

Digital Curation

A How-To-Do-It Manual®

Ross Harvey

HOW-TO-DO-IT MANUALS®

NUMBER 170

Neal-Schuman Publishers, Inc.

New York

London



Don't miss the companion website for this book!
Download checklists and templates for developing
digital curation plans and procedures at:
www.neal-schuman.com/curation

Published by Neal-Schuman Publishers, Inc.
100 William St., Suite 2004
New York, NY 10038

Copyright © 2010 Neal-Schuman Publishers, Inc.

“A How-To-Do-It Manual®” and “A How-To-Do-It Manual for Librarians®” are registered trademarks of Neal-Schuman Publishers, Inc.

All rights reserved. Reproduction of this book, in whole or in part, without written permission of the publisher, is prohibited.

Printed and bound in the United States of America.

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 Z39.48-1992.

Library of Congress Cataloging-in-Publication Data

Harvey, D. R. (Douglas Ross), 1951-
Digital curation : a how-to-do-it manual / Ross Harvey.
p. cm. — (How-to-do-it manuals ; no. 170)
Includes bibliographical references and index.
ISBN 978-1-55570-694-4 (alk. paper)
1. Digital libraries. 2. Digital preservation. 3. Digital libraries—Management. I.
Title.

ZA4080.H37 2010
025.00285—dc22

2010020400

Contents

List of Figures	ix
List of Abbreviations	xi
Preface	xv
Acknowledgments	xxi
Part I. Digital Curation: Scope and Incentives	1
Chapter 1. Introduction	3
Why There Is a Need for Digital Curation	3
What Digital Curation Is	5
Why We Should Be Interested in Digital Curation	8
Incentives for Digital Curation	11
Direct Benefits to Data Creators	12
“Public Good” Obligations	13
Compliance Reasons	13
Digital Curators	14
Summary: Main Characteristics of Digital Curation	15
References	16
Chapter 2. The Changing Landscape	19
Cyberscholarship: New Ways of Working	19
Cyberscholarship in Practice	21
E-science	21
Cyberscholarship’s Requirements and Challenges	22
Content	22
Tools and Services	23
Expertise	23
Digital Curation: A New Profession, New Requirements	25
Educating and Training Digital Curators	28
Summary: Meeting the New Demands	29
References	29
Chapter 3. Conceptual Models	33
The DCC Curation Lifecycle Model	34
The Digital Curation Centre	35
Other Lifecycle Models	36

The OAIS Reference Model	38
OAIS Functions	38
Actors and Objects	39
OAIS Information Packages	40
OAIS and the DCC Curation Lifecycle Model	41
Summary: The Importance of Models	43
References	43
Chapter 4. Defining Data	45
Data as Digital Heritage	45
Born-Digital <i>and</i> Digitized Data	46
Data—And Much More	47
Metadata Is Data Too	49
Databases	50
Summary: New Kinds of Data	50
References	51
Part II. Key Requirements for Digital Curation	53
Chapter 5. Curation and Curators	55
Aims of Digital Curation	55
Scope of Digital Curation	56
Ensuring Longevity	57
Ensuring Integrity	57
Maintaining Accessibility	57
Roles of Digital Curators	58
Funding Bodies	59
Discipline Groups	59
Data Creators	60
Data Users and Reusers	60
Data Curators	60
Summary: Managing Curation	62
References	62
Chapter 6. Description and Representation Information	65
The Need for Description and Representation Information	66
Definitions	67
Standards for Description and Representation Information	68
Description Information	69
Preservation Metadata	70
Persistent Identifiers	72
Metadata Schemas and Standards	72
PREMIS	73
METS	74
MODS and MADS	75
Representation Information	76
OAIS and Representation Information	77
Sharing Representation Information	78
Policies for Description and Representation Information	79

