

---

# BUILDING DIGITAL LIBRARIES

A How-To-Do-It Manual®

TERRY REESE, JR.  
KYLE BANERJEE

***HOW-TO-DO-IT MANUALS®***

---

***NUMBER 153***

NEAL-SCHUMAN PUBLISHERS, INC.  
New York London

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

Copyright © 2008 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**

ISBN: 978-1-55570-617-3

Library of Congress Cataloging-in-Publication Data

Reese, Terry.

Building digital libraries : a how-to-do-it manual / Terry Reese, Jr., Kyle Banerjee.

p. cm. — How-to-do-it manuals ; no. 153)

Includes bibliographical references and index.

ISBN 978-1-55570-617-3 (alk. paper)

1. Digital libraries. I. Banerjee, Kyle. II. Title.

ZA4080.R44 2008

025.00285—dc22

2008024103

# CONTENTS

<b>List of Figures</b> .....	ix
<b>Preface</b> .....	xiii
<b>1. Planning a Digital Repository</b> .....	1
What Is a Digital Repository? .....	1
The Decision to Build a Digital Repository .....	2
Advantages of a Digital Repository .....	6
Selling the Project .....	7
Understanding the Purpose of the Repository .....	8
Anticipating Use Patterns .....	10
Image and Text Processing .....	12
Archival Image Formats .....	15
Display Image Formats .....	15
Sample Archival Settings .....	16
Rights Management .....	17
Accommodating Changing Formats and Data Structures .....	19
Protecting Integrity of Resources .....	23
LOCKSS (Lots of Copies Keep Stuff Safe) .....	24
Disaster Planning and Security .....	26
Managing with Available Resources and Change ...	26
Further Information .....	29
Summary .....	30
References .....	31
<b>2. Acquiring, Processing, Classifying, and     Describing Digital Content</b> .....	33
Planning Workflow .....	33
Developing the Collection .....	36
Acquiring Digital Content .....	39
Processing and Organizing Digital Content .....	43

Organizing Related Works and Subcollections . . . . .	46
Rights Management . . . . .	48
Batch Processes . . . . .	49
Ergonomics . . . . .	52
Summary . . . . .	54
<b>3. Choosing a Repository Architecture. . . . .</b>	<b>55</b>
Questions to Ask before Choosing an Architecture . . . . .	55
Required Features . . . . .	56
Desirable Features . . . . .	57
Frameworks for Digital Repositories . . . . .	58
Platforms Optimized for Specific Purposes . . . . .	61
Evaluating Repository Functionality . . . . .	69
Resource Identification and Ingestion. . . . .	69
Automating Management and Organization . . . . .	76
Indexing for Easy Retrieval . . . . .	77
Other Storage Challenges. . . . .	80
Repository Administration . . . . .	81
Summary . . . . .	83
References . . . . .	83
<b>4. General Purpose Technologies Useful for     Digital Repositories. . . . .</b>	<b>85</b>
The Changing Face of Metadata . . . . .	85
XML in Libraries . . . . .	86
XHTML . . . . .	88
XPath. . . . .	89
XForms . . . . .	90
XSLT. . . . .	90
XLink . . . . .	90
XQuery . . . . .	91
XPointer . . . . .	91
XML Schema . . . . .	91
Why Use XML-based Metadata? . . . . .	97
XML Is Human-Readable . . . . .	97
XML Offers a Quicker Cataloging Strategy . . . . .	102
XML Can Represent Multi-formatted and Embedded Documents. . . . .	103

