not Only about Access	4
Digital Preservation Is Not an Afterthought	5
Elements of Digital Preservation	6
Why Digital Preservation?	6
Digital Preservation: A Management Issue	10
Why Libraries, Archives, Museums?	12
Conclusion	13
2 Getting Started with the Digital Preservation Triad	17
Steps in the Digital Preservation Process	19
The Digital Preservation Triad	21
Management	21
Policies and Planning for Digital Preservation	21
Technology Decisions	23

The Question of Rights	24
Resource Issues	25
Outreach and Sustainability	29
Technology	30
Trustworthy Digital Preservation Systems	30
Metadata	33
File Formats	34
Content	34
Copyright Issues	35
Content-Related Challenges	36
Conclusion	37

Part II: Management Aspects

3 The OAIS Reference Model	43
History	44
OAIS Reference Model Components	44
Vocabulary	45
Information Model	45
OAIS Functional Model	46
OAIS Required Responsibilities	48
Conclusion	52
4 Human Resources and Education	55
Human Resources	55
Categories of Human Resources	56
Education for Digital Preservation	57
Digital Preservation and Digital Curation: What's in a Name?	58
University-Level Education for Digital Preservation	58
Continuing Education for Digital Preservation	61
Research in Digital Preservation	62
Conclusion	65
5 Sustainable Digital Preservation	67
Digital Preservation as Risk Management	68
Involvement in the Creation Process	69
Open and/or Well-Documented Standards and Systems	69
Documentation of Decisions	69
Accepted Standards for Metadata Schemas	69
Needs of the User	70
Exit Strategy	70
Succession Planning	71
Other Considerations for Risk Management	71
Sustainable Digital Resources	72

Blue Ribbon Task Force on Sustainable Digital Preservation and Access	73
Five Conditions Necessary for Digital Preservation Sustainability	74
Factors Affecting Digital Preservation Sustainability	75
Organizational Factors	76
Financial Factors	77
Social and Societal Factors	80
Technological Factors	81
Homegrown, Open Source, and Proprietary Software	
Development Models	82
Memorandums of Understanding (MOUs)	84
Conclusion	89

Part III: Technology Aspects

6 The Digital Preservation Repository and Trust	95
Trust	96
Trusted Repository Criteria and Checklists	97
European Framework for Audit and Certification of	
Digital Repositories	98
TRAC, TRD, and ISO 16363	103
DRAMBORA	106
Conclusion	107
7 Metadata and Metadata for Digital Preservation	111
Metadata in Digital Librarianship	112
Descriptive Metadata	114
Administrative Metadata	114
Technical Metadata	114
Structural Metadata	115
Markup Languages	115
Structure of Metadata Files	117
Metadata Schema	118
Application Profiles	122
Converting Records and Data to a New Format	122
Metadata Generation and Creation	123
Documentation	125
Metadata Necessary for Digital Preservation	125
Preservation Description Information (PDI)	127
Digital Preservation Metadata	131
Metadata Specific to Digital Preservation	132
PREMIS Model	133
Metadata Encoding and Transmission Standard (METS)	136
METS Profiles	137
Conclusion	137

8	File Formats and Software for Digital Preservation	143
	File Formats	144
	File Formats for Digital Preservation	145
	Evaluating File Formats for Digital Preservation	148
	Determining File Formats	155
	File Extensions	155
	MIME Internet Media Types	156
	File format Registries	156
	Why Are Registries So Difficult?	159
	Software to Help Identify File Formats	159
	Generic Tools	159
	File Type Specific Tools	162
	Conclusion	164
Part IV:	Content-Related Aspects	
9	Collection Development	171
	Criteria	173
	Existing Collections	173
	New Collections	173
	Conclusion	175
10	Preserving Research Data	179
	Research Data	180
	Research Data Life Cycle	180
	Big Data	183
	Small Data as Big Data's Counterpart	185
	Metadata Schema for Science Data	185
	Directory Interchange Format (DIF)	185
	The Content Standard for Digital Geospatial Metadata (CSDGM)	186
	Darwin Core Schema	186
	Core Scientific Metadata Model	186
	Harvestable Scientific Metadata	187
	Open Data Initiatives	187
	The U.S. National Science Foundation	188
	The U.S. National Institute of Health	189
	Other U.S. Initiatives	189
	English-Speaking Countries: Approaches to Open Data	190
	Human Subjects and Data Preservation	191
	Challenges with Preserving Human Subject Data	191
	Conclusion	192
11	Preserving Humanities Content	197
	Computerizing the Humanities	199
	Big Data in the Digital Humanities	199

