How to Build a Digital Library

Second Edition

MK
MORGAN KAUFMANN

Ian H. Witten | David Bainbridge | David M. Nichols

Contents

Preface ..... *xv*

The Greenstone Software xvi
Updated and Revised Content xvii
How the Book Is Organized xviii
What the Book Covers xviii
About the Web Site xxi
Acknowledgments xxi

Part I Principles and Practices *1*

Chapter 1 Orientation: The world of digital libraries *3*

Example One: Supporting Human Development 3
Example Two: Pushing on the Frontiers of Science 4
Example Three: Preserving a Traditional Culture 5
Example Four: Exploring Popular Music 6
The scope of digital libraries 6
1.1 Libraries and Digital Libraries 7
1.2 The Changing Face of Libraries 9
In the beginning 11
The information explosion 12
The Alexandrian principle 13
Early technodreams 15
The library catalog 16
The changing nature of books 17
1.3 Searching for Sophocles 20
1.4 Digital Libraries in Developing Countries 25
Disseminating humanitarian information 26
Disaster relief 26
Preserving indigenous culture 27
Locally produced information 27
The technological infrastructure 28
1.5 The Pen Is Mighty: Wield It Wisely 29
Copyright law 29
The public domain 30

Relinquishing copyright 32
Digital rights management 33
Copyright and digitization 34
Collecting from the Web 35
Illegal and harmful material 37
Cultural sensitivity 38
1.6 Planning a Digital Library 38
1.7 Implementing a Digital Library: The Greenstone Software 41
1.8 Notes and Sources 41

Chapter 2: People in digital libraries 47
2.1 Roles 49
Global users 50
Roles of librarians 51
Change 52
2.2 Identity 54
Anonymous use 54
Authenticated use 56
Recording usage data 57
2.3 Help and User Support Services 61
2.4 Working with Digital Collections 63
Using information from digital libraries 64
Referring to objects in a digital library 65
Berry-picking 65
2.5 User Contributions 67
Annotations 67
Keywords 67
Ratings 68
Corrections 68
New documents 68
Partial and fluid documents 68
2.6 Notes and Sources 70

Chapter 3: Presentation: User interfaces 73
From People to Presentation 73
3.1 Presenting Textual Documents 74
Documents, chapters, sections 74
Unstructured text documents 76
Page images 79
Images with text 81
Realistic books 84
3.2 Presenting Multimedia Documents 86
Sound and pictures 86
Video 88
Music 88

3.3 Document Surrogates 90
Metadata 90
Multimedia surrogates 93
3.4 Searching 93
Types of queries 95
Case-folding and stemming 98
Phrase searching 100
Query interfaces 102
Searching multimedia 104
3.5 Metadata Browsing 110
Lists 111
Dates 113
Hierarchies 114
Facets 114
3.6 Putting It All Together 116
An institutional repository 116
3.7 Notes and Sources 123

Chapter 4: Textual documents: The raw material 127

4.1 Representing Textual Documents 130
ASCII 130
Unicode 132
Plain text 133
Indexing 134
Word segmentation 137
4.2 Textual Images 137
Scanning 139
Optical character recognition 140
Page handling 146
Planning an image digitization project 147
Inside an OCR shop 148
An example project 149
4.3 Web Documents: HTML and XML 152
Markup and stylesheet languages 153
Basic HTML 155
Using HTML in a digital library 158
Basic XML 159
Parsing XML 162
Using XML in a digital library 162
4.4 Presenting Web Documents: CSS and XSL 163
CSS 163
Extensible stylesheet language 170
4.5 Page Description Languages: PostScript and PDF 177
PostScript fundamentals 177
Fonts 182

Text extraction 185
Using PostScript in a digital library 189
Portable Document Format: PDF 190
PDF and PostScript 195
4.6 Word-Processor Documents 195
Rich Text Format: RTF 197
Native Word formats 202
Office Open XML: OOXML 203
Open Document format: ODF 204
Scientific documents: LaTeX 207
4.7 Other Documents 210
Spreadsheets and presentation files 210
E-mail 210
4.8 Notes and Sources 211

Chapter 5: Multimedia: More raw material 215

5.1 Introducing Compression and Transforms 216
Basic compression techniques 217
Transforms 219
The Fourier transform 219
5.2 Audio 221
Pulse code modulation: PCM 222
Variants of PCM 224
Early formats: WAV, AIFF, AU 226
MPEG audio: MP3 and its siblings 228
Post-MP3 formats: AAC, Ogg Vorbis, FLAC 229
Replaying audio 231
An audio digital library 231
5.3 Images 235
Lossless compression: GIF and PNG 236
Lossy compression: JPEG 237
Progressive refinement 242
Archiving images: JPEG 2000 and TIFF 245
A digital library of photographs 248
Vector graphics images 252
5.4 Video 258
Codecs 258
Multimedia compression: MPEG 259
High Definition Digital Television 264
Proprietary formats 264
Streaming 266
Ogg Theora 266
Using multimedia in a digital library 267
A video digital library 268
Reflection 268

5.5 Rich Media 271
Synchronized Multimedia Integration Language: SMIL 271
Adobe Flash 275
5.6 Music 277
Musical Instrument Digital Interface: MIDI 278
Digital music libraries 279
5.7 Notes and Sources 282
Audio 282
Images 283
Video 283
Rich Media 284
Music 284