---

XML Metadata Becomes “Smarter” . . . . .	105
XML Is Not Just a Library Standard . . . . .	105
Future of Software Development . . . . .	106
Web Services and SOAP . . . . .	108
Sharing Your Services . . . . .	113
Summary . . . . .	114
References . . . . .	115
<b>5. Metadata Formats . . . . .</b>	<b>117</b>
Metadata Primitives . . . . .	117
MARC . . . . .	119
MARC21XML . . . . .	121
Dublin Core . . . . .	123
History . . . . .	123
Elements . . . . .	125
Strengths . . . . .	128
Challenges . . . . .	129
MODS (Metadata Object Description Schema) . . . .	130
History . . . . .	130
Strengths . . . . .	132
Challenges . . . . .	133
METS (Metadata Encoding and Transmission Standard) . . . . .	134
History . . . . .	134
METS at a Glance . . . . .	135
Applications . . . . .	136
Semantic Web . . . . .	138
Application Profiles . . . . .	142
Summary . . . . .	144
References . . . . .	145
<b>6. Sharing Data: Metadata Harvesting and Distribution . . . . .</b>	<b>147</b>
The Evolving Roles of Libraries . . . . .	147
Metadata Wants to Be Free . . . . .	149
Sharing Metadata . . . . .	151
XSLT (eXtensible Stylesheet Transformation) . . .	151
Metadata Crosswalking . . . . .	157
Open Archives Imitative Protocol for Metadata Harvesting (OAI-PMH) . . . . .	161

OAI-PMH Verbs . . . . .	162
OAI-PMH Application . . . . .	169
Facilitating Third-Party Indexing . . . . .	169
Metadata Repurposing . . . . .	170
OSU Electronic Theses Process . . . . .	171
Microformats . . . . .	173
COinS . . . . .	175
UNAPI . . . . .	177
Summary . . . . .	178
References . . . . .	179
<b>7. Federated Searching of Repositories . . . . .</b>	<b>181</b>
What Is Federated Searching? . . . . .	182
Federated Search and Digital Libraries. . . . .	184
Federated Search versus Traditional Search Engines . . . . .	185
Current Research . . . . .	186
Recommender/Collaborative Filtering . . . . .	186
Deduplication of Results . . . . .	186
Knowledge-Base Management. . . . .	187
Automatic Data Classification . . . . .	187
Ranking Systems . . . . .	187
Need for Speed . . . . .	188
Searching Protocols . . . . .	188
Z39.50 . . . . .	189
SRU/SRW (Search/Retrieval/URL and Search/Retrieval Web Service). . . . .	192
OpenSearch . . . . .	197
Linking Protocols . . . . .	201
OpenURL . . . . .	202
DOI (Digital Object Identifiers) . . . . .	204
Search Engine Support. . . . .	206
Service Registries . . . . .	207
Service. . . . .	209
Agent. . . . .	210
Collection . . . . .	211
Evaluating Needs . . . . .	215
Developmental Needs . . . . .	215
User Needs . . . . .	215

---

Summary . . . . .	216
References . . . . .	217
<b>8. Access Management . . . . .</b>	<b>219</b>
Copyright Issues . . . . .	220
Copyright as Organizational Policy . . . . .	222
Can It Be Archived? Can It Be Distributed? . . . . .	226
Long-Term Rights Management . . . . .	227
Allowing/Restricting Access . . . . .	228
CONTENTdm . . . . .	229
DSpace . . . . .	231
Control Mechanisms . . . . .	232
LDAP . . . . .	233
Shibboleth . . . . .	234
OpenID . . . . .	235
Athens . . . . .	235
Monitoring Repository Use and Statistics . . . . .	236
Intellectual Property . . . . .	236
Service Usability . . . . .	236
Statistical Analysis . . . . .	238
Web Spiders/Harvesters . . . . .	238
Item Prefetching . . . . .	239
Summary . . . . .	239
References . . . . .	240
<b>9. Planning for the Future . . . . .</b>	<b>241</b>
Providing Information That People Need . . . . .	241
Libraries' New Roles . . . . .	243
Learning from the Past . . . . .	244
Adapting to Change . . . . .	247
Consolidation and Specialization . . . . .	249
Federated Collection Management . . . . .	251
Federated Vocabularies . . . . .	255
Summary . . . . .	256
References . . . . .	257
<b>10. Conclusions . . . . .</b>	<b>259</b>
<b>Index . . . . .</b>	<b>267</b>
<b>About the Authors . . . . .</b>	<b>277</b>