Summary: Curation Needs Metadata	79
References	80
Chapter 7. Preservation Planning and Policy	83
Risk Management as the Context for Preservation Planning	83
Key Planning Steps	84
Planning for Sequential Actions	85
Planning for Full Lifecycle Actions	85
Policy for Curation	86
What Policies Address	86
Kinds of Policies Required	88
Costs of Curation	89
Summary: Planning for Active Management	91
References	91
Chapter 8. Sharing Knowledge and Collaborating	93
Keeping Up-to-Date	93
Starting Points	94
Online Tutorials	94
Project Websites	95
Blogs and E-mail Lists	95
Online Journals	96
Other Sources	96
Collaboration: Intrinsic to Digital Curation	96
Standards: Essential for Digital Curation	98
Tools and Toolkits	99
Summary: Collaboration Is the Key	100
References	101
Part III. The Digital Curation Lifecycle in Action	103
Chapter 9. Designing Data	105
Designing Curation-Ready Data	106
Importance of Standards	108
Designing Projects with Curation in Mind	109
Three Examples	110
Summary: Planning Data for Curation	112
References	112
Chapter 10. Creating Data	115
Policies for Creating and Receiving Data	116
Creating Data for Curation	117
Structuring Data for Use and Reuse	118
Open Formats and Open Source	119
Significant Properties and Authenticity	120
Documentation	122
Influencing Data Creators	123
Structuring Data for Management	124
Data Management	124
Data Quality	125

Structuring Data for Discoverability	126
Receiving Data for Curation	127
Summary: The Positive Effects of “Good” Data	128
References	128
Chapter 11. Deciding What Data to Keep	131
What Is Appraisal?	131
What Data Do We Want to Keep?	133
Drivers for Keeping Data	134
Why We Can’t Keep Everything	136
How Long Do We Want to Keep Those Data?	137
Appraisal and Selection Policies	138
Who Decides?	139
Appraisal Tools	140
Two Examples	142
Selection and Archiving of Websites	142
Appraisal of Scientific Data Sets	143
Reappraisal	144
Disposal of Data	145
Transfer of Data	146
Destruction of Data	146
Summary: The Necessity for Appraisal and Selection	147
References	147
Chapter 12. Ingesting Data	151
OAIS and Ingest	152
Ingest Processes in More Detail	152
Submitting SIPs	153
Receiving SIPs	153
Generating AIPs	155
Ingest Tools	156
Policies for Ingest	158
Summary: Automation Is the Key	158
References	158
Chapter 13. Preserving Data	161
Digital Preservation Methods	162
Migration in Practice	165
Implementing Migration	170
Migration Changes Data	170
Automating Preservation Actions	171
Tools	171
Metadata Tools	173
Format Validation, Format Registry, and Obsolescence	
Notification Tools	173
Normalizing and Encapsulation Tools	174
Migration Tools	174
Emulation Tools	175
Web Archiving Tools	175
Other Curation Tools	175
Tool Development	176

Summary: Methods and Tools	178
References	178
Chapter 14. Storing Data	181
Storage Requirements	182
Organizational Structure and Continuity	183
Technical Infrastructure and Practices	184
Best Practice in Data Storage	184
Ensuring Quality of Storage	186
OAIS Reference Model	186
Trusted Digital Repositories	187
Audit and Certification	188
Backing Up Data	188
Data Security	190
Physical Security	191
Network and File Security	191
Repository Software and Storage Solutions	192
Fedora, DSpace, EPrints, and Other Repository Software	192
Storage Solutions	194
New Models for Collaboration	195
Summary: Storing Data Securely	196
References	197
Chapter 15. Using and Reusing Data	199
Access, Use, and Reuse	199
Sharing Data	200
Building Blocks for Sharing and Reusing Data	202
Standards	203
Standards for Repository Functionality	203
Standards for Interoperability	204
Structuring Data for Access	204
Citing Data	205
Legal Issues	206
Collaboration Processes	207
Annotation	207
Provenance	209
Access Controls and Authentication Procedures	210
Transform	211
Migration	211
Creating New Data	213
Conclusion: The Lifecycle Continues	213
References	214
Index	217
About the Author	225