Funding for the Digital Humanities	200
Humanities Sources	200
Metadata Schema for Published Texts	201
Metadata Schema for Digital Texts	202
Metadata Schema for Encoding Visual Resources: Museum Artifacts	203
Metadata Schema for Encoding Video and Sound	205
Conclusion	206
12 Conclusion	211
Appendix A Select Resources in Support of Digital Preservation	213
Selected Digital Preservation Organizations	213
Selected Digital Preservation Consortium/Group Initiatives	214
Data Preservation	214
Other Initiatives	215
Reports	216
General Reports on Digital Preservation	216
Archives	217
Museums	217
Metadata	217
File Formats	217
Moving Images	218
Music	218
Webliographies and Webinars	218
Webliographies	218
Webinars	219
Books, Guides, and Textbooks	219
Online Digital Preservation Glossaries	220
Directories for Digital Preservation Education	220
Centers Supporting Research and Teaching in Digital Preservation	221
Conferences and In-Person Events	221
Core Conferences on Digital Preservation	221
Related Conferences on Digital Preservation	222
Glossary	223
Bibliography	235
Index	259

FOR EDUCATIONAL PURPOSES ONLY

Figures

2.1.	The Digital Preservation Triad	18
2.2.	Cutting an Apple Crosswise with a Paring Knife	24
2.3.	The LIFE Model v2.1	28
3.1.	Interaction of OAIS Functional Entities	48
6.1.	The Data Seal of Approval	100
7.1.	Automatically Generated Digital Photography Metadata	116
7.2.	How Standards Proliferate	120
7.3.	Trove Homepage	124
7.4.	Events for an Information Package in a Digital Preservation System	132
7.5.	Caplan's Figure of PREMIS as a Subset of Preservation Metadata	136
8.1.	A Corrupt Image File	144
8.2.	File format Information about Various Files Detected by the DROID Software	160
8.3.	Results Screen of a File Format Demonstrating the Normalization Process within the Xena Digital Preservation (Open Source) Software	161
8.4.	Output Produced by Executing the FFprobe Command Line Program on an Audio File	164
9.1.	Collection Development Models for Digital Preservation	172
10.1.	The UK Data Archive Research Data Lifecycle Model	181
10.2.	The Scope of Darwin Core and Its Relation to Other Schema and to Relevant Domains	187

FOR EDUCATIONAL PURPOSES ONLY

Tables

1.1.	JISC's Key Aspects of the DPC Digital Preservation Definition	7
3.1.	OAIS Functional Entities	46
3.2.	Categories Related to Having Sufficient Control of Content for Preservation	50
5.1.	Example MOU Worksheet	88
6.1.	Data Seal of Approval Compliance Levels	100
6.2.	Data Seal of Approval 2014–2015 Guidelines (Version 2)	102
7.1.	Four Basic Kinds of Metadata	113
7.2.	OAIS Reference Model Information Necessary for Preservation Description Information (PDI) and Examples	128
7.3.	Common Algorithms Used to Generate Checksums	131
7.4.	PREMIS Data Model Entities	134
8.1.	Binghamton University's Digital Preservation Support Based on File Type	150
9.1.	Factors that May Influence Collection Development Policies	174

FOR EDUCATIONAL PURPOSES ONLY

Foreword

Michael Lesk

Digital preservation is not a problem; it is an opportunity. Until recently we accepted that many creative activities, from poetry reading to broadcast interviews, would be transitory. Even the average written piece of paper would be lost, not because the paper would necessarily turn yellow (we have learned how to make acid-free paper) but because nobody could afford the costs of retaining the paper, describing what was on it, and remembering where it was. Today digital technology is cheap and accessible to everyone. Architects today neither have to worry about the space required to store models of buildings nor about the permanence of cardboard, balsa wood, and foamboard; instead, computer-aided design (CAD) models are universally used and stored. Digital cameras today are so small and cheap that the BBC put cameras on the collars of fifty cats in a rural town and recorded what the cats did all day, producing a program called *The Secret Life of the Cat* (BBC Horizon).

The explosion in quantity produces an explosion in our need to preserve and organize. The cats may be able to take pictures but not yet to tag these pictures with descriptions (and, my wife observed, these cats need to learn about composition). I'm not worried about the BBC, which has an admirable record of retaining its history. We can still hear what William Butler Yeats sounds like because he read his poems on BBC radio in the 1930s. But how does one make this kind of preservation happen?

Unfortunately a large fraction of what has been said about digital preservation has focused on technology: tapes wear out, disks have head crashes, and so on. I am one of the authors who wrote too much about this twenty years ago, not realizing that the media problems would become insignificant compared to the organizational issues. Digital copies are perfect: they are exactly the same as the original, and so multiple copies are nearly always the best answer to the fear of information loss. And so long as the price of disk drives declines by half every eighteen months we can afford to keep the copies of anything we could afford to copy in the first place. But, to repeat,

the problem is not about the weaknesses of media; it is about the weaknesses of organizations and knowledge.