# LIST OF FIGURES

Figure 1-1.	1964 Flood in Corvallis, Oregon . . . . .	4
Figure 1-2.	Flooded Housing Area . . . . .	4
Figure 1-3.	Oregon State University Digital Library Resources Found Using Google . . . . .	5
Figure 1-4.	Oregon State University Libraries Image Capture Guidelines . . . . .	17
Figure 1-5.	Digital Access Model . . . . .	21
Figure 1-6.	Lot of Copies Keep Stuff Safe (LOCKSS) . . . . .	25
Figure 2-1.	Digital Repository Workflow . . . . .	35
Figure 2-2.	Serving Digital Content . . . . .	45
Figure 2-3.	Repository Acquisitions Tool . . . . .	50
Figure 2-4.	Repository Acquisitions Tool . . . . .	53
Figure 3-1.	Fedora Service Architecture . . . . .	60
Figure 3-2.	CONTENTdm Page Addition Form . . . . .	63
Figure 3-3.	DSpace Page Addition Form . . . . .	64
Figure 3-4.	CONTENTdm Controlled Vocabulary Management . . . . .	65
Figure 3-5.	CONTENTdm Field Definition . . . . .	66
Figure 3-6.	CONTENTdm Field Registry . . . . .	67
Figure 3-7.	DSpace Field Registry . . . . .	68
Figure 4-1.	XML Family Tree . . . . .	87
Figure 4-2.	XML Display from an ILS . . . . .	95
Figure 4-3.	XML Display from CONTENTdm . . . . .	96
Figure 4-4.	MARC Record . . . . .	98
Figure 4-5.	XML Representation of Figure 4-4 . . . . .	99
Figure 4-6.	Unreadable MARC Record . . . . .	101
Figure 4-7.	Sample Data Dictionary . . . . .	102
Figure 4-8.	Part of an EAD Record . . . . .	104
Figure 4-9.	eBay + Google Maps . . . . .	107
Figure 4-10.	The U.S. Library of Congress SRU Response . . . . .	109

Figure 4-11.	Personalized Google Home Page with CONTENTdm Widget . . . . .	110
Figure 4-12.	CONTENTdm Google Widget Snippet, in PHP . . . . .	111
Figure 4-13.	Ruby WSDL Example . . . . .	112
Figure 5-1.	Example of MARC21 Record . . . . .	119
Figure 5-2.	Plain Text View of MARC Record . . . . .	121
Figure 5-3.	MARC21XML Record. . . . .	122
Figure 5-4.	Dublin Core Meta Tag Example. . . . .	124
Figure 5-5.	Unqualified Dublin Core . . . . .	126
Figure 5-6.	DSpace Dublin Core Display . . . . .	128
Figure 5-7.	Example of a MODS Record . . . . .	131
Figure 5-8.	Example of METS with Dublin Core . . . . .	136
Figure 5-9.	Example of METS Navigator. . . . .	137
Figure 5-10.	Example of RDF/XML Encoding of Dublin Core Data. . . . .	140
Figure 6-1.	XML Document Transformed Using XSLT . . . . .	152
Figure 6-2.	Using XSLT to Transform an XML Document to HTML Sorted by Title . . . . .	153
Figure 6-3.	XSLT/XML Transformed Output . . . . .	155
Figure 6-4.	XSLT Excerpt from a Larger XSLT Document. . . . .	156
Figure 6-5.	Dublin Core to MARC21 Author Crosswalk . . . . .	160
Figure 6-6.	Response to a Get Record Request. . . . .	162
Figure 6-7.	Response to an Identify Request . . . . .	164
Figure 6-8.	Response to a Request Using List Metadata Formats . . . . .	165
Figure 6-9.	Response (Truncated) to a Request Using List Identifiers . . . . .	166
Figure 6-10.	Response to a Request Using List Records. . . . .	167
Figure 6-11.	Response to a Request Using List Sets. . . . .	168
Figure 6-12.	MarcEdit OAI-PMH Harvester. . . . .	171
Figure 6-13.	Generate Records in MarcEdit's MarcEditor . . . . .	172
Figure 6-14.	hCalendar Entry in an HTML/XHTML Document . . . . .	173
Figure 6-15.	COinS in the Browser . . . . .	176
Figure 6-16.	Code with COinS Object for a Book Embedded. . . . .	176
Figure 6-17.	Code with COinS Object . . . . .	177
Figure 6-18.	UNAPI Call . . . . .	177
Figure 7-1.	Federated Search Diagram . . . . .	183
Figure 7-2.	Hybrid Federated Search Diagram . . . . .	183