List of Figures

Figure 1.1	Threats to Digital Continuity	9
Figure 2.1	Core Skills for Data Management	26
Figure 2.2	Comparison of Skills Required for Digital Curation	27
Figure 3.1	The Life Cycle of Research Knowledge Creation	36
Figure 3.2	Digital Archives and the Records Cycle	37
Figure 3.3	OAIS Functional Entities	40
Figure 3.4	Correlations between the DCC Curation Lifecycle Model and the OAIS Reference Model	42
Figure 6.1	Description Information and Its Functions	70
Figure 10.1	Selecting File Formats for Data Curation	121
Figure 11.1	Risks Matrix	135
Figure 13.1	Comparison of the Main Digital Preservation Methods	164
Figure 15.1	DISC-UK DataShare: Data Sharing Continuum	212

List of Abbreviations

ACE	Audit Control Environment
AHDS	Arts and Humanities Data Service
AIFF	Audio Interchange File Format
AIP	Archival Information Package
ALA	American Library Association
AONS	Automatic Obsolescence Notification System
API	Application Programming Interface
APSR	Australian Partnership for Sustainable Repositories
ARCHER	Australian Research Enabling Environment
BADC	British Atmospheric Data Centre
BAT	BnF Arc Tools
BBSRC	Biotechnology and Biological Sciences Research Council
BMP	Bitmap
CAD	Computer-Aided Design
CAIRO	Complex Archive Ingest for Repository Objects
CASPAR	Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval
CHIN	Canadian Heritage Information Network
CIC	Committee on Institutional Cooperation
CLADDIER	Citation, Location, and Deposition in Discipline and Institutional Repositories
CRC	Cyclic Redundancy Checks
CRiB	Conversion and Recommendation of Digital Object Formats
Data-PASS	Data Preservation Alliance for the Social Sciences
DCC	Digital Curation Centre
DDI XML	Data Documentation Initiative XML
DIFFUSE	Dissemination of Informal and Formal Useful Specifications and Experiences
DigCCurr	Digital Curation Curriculum
DIP	Dissemination Information Package

DISC-UK	Data Information Specialists Committee—United Kingdom
DOI	Digital Object Identifier
DPC	Digital Preservation Coalition
DPE	Digital Preservation Europe
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DROID	Digital Record Object Identification
DTD	Document Type Definition
EAD	Encoded Archival Description
ECDL	European Conference on Digital Libraries
EML	Ecological Markup Language
EROS	Earth Resources Observation and Science
ERPANET	Electronic Resource Preservation and Access Network
ESDS	Economic and Social Data Services
ESRC	Economic and Social Research Council
EU	European Union
Exif	Exchangeable Image File Format
FAT	File Allocation Table
Fedora	Flexible Extensible Digital Object and Repository Architecture
FITS	File Information Tool Set
FLAC	Free Lossless Audio Codec
GIF	Graphics Interchange Format
GIS	Geographic Information System
GPS	Global Positioning System
GRATE	Global Remote Access to Electronic Services
HATII	Humanities Advanced Technology and Information Institute
HOPPLA	Home and Office Painless Persistent Long-Term Archiving
HTML	Hypertext Markup Language
IB	Integrative Biology
ICADL	International Conference on Asian Digital Libraries
IIPC	International Internet Preservation Consortium
InSPECT	Investigating the Significant Properties of Electronic Content over Time
InterPARES	International Research on Permanent Authentic Records in Electronic Systems
IP	Internet Protocol
iPres	International Conference on Preservation of Digital Objects

List of Abbreviations

ISMS	Information Security Management System
ISO	International Organization for Standardization
JCDL	Joint Conference on Digital Libraries
JHOVE	JSTOR/Harvard Object Validation Environment
JISC	Joint Information Systems Committee
JPEG	Joint Photographic Experts Group
KEEP	Keeping Emulation Environments Portable
koLibRI	kopal Library for Retrieval and Ingest
LIFE	Lifecycle Information for E-literature
LOCKSS	Lots of Copies Keep Stuff Safe
MADS	Metadata Authority Description Schema
MARC	MAchine Readable Cataloging
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema
MPEG	Moving Picture Experts Group
NDIIPP	National Digital Information Infrastructure and Preservation Program
NERC	National Environment Research Council
NIH	National Institutes of Health
NISO	National Information Standards Organization
NISPOM	National Industrial Security Program Operating Manual
NLNZ	National Library of New Zealand
NOST	NASA/Science Office of Standards and Technology
NSF	National Science Foundation
NTFS	New Technology File System
OAIS	Open Archive Information System
ODF	Open Document Format
PADI	Preserving Access to Digital Information
PARADIGM	Personal Archives Accessible in Digital Media
PARSE	Permanent Access to the Records of Science in Europe
PDF	Portable Document Format
PDF-A	Portable Document Format—Archival
PDI	Preservation Description Information
PeDALS	Persistent Digital Archives and Library System
PGP	Pretty Good Privacy
PLANETS	Preservation and Long-Term Access through Networked Services
PLATTER	Planning Tool for Trusted Electronic Repositories
PLN	Private LOCKSS Network

PNG	Portable Network Graphics
PoWR	Preservation of Web Resources
PREMIS	Preservation Metadata: Implementation Strategies
PURL	Persistent Uniform Resource Locator
RIN	Research Information Network
RODA	Repositório de Objectos Digitais Autênticos
RTF	Rich Text Format
SAA	Society of American Archivists
SAS	Statistical Analysis System
SHAMAN	Sustaining Heritage Access through Multivalent ArchiviNg
SHERPA	Securing a Hybrid Environment for Research Preservation and Access
SIP	Submission Information Package
SPSS	Statistical Package for the Social Sciences
TDR	Trusted Digital Repository
TIFF	Tagged Image File Format
TRAC	Trusted Repositories Audit and Certification
URL	Uniform Resource Locator
URN	Uniform Resource Name
UVC	Universal Virtual Computer
WAV	Waveform Audio File Format
Xena	XML Electronic Normalizing for Archives
XML	eXtensible Markup Language

Preface

We live in an environment where data (information in binary digital form) surrounds us and is essential for most activities in which we participate. Librarians and archivists act as data creators, data users and reusers, and/or data curators in increasingly digitally oriented environments. Despite this, professional practice has not caught up with digital practice in many respects. Caring for data, ensuring its usability and reuse in the future, and ensuring its accessibility and understandability over time require new strategies, practices, and tools. Traditional library and archival practices developed in a predigital and largely paper-oriented environment do not automatically transfer to the current digitally oriented environments. Although the past decade has seen the rapid development of new strategies, practices, and tools, these are not yet sufficiently mature. Consider these facts:

- Immense quantities of information in binary digital form are being generated in all walks of life.
- The quantities are increasing at a rapid rate.
- The scientific, scholarly, and research communities increasingly rely on networked computing, as trends such as the move from *in vitro* to *in silico* science and the development of large digital libraries in the humanities become dominant.
- Computer technology (hardware, software, and communications networks) quickly becomes obsolete.

All of these place data at risk from factors such as technology obsolescence, digital object fragility, a lack of understanding about what constitutes good practice, insufficient resources, and inappropriate organizational infrastructure.

Although the body of practice known as *digital preservation* is developing to address the factors that place data at risk, it is starting to be commonly accepted that its outcomes provide only part of the answer. For example, it is relatively straightforward to maintain a bit stream over time: there are more than 40 years of practice to call on in this respect. However, there is no guarantee that the data represented in this bit stream have the characteristics that allow them to be used and understood in the future, and to remain unchanged. How can these

characteristics be retained in the data that is maintained for use in the future?

To answer this question effectively, more than simply a focus on maintaining the data (i.e., digital preservation) is required. What must also be considered is what comes before preservation and what comes after—that is, how the data are created and how they are used before they get to an archive or library and how they will be used, and by whom, in the future. This requires a focus on data that differs from that applied to physical artifacts such as books, manuscripts, and photographic prints in a predigital environment.

Digital curation is a developing set of techniques that address these issues, emphasizing the maintenance of data and adding value to these data for current and future use. Because it is still developing, digital curation is not yet described in detail in the literature. *Digital Curation: A How-To-Do-It Manual* therefore makes a significant contribution by describing in detail, in one place, the basics and current practices of digital curation.

Various models of the lifecycle of data are available. These typically begin with the creation of data and move through its various stages, ending with data use. The *Curation Lifecycle Model* developed by the Digital Curation Centre (DCC) (DCC, 2008; Higgins, 2008) is one of these. It was developed by the DCC to describe the processes involved in digital curation. The DCC Curation Lifecycle Model encompasses data from their conceptualization and creation through all aspects of their selection, archiving, maintenance, and use, to their reuse in the future. The DCC Curation Lifecycle Model provides an action-oriented structure for this book.

Digital Curation: A How-To-Do-It Manual is intended for anyone who creates data, anyone who uses and reuses data, and anyone who curates data. In essence, this means everyone who uses computers. More specifically, this book is intended to be read by librarians and archivists and by students of these professions. It should also have wider appeal, for example, to scientists and scholars who plan research and collect and use data. Whoever its readers, it will assist them to incorporate curation procedures, where relevant, into their own practice, figure out where to start when developing and implementing digital curation processes, and explore digital curation issues by providing a context for digital curation.

Digital Curation: A How-To-Do-It Manual is designed to be read in several ways. Its chapters can be read consecutively as an overview of digital curation, or they can be dipped into for general background and for advice on specific actions. The book's accompanying website (www.neal-schuman.com/curation) provides checklists that can be used separately as guidance and reminders about the tasks that comprise digital curation actions and templates that can be downloaded and used as the basis for developing digital curation plans and procedures for specific libraries, archives, and other organizations, as well as providing guidance for informing individual practice.

This book is based on the author's extensive international experience as a researcher, author, and presenter in the field of digital preservation

and digital curation. It draws in particular on his experience with digital curation in Australia (which is widely acknowledged as representing international best practice) and in the European Union context (including a period based at the Humanities Advanced Technology and Information Institute [HATII] at the University of Glasgow), and his current work at Simmons College in Boston. The book's content was developed through observing digital curation practice, attendance at relevant international conferences, developing material for the DCC, and investigating the real-life experiences of digital curators, especially in the United States. It is also informed by a series of in-depth interviews with digital preservation professionals, which the author carried out as part of the preparation of his book *Preserving Digital Materials* (Harvey, 2005).

Examples of digital curation practice from the United Kingdom and Europe are well represented, as are examples from the United States and from other countries. It may on the surface seem surprising to the American reader that there are not more examples from the United States. There are three main reasons for this. The first is that digital curation is highly international and collaborative to an extent that is perhaps unprecedented in library science and archival practice (as noted in more detail in Chapter 8). This means that developments and practice in the field in one country are keenly observed and adopted, with modifications to suit local requirements, in other countries. The second is that, as Jordan and his colleagues note, "U.S. funding dedicated to digital preservation has traditionally lagged behind that available in the European and British contexts in particular" (Jordan et al., 2008). This means that the large majority of documented examples of digital curation developments and practice have to date come from outside the United States. This is likely to change in the next two to three years as funding such as the National Science Foundation's DataNet program comes on stream. The third is that the more centralized U.K. and European environments have required freely available documentation of any digital curation activities funded by public money. This point applies in particular to the material available on the DCC's website, which represents the only public documentation of a prolonged effort to identify and describe digital curation and to investigate practice in the field. This book therefore makes heavy use, with the permission of the DCC, of the materials accessible through the Centre's website.