Chapter 6: Metadata: Elements of organization 285

6.1 Characteristics of Metadata 286
6.2 Bibliographic Metadata 288
MARC 289
MARCXML 293
Dublin Core: DC 294
Qualified Dublin Core 295
Metadata Object Description Schema: MODS 297
BibTeX 297
EndNote 298
6.3 Metadata for Multimedia 299
Image metadata: TIFF 300
Image metadata: EXIF, XMP, IPTC, and MIX 302
Audio metadata 304
Video metadata 306
Multimedia metadata: MPEG-7 307
Multimedia application metadata: MPEG-21 309
6.4 Metadata for Compound Objects 310
Resource Description Framework: RDF 310
Metadata Encoding and Transmission Standard: METS 313
Collection-level metadata 316
Open Archives Initiative Object Reuse and Exchange: OAI-ORE 319
Metadata for education: LOM and SCORM 319
Metadata for eResearch 321
6.5 Metadata Quality 323
Authority control: Names 324
Authority control: Subjects 327
Controlling metadata values 329
Metadata tools 330
6.6 Extracting Metadata 330
Extracting document metadata 332
Generic entity extraction 332

Bibliographic references 334
Language identification 334
Acronym extraction 335
Key-phrase metadata 336
6.7 Notes and Sources 339

Chapter 7: Interoperability: Protocols and services 343
7.1 Z39.50 Protocol 344
7.2 Open Archives Initiative 345
OAI Protocol for Metadata Harvesting: OAI-PMH 346
Serving 348
Harvesting 350
7.3 Object Identification 350
Handles 351
Digital object identifiers: DOIs 352
OpenURLs 353
Persistence 353
7.4 Web Services 354
Search/Retrieval via URL: SRU 357
7.5 Authentication and Security 359
7.6 DSpace and Fedora 361
DSpace 361
Fedora 364
7.7 Notes and Sources 369

Chapter 8: Internationalization: The global challenge 371
8.1 Multilingual Interfaces and Documents 372
8.2 Unicode 375
Composite and combining characters 381
Unicode character encodings 384
Using Unicode in a digital library 387
8.3 Hindi and Indic Scripts 389
ISCII: Indian Script Code for Information Interchange 389
Unicode for Indic scripts 390
Problems with the adoption of Unicode 392
8.4 Word Segmentation and Sorting 394
Segmenting words 394
Sorting Chinese text 396
8.5 Notes and Sources 398

Chapter 9: Visions: Future, past, and present 401
9.1 Libraries of the Future 402
Today's visions 402
Tomorrow's visions 404
Working inside the digital library 407

9.2 Preserving the Past 408
The problem of preservation 410
A sorry tale 411
Preservation strategies 415
9.3 Trends in Digital Libraries 420
Mobility: Portable collections 420
Knowledge-based information retrieval 424
9.4 Digital Libraries for Oral Cultures 427
9.5 Notes and Sources 429

Part II Greenstone Digital Library Software 433

Chapter 10: Building collections 435
10.1 The Reader's Interface 437
The Greenstone digital library 437
Exploring the Demo collection 438
Browsing 438
Searching 440
Preferences 441
10.2 The Librarian Interface 442
Users and functions 442
A walk-through 443
10.3 Working with Documents 454
HTML documents 454
Word and PDF files 456
Enhanced Word document handling 458
Enhanced PDF document handling 461
Enhanced HTML document handling 464
Scaling up 466
10.4 Formatting 469
The Format panel 469
Format Features 470
Default format statements 472
Format strings 473
Formatting exercise 1: Tudor collection 476
Formatting exercise 2: Word and PDF collection 481
Formatting exercise 3: Branding your collection 483
10.5 Dealing with Metadata 485
The Enrich panel 486
How metadata is stored 488
Collections of bibliographic information 490
Working with individual metadata records 491
Combining metadata and source documents 494
10.6 Non-Textual Documents 495
Images 495

Textual images 497
Multimedia 504
10.7 Learning More 509
Sources of information 509
The user community 511
When things go wrong ... .511

Chapter 11: Operating and interoperating 513

11.1 Inside Greenstone 514
Updating the software 514
Files and folders 515
Collections 517
Greenstone CD-ROM/DVDs 518
11.2 Operational Aspects 519
Configuration files 519
Logging 520
Administration facility 521
Authentication 521
Protecting a collection 522
11.3 Command-Line Operation 524
Getting started 524
Making a framework 525
Importing documents 526
Building indexes 528
Installing the collection 528
11.4 Under the Hood 529
Importing and building 529
Incremental building 529
Scheduled rebuilding 530
Archive formats 531
Document identifiers 533
Plug-ins 534
Search indexes 536
11.5 Interoperating 539
Downloading Web sites 539
Metadata protocols 540
Serving OAI 541
Exporting collections 543
Interoperating with DSpace 543
11.6 Distributed Operation 545
Remote Librarian interface 545
Institutional repositories 549
11.7 Large-Scale Usage 554
Limitations of the Librarian interface 554
Large collections 554

A very large collection 555
Distributed serving 558

Chapter 12: Design patterns for advanced user interfaces 559

12.1 Format Statements and Macros 560
Format statements 561
Macros 563
Commonly used macros 565
12.2 Design Patterns 567
Design pattern 1: Additional static pages 567
Design pattern 2: Using JavaScript to adjust presentation 569
Design pattern 3: Making format statements reusable through macro definitions 571
Design pattern 4: Dynamic HTML 573
Design pattern 5: Exploiting Asynchronous JavaScript and XML (AJAX) 578
12.3 The Greenstone Research Project 585
Research with Greenstone3 585
Reconciling research and production values 586
Closing words 587

Glossary 589

References 597

Index 607

Preface

On the top floor of the Tate Modern Art Gallery in London is a meeting room with a magnificent view over the River Thames and down into the open circle of the reconstructed Globe Theatre nearby. Here, at a gathering of senior administrators who fund digital library projects internationally, one of the authors stood up to introduce himself and ended by announcing that he was writing a book entitled *How to Build a Digital Library*. As he sat down, his neighbor nudged him and asked with a grin, "A work of fiction, eh?" A few weeks earlