

# Putting Content Online: A Practical Guide for Libraries

Mark Jordan



CHANDOS PUBLISHING

# Putting Content Online: A Practical Guide for Libraries

**CHANDOS**  
**INFORMATION PROFESSIONAL SERIES**

Series Editor: Ruth Rikowski  
(email: Rikowskigr@aol.com)

Chandos' new series of books are aimed at the busy information professional. They have been specially commissioned to provide the reader with an authoritative view of current thinking. They are designed to provide easy-to-read and (most importantly) practical coverage of topics that are of interest to librarians and other information professionals. If you would like a full listing of current and forthcoming titles, please visit our web site [www.chandospublishing.com](http://www.chandospublishing.com) or contact Hannah Grace-Williams on email [info@chandospublishing.com](mailto:info@chandospublishing.com) or telephone number +44 (0) 1865 884447.

**New authors:** we are always pleased to receive ideas for new titles; if you would like to write a book for Chandos, please contact Dr Glyn Jones on email [gjones@chandospublishing.com](mailto:gjones@chandospublishing.com) or telephone number +44 (0) 1865 884447.

**Bulk orders:** some organisations buy a number of copies of our books. If you are interested in doing this, we would be pleased to discuss a discount. Please contact Hannah Grace-Williams on email [info@chandospublishing.com](mailto:info@chandospublishing.com) or telephone number +44 (0) 1865 884447.

# Putting Content Online: A Practical Guide for Libraries

**MARK JORDAN**



**Chandos Publishing**  
*Oxford · England*

Chandos Publishing (Oxford) Limited  
Chandos House  
5 & 6 Steadys Lane  
Stanton Harcourt  
Oxford OX29 5RL  
UK  
Tel: +44 (0) 1865 884447 Fax: +44 (0) 1865 884448  
Email: [info@chandospublishing.com](mailto:info@chandospublishing.com)  
**[www.chandospublishing.com](http://www.chandospublishing.com)**

---

First published in Great Britain in 2006

ISBN:  
1 84334 176 X (paperback)  
1 84334 177 8 (hardback)  
978 1 84334 176 5 (paperback)  
978 1 84334 177 2 (hardback)

© Mark Jordan, 2006

British Library Cataloguing-in-Publication Data.

A catalogue record for this book is available from the British Library.

All rights reserved. No part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording or otherwise) without the prior written permission of the Publishers. This publication may not be lent, resold, hired out or otherwise disposed of by way of trade in any form of binding or cover other than that in which it is published without the prior consent of the Publishers. Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The Publishers make no representation, express or implied, with regard to the accuracy of the information contained in this publication and cannot accept any legal responsibility or liability for any errors or omissions.

The material contained in this publication constitutes general guidelines only and does not represent to be advice on any particular matter. No reader or purchaser should act on the basis of material contained in this publication without first taking professional advice appropriate to their particular circumstances.

Typeset by Domex e-Data Pvt. Ltd.

Printed in the UK and USA.

---

# Contents

<i>Acknowledgements</i>	<i>xi</i>
<i>List of figures</i>	<i>xiii</i>
<i>List of tables</i>	<i>xv</i>
<i>About the author</i>	<i>xvii</i>
<b>1. Introduction</b>	<b>1</b>
This book's intended audiences	2
The big(ger) picture	3
A word on terminology	4
More on digital collections	6
Libraries as publishers	10
Some current trends	10
Keeping current	13
Conventions used in this book	15
Notes	16
<b>2. Preliminary tasks</b>	<b>17</b>
Rationales for digitising	17
Defining collection goals, scope and objectives	22
Evaluating and selecting source material	26
Cultural sensitivity and privacy issues	32
Collection policies	33
Summary: preliminary tasks	33
Notes	34
<b>3. Copyright and digital library collections</b>	<b>37</b>
National copyright law and digital collections	38