The late Jim Gray used to say, “May all your problems be technical,” expressing his frustration with the complexities of economic, legal, social, and organizational issues. Digital preservation is a fine example: it is not about knowing the mean time to failure of a flash drive but about creating an organizational system that will make our information available in the future. Carving hieroglyphic inscriptions into stone blocks on pyramids did not guarantee intelligibility centuries later; only the accidental survival of the Rosetta Stone, with the same text in both hieroglyphs and Greek, enabled that. Worse yet, we still have difficulty with ancient Mayan texts as a result of deliberate destruction of most of the codices after the conquest of Mexico. Preservation today similarly requires organizational survival, knowledge of formats, understanding of content, and competence in technology.

As a contrast, there are two versions of the U.S. census that have posed preservation issues. The 1890 census records were destroyed by a fire in 1921. More frequently we read about the loss of some digital information from the 1960 census, the first to use digital magnetic tape. The tapes were from an early Univac system, and the drives to read them became obsolete quickly. However, we lost less than 1 percent of the census data, and that mostly because two of the tapes were physically lost. The response to the 1921 fire was in part a new organization, the National Archives. And the response to the tape problems was a managed program of backup copies, now that it was recognized that the very detailed data was in fact worth keeping. Until this episode, the census had routinely discarded the “micro-data” as not worthy of preservation. So, in both cases, the answer is organizations and procedures, not a discussion of sprinklers as opposed to night watchmen or tape durability compared to disk.

The greatest danger to digital materials is that we forget the meaning of them. Preservation depends on our knowledge: we may have bits but be unable to interpret them. Keeping knowledge, rather than objects, is an organizational problem. This book is an excellent description of the issues involved in developing a digital preservation program. It will be useful to people who work in cultural heritage institutions—libraries, archives, and museums—or in institutions that perhaps have not been focused on preservation, such as theater companies or orchestras, but wish to exploit their legacy.

Both the knowledge and organizational issues described in this book are complex and well-explained. A variety of kinds of knowledge must come together in a digital preservation program: knowledge of the content, knowledge of the technology, and knowledge of the procedures used. This poses issues for human resources and educators, and one of the most valuable aspects of this book is its ample references to courses, conferences, and other resources for learning about digital preservation. Even if an organization follows a teamwork model in which different people are handling each aspect of the digital preservation process, it is still important to understand what the other team members are doing.

The importance of copies and of searching in digital preservation makes the organizational problems more serious. To enable other organizations to share copies of material, and to have search engines operate across all of our stored resources, we need interoperable representations and common protocols. This book describes the interworking of the various standards bodies, professional associations, and government or university groups that have created procedures and policies to encourage and facilitate sharing. These policies also reduce the workload of individual organizations and increase the chance of long-term survival.

The book also touches on many of the most delicate organizational issues: legal permissions, sustainable funding, and institutional survival. The habit of doing digitization as “soft money” has led to fears for long-term survival. Examples are the end of funding for the Arts and Humanities Data Service in the United Kingdom (taken over by Kings College London) and the Arabadopsis Information Resource (becoming a consortium). Various strategies are mentioned, but we don’t have a general answer yet.

Sometimes there is an organizational tension between access and preservation. Libraries have always seen this tension when they acquire personal papers that must be kept confidential for a long period; some of Mark Twain’s papers were under a 100-year embargo, requiring preservation activities helping no current users. A 1993 British Library strategic review noted that the library did both access and preservation, access for today’s users and preservation for tomorrow’s users. Only today’s users, however, helped pay the bills. A preservation plan must balance priorities over time.

Finally, the book ends with some of the most important opportunities in the area of data preservation. As of mid-2013, the “big data” craze has demonstrated the importance of keeping large raw-data files from many areas around, and that subject has merged with national policies for preservation of research data that apply in the United States and other countries. Institutional staff not historically concerned with the details of scientific research projects may find themselves with enormous files of data. For example, the Sloan Digital Sky Survey primary site has moved from Fermilab to the University of Chicago Library. That’s 100 terabytes of data, which is far more than the number of bytes you would get if you typed out every book that library owns. Its management involves a knowledge of astronomy and instrumentation and has to be coordinated with astrophysicists around the world.

The authors have tackled the complexity of digital preservation in an intelligible and useful way. Their recommendations apply to both large and small organizations, since they deal with the strategic and policy problems impacting long-term access and storage. The prospects for digital preservation of “big data” may be daunting, but they are exciting. If you wish to learn the area, there is no better introduction than this book.

Dr. Michael Lesk, professor at Rutgers University, has been at the forefront of research in digital libraries since completing his Ph.D. at Harvard in the 1960s. Prior to joining Rutgers University, he headed the Division of Information and Intelligent

Systems at the National Science Foundation. Dr. Lesk received the Flame award for lifetime achievement from USENIX in 1994, is a fellow of the Association for Computing Machinery, and in 2005 was elected to the National Academy of Engineering. He has written extensively on digital libraries and on issues relating to digital preservation, including his 1997 book, *Practical Digital Libraries: Books, Bytes, and Bucks*, and his 2004 book, *Understanding Digital Libraries*, now in its second edition.

FOR EDUCATIONAL PURPOSES ONLY