DIGITAL PRESERVATION FOR LIBRARIES, ARCHIVES, & MUSEUMS

EDWARD M. CORRADO AND HEATHER LEA MOULAISON

Digital Preservation for Libraries, Archives, and Museums

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

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

Tables

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 "microdata" 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.