---

Figure 7-3.	OSI Model . . . . .	190
Figure 7-4.	SRU Explain Response . . . . .	194
Figure 7-5.	SRU Subject Query . . . . .	196
Figure 7-6.	Example of Plugin . . . . .	198
Figure 7-7.	OpenSearch Browser Integration . . . . .	201
Figure 7-8.	OpenURL Resolution Diagram . . . . .	203
Figure 7-9.	OpenURL Request for an Article . . . . .	204
Figure 7-10.	Modified OpenURL Diagram with DOI Resolution . . .	206
Figure 7-11.	Infrastructure with Registry Integration . . . . .	208
Figure 7-12.	Ockham Initiative User Interface . . . . .	209
Figure 7-13.	Service Node Example. . . . .	210
Figure 7-14.	Agent Node Example. . . . .	211
Figure 7-15.	Canary Database from Oregon State University IP. . . .	213
Figure 7-16.	Canary Database–Unknown IP. . . . .	214
Figure 8-1.	Microsoft/Creative Commons License Wizard . . . . .	225
Figure 8-2.	CONTENTdm Administration Module. . . . .	230
Figure 8-3.	DSpace Groups . . . . .	231
Figure 8-4.	LDIF Response Format . . . . .	234
Figure 8-5.	Oregon State University CONTENTdm Archive Collection Click Map . . . . .	237
Figure 9-1.	Traditional Library Service Model . . . . .	252
Figure 9-2.	Next-Generation Library Service Model. . . . .	253



# PREFACE

In the digital age, libraries need to preserve and provide access to digital resources, but traditional library procedures and tools from the brick-and-mortar type library are often not suited to this task. Physical libraries and their access mechanisms rely upon a publishing model that has slowly evolved over the past 500 years. In this model, each resource (book, globe, audiotape) consists of an object or objects in a single format, and each object remains static over time. Methods of building and cataloging physical library collections depend on these constants.

A digital library exists within a very different framework. A single resource (e.g., a portal) may consist of objects in many formats (full-text articles, databases, etc.), yet each of these objects is a resource in its own right. These objects may be updated frequently, and their original formats may become obsolete as technological developments lead to new types of information resources. Due to these differences, creating a digital library requires a new set of skills. *Building Digital Libraries: A How-To-Do-It Manual*<sup>®</sup> is a tool kit for the new world of digital libraries. It demystifies the challenges of designing, constructing, and maintaining a digital repository.

Although there have been many books written on electronic resources management, most have assumed that librarians will be managing subscription databases or collections created by others. *Building Digital Libraries* focuses on locally created digital repositories. A locally created digital repository is designed or maintained by the host institution. Such a collection may contain institutional or archival materials, or it may contain resources created partly or exclusively by parties outside the institution.

The few books that have focused on locally created repositories have been heavily weighted towards the interests of archivists. Archival collections may be the most obvious candidates for digitization, but digital libraries are far from just an archival concern. With the ease of finding general resources on the Web, digital collections have become one of the more unique and valuable parts of many libraries' holdings. In *Building Digital Libraries*, we have tried to speak to the full range of librarians who are, or who will be, involved in digital library projects: Systems librarians, project managers, and students, many of whom will find themselves starting, updating, or maintaining digital collections in years to come.