What libraries can put online	40
Managing permissions	44
Additional issues	46
Summary: making decisions about intellectual property	48
Further reading	48
Notes	50
<b>4. Metadata for digital collections</b>	<b>53</b>
Types of metadata	54
Selected major standards	65
Collection description	77
Subject access and authority control	78
Persistent identifiers	79
Interoperability	82
Native vs. derived metadata	84
Sources of metadata	86
Strategies for metadata creation and maintenance	87
Summary: making decisions about metadata	89
Further reading	89
Notes	90
<b>5. File formats</b>	<b>93</b>
Master vs. derivative versions	94
Open vs. proprietary formats	94
Formats for still images	96
Resolution, colour depth and compression	97
Formats for text	101
Formats for sound and video	106
Formats for data sets	109
Formats for complex documents	111
Factors in determining which formats to use	111

Summary: making decisions about file formats	117
Further reading	117
Notes	119
<b>6. Search and display</b>	<b>121</b>
Still images	122
Textual documents	128
Moving images and sound	134
Data sets	137
Mixed document collections	138
Planning your collection's interface	139
Summary: making decisions about search and display	147
Further reading	148
Notes	149
<b>7. Content management systems</b>	<b>151</b>
Types of CMSs	151
Examples of DLCMSs	153
Evaluating DLCMSs	159
Summary: making decisions about content management systems	173
Further reading	173
Notes	174
<b>8. Project management</b>	<b>177</b>
Sequence and timing	178
Planning the implementation	180
Additional planning considerations	194
Executing the project	199
Proposal writing	199
Reporting	203
Monitoring	204
Evaluating the production phase of the project	204

Evaluating the overall project	205
Evaluating the collection	207
Multi-institution projects	208
Summary: managing digital collection projects	209
Further reading	210
Notes	211
<b>9. Project operations</b>	<b>213</b>
A closer look at staffing	214
Documentation in project operations	219
Creating metadata	221
Quality control	224
General hardware and software used in project operations	228
Capturing and converting printed content	232
Capturing sound and video	238
Evaluating and acquiring specialised hardware and software	240
File management	246
Summary: making decisions about project operations	250
Further reading	250
Notes	251
<b>10 Developing workflows</b>	<b>255</b>
The workflow development cycle	257
Outlining techniques	259
Diagramming techniques	260
Selected workflows from the literature and the field	267
A general workflow modelling technique	269
Summary: developing workflows for digital collections	279
Further reading	280
Notes	281

<b>11. Preservation strategies</b>	<b>283</b>
The problems	283
Approaches to digital preservation	285
OAIS Reference Model	288
Trusted digital repositories	290
Preservation metadata	290
Preservation policies	293
Practical technologies	295
Practical things you can do	298
Summary: making decisions about digital preservation	301
Further reading	301
Notes	303
<b>12. A case study</b>	<b>305</b>
Description of the collection and the project	306
Defining the collection's goals	307
Evaluating the content and clearing permissions	308
Planning the project	310
Developing the workflows	319
Staffing	330
Documentation and procedures	331
Budgeting	331
Final preparations	337
Executing the project: doing the work	337
Evaluating the project	338
Notes	339
<b>Index</b>	<b>341</b>

This page intentionally left blank

---

## Acknowledgements

I would like to thank Lynn Copeland (University Librarian and Dean of Library Services) and Brian Owen (Associate University Librarian for Processing and Systems) at Simon Fraser University for granting me a study leave to complete this book during an extremely busy period at the Library, and for getting me involved in lots of interesting projects. I would also like to thank my colleagues in Library Systems for taking on extra work caused by my absence.

David Kisly (Target Library and Information Services) and Gwen Bird (Simon Fraser University) read drafts of each chapter as I completed them and offered frank and insightful comments. Nina Saklikar (Simon Fraser University) read the entire manuscript and supplied a number of characteristically sensible suggestions. Despite these three people's contributions I claim credit for all remaining errors and omissions.

I would also like to thank my students in 'Developing Digital Collections' at the University of British Columbia's School of Library, Archival and Information Studies for challenging me to explain many of the topics covered in the following pages. Acknowledgements are due also to the following people for sharing with me their expertise, advice and encouragement: Ian Song, Elaine Fairey, Peter Van Garderen, Joe Tennis and Kim Lawson.

Lastly, thanks to Michele for her support and for tolerating my obsessive working habits.

This page intentionally left blank

---

## List of figures

4.1	Simple METS fileSec metadata represented in XML	62
4.2	A view of Vancouver's Lions Gate Bridge	69
4.3	A simple description encoded in MARC	69
4.4	A simple MODS description	73–74
6.1	List and grid views of the same result set, as provided by the Streetprint Engine content management system. Used with permission of the CRC Humanities Computing Studio, University of Alberta.	124
6.2	Screen space filled by the same image on monitors of two different sizes: (a) 1024 × 768 resolution monitor; (b) 1280 × 1024 resolution monitor.	125
6.3	A double page image (above) from <i>The Latch Key of My Bookhouse</i> , and a magnified view (below), as viewed using the Internet Archives' DjVu applet. Used with permission of the Internet Archive.	127
6.4	A simple page turner from a collection of texts that provides one image for each page of text. Used with permission of Simon Fraser University Library.	132
6.5	A simplified diagram of how standard newspaper layout can pose challenges for navigating between parts of a single article.	133
6.6	A single newspaper article removed from its original context.	134
6.7	A simple storyboard illustrating the steps required to get from the search page to the display of an image.	144
9.1	An image with a border around the page edges	234
9.2	Example directory structure (in Mac OS X) for managing different versions of the same master files	248

10.1 A simple, hand-drawn block diagram (left) and a slightly more sophisticated version drawn using Microsoft Visio (right)	263
10.2 A flowchart representing a workflow for scanning still images	264
10.3 A UML activity diagram illustrating forks and branches	266
10.4 A UML activity diagram with swimlanes and agents	267
10.5 A UML activity diagram representing the workflow for the diary digitisation project	280

---

## List of tables

3.1	Summary comparison of term, fair dealing/use, and orphan works in UK, US and Canadian copyright law	39
4.1	The same METS fileSec metadata represented in a spreadsheet	62
4.2	Simple METS structMap metadata represented in a spreadsheet, showing the books and chapters in H. G. Wells' <i>War of the Worlds</i>	62
4.3	A simple description encoded in unqualified Dublin Core	71
4.4	A simple description encoded in qualified Dublin Core	71
4.5	Sample persistent identifiers	80
4.6	Mappings between MARC, MODS and unqualified Dublin Core	85
5.1	Effects of resolution, dimensions, colour depth and compression on image file size	98
5.2	Sample disk space requirements for image files	100
5.3	Summary of best practices for selecting file formats for text	106
5.4	Summary of best practices for selecting formats based on collection goals	114
6.1	A sample use case	142
7.1	Summary of the advantages and disadvantages of the three types of DLCMSs	162
8.1	Sample calculations for estimating extent of source material	182
8.2	Sample production milestones	182
8.3	Workflow tasks and average times for processing sample audio recordings	193
8.4	Estimated time required to digitise entire collection and create metadata	195

8.5	Tasks and salaries for the Web Developer and Supervisor	195
8.6	Total staff costs for processing the audio recordings	196
8.7	Hours of work per day required to complete all work	198
8.8	Positions required to complete all work in days available (FTE)	198
9.1	A sample item status board	216
9.2	A sample decision table	220
9.3	Typical capacities for the most common types of hard disk storage	229
9.4	Examples of processes and corresponding software used on still images and textual documents	235
9.5	Types of scanners used in digital library content production	236
9.6	Examples of processes and corresponding software used on sound and video files	239
9.7	Sample publicised versus tested scan times	244
9.8	Sample version file names using single-letter indicators to distinguish versions	247
10.1	The general workflow model template	271
10.2	General workflow model template with specific activities	274
11.1	Summary of metadata from PREMIS Data Dictionary example 6	292

---

## About the author

**Mark Jordan** is Head of Library Systems at Simon Fraser University in Burnaby, British Columbia, Canada. His areas of expertise include digitisation, data integration and managing the development of open-source library software. He has published articles on electronic journal management, systems librarianship, metadata and open-source development tools and has presented at numerous Canadian library and information technology conferences.

Mark is also an Adjunct faculty member at the University of British Columbia's School of Library, Archival and Information Studies, and is the maintainer of Digitizationblog (*<http://digitizationblog.interoperating.info>*).

The author may be contacted by email: [mjordan@sfu.ca](mailto:mjordan@sfu.ca)

This page intentionally left blank