Organization

Digital Curation: A How-To-Do-It Manual is organized in three parts. "Part I. Digital Curation: Scope and Incentives" provides a broad context for digital curation by introducing the main concepts and providing an overview. Chapter 1 indicates the reasons why digital curation is necessary, identifies what digital curation encompasses, suggests why one should be interested in digital curation, notes the main incentives for digital curation, and examines who does digital curation and what tasks they carry out. Chapter 2 notes the changing landscape in which

librarians, archivists, researchers, and scholars work, its requirements for different ways of working and new kinds of infrastructure, and the different skill sets for data curation. Chapter 3 examines an important conceptual model for digital curation, the DCC Curation Lifecycle Model, on which this book is based, and a key standard, the OAIS Reference Model. Chapter 4 investigates in more detail what is meant by the term *data* and other related terms. This is important to think about because it allows us to address better an important question—What exactly is it that we want to curate?

“Part II. Key Requirements for Digital Curation” examines the DCC Curation Lifecycle’s Full Lifecycle Actions—the essential basic requirements for all aspects of digital curation, which apply to all of the Sequential Actions noted in Part III of this book. Chapter 5 covers *Curate and Preserve*, one of four Full Lifecycle Actions, noting how digital preservation and digital curation differ, examining the aims of digital curation, and describing how these aims are achieved. Chapter 6 examines another Full Lifecycle Action, *Description and Representation Information*, the metadata and other information required for effective data curation. Chapter 7 notes the essential nature of planning and policy in data curation by describing a third Full Lifecycle Action, *Preservation Planning*. Chapter 8 completes the examination of Full Lifecycle Actions by describing *Community Watch and Participation* and noting the high value placed in digital curation on sharing knowledge and on collaboration.

“Part III. The Digital Curation Lifecycle in Action” is based on the DCC Curation Lifecycle’s Sequential Actions and also notes its Occasional Actions. Chapter 9 notes the Sequential Action “Conceptualise,” stressing the need to think about curation at the very first stages of planning research or creating digital objects. Chapter 10 examines the second Sequential Action, *Create or Receive*, noting the requirements for curation-ready digital objects. Chapter 11 describes *Appraise and Select*, the third Sequential Action, noting the importance of selection of the digital objects to be curated. This chapter also notes the Occasional Actions *Reappraise and Dispose*. The fourth Sequential Action, *Ingest*—the actions required when digital objects are taken into an archiving system—is the topic of Chapter 12. Chapter 13 discusses the preservation strategies and actions associated with *Preservation Action*, the fifth Sequential Action. Also included in this chapter is the Occasional Action *Migrate*. Chapter 14 focuses on the sixth Sequential Action, *Store*, which is concerned with what is required to provide acceptable data storage in the archiving system. Chapter 15 notes the seventh Sequential Action, *Access, Use, and Reuse*, examining the requirements for successful sharing and reuse of data in the future. It also notes the eighth Sequential Action, *Transform*, thus completing the data Curation Lifecycle Model.

A decade ago, little was known about how to assess and preserve the immense body of digitized material that can double in size in a matter of a few short years. Today, we have a body of international experience and expertise to draw on. *Digital Curation: A How-To-Do-It Manual* and its companion website (www.neal-schuman.com/curation) are designed

as a comprehensive resource for best practices in this area. Preserving knowledge is a sacred trust; this resource will enable practitioners in all areas of human experience to better succeed at this crucial task.

References

- Digital Curation Centre. 2008. "The DCC Curation Lifecycle Model." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed April 26, 2010).
- Harvey, Ross. 2005. *Preserving Digital Materials*. Munich: K. G. Saur.
- Higgins, S. 2008. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation* 3, no. 1: 134–140. Available: www.ijdc.net/index.php/ijdc/article/viewFile/69/48 (accessed April 26, 2010).
- Jordan, Christopher, Ardys Kozbial, David Minor, and Robert H. McDonald. 2008. "Encouraging Cyberinfrastructure Collaboration for Digital Preservation." Paper presented at iPres 2008, British Library, London, September 30, 2008. Available: www.bl.uk/ipres2008/presentations_day2/39_Jordan.pdf (accessed April 26, 2010).