*Building Digital Libraries* covers both the fundamentals of digital library theory and the details of how to implement a digital collection. No

specific technical knowledge is required. Each chapter discusses the capabilities and limitations of specific technologies and reflects important developments of the last few years, with a focus on tools that are applicable and appropriate to a variety of environments. The most recent and useful third-party technologies are highlighted, including service repositories, Search/Retrieval URL and Search/Retrieval Web Service (SRU/SRW), federated searching, digital object identifiers (DOIs), and widespread adoption of OpenURLs.

After completing the book, readers will have sufficient knowledge to identify and implement the technical components necessary to construct a digital repository from scratch. Our aim is to explain and clarify both the technical and conceptual aspects of digital repositories so that readers can thoroughly understand how to create such a valuable resource for your library.

---

## ORGANIZATION

Each chapter in *Building Digital Libraries* focuses on a step in the process, addressing both how to execute that step and how to combat challenges encountered along the way.

Chapter 1, “Planning a Digital Repository,” gives a broad overview of issues surrounding the construction of a digital library. It provides the reader with an understanding of how the integrity of information can be protected over time, how to safeguard a repository against natural and man-made disasters, and how to accommodate the problem of constantly changing formats.

Chapter 2, “Acquiring, Processing, Classifying, and Describing Digital Content,” discusses specialized access mechanisms, processing and acquisitions, and maintenance; it also emphasizes the critical importance of good workflow. Chapter 3, “Choosing a Repository Architecture,” describes several frameworks for digital libraries and outlines the strengths and limitations of various hardware and software architectures. Choosing an appropriate hardware and software platform is critical to the success of a repository, so it is important to understand how the choice of platform influences how information can be stored and retrieved, which systems the collection can interact with, and how functionality can be enhanced in the future.

Chapter 4, “General Purpose Technologies Useful for Digital Repositories,” introduces metadata, particularly the group of technologies associated with eXtensible Markup Language (XML). Chapter 5, “Metadata Formats,” explores in greater detail generic technologies and critical standards, such as MARC, Dublin Core, Metadata Object Description Schema

(MODS), and Metadata Encoding and Transmission Standard (METS), providing examples that help the reader understand how these standards can be leveraged to provide services with relatively little effort.

As the number of information providers continues to grow, a repository cannot simply be a silo on the Internet. Chapter 6, “Sharing Data: Metadata Harvesting and Distribution,” looks at the role individual repositories play in a shared environment, how they can be normalized and shared for use by diverse systems, and how to make repositories searchable as part of federated collections and make their resources visible to search engines.

Chapter 7, “Federated Searching of Repositories,” investigates a wide array of protocols and technologies used for searching materials located in a vendor database or scattered across Web pages. From Z39.50—the original metasearch protocol for libraries—to the latest methods, readers will learn how to layer different search technologies to provide seamless access to diverse resources stored in different systems.

Chapter 8, “Access Management,” examines digital rights, protection of intellectual property rights, and monitoring of repository use in long-term repositories. Control mechanisms such as LDAP (Lightweight Directory Access Protocol) Shibboleth, OpenID, and Athens are also discussed.

Maintaining a repository is an ongoing endeavor. Chapter 9, “Planning for the Future,” is devoted to managing a living repository and anticipating future needs, as well as issues of updating as technologies and patron needs change.

Chapter 10, “Conclusions,” offers a clear outline of the process from start to finish and highlights the global importance of points touched upon in previous chapters.

It is not just the information itself, but the organization, structure, and presentation of that information, that give a repository its value. Digital libraries enhance the value of information resources by allowing users to locate information in contexts that suit their needs. We believe that these benefits of digitization have potential for a wide range of different types of collections and institutions. In *Building Digital Libraries*, we presume nothing except the desire to learn how to help bring libraries into the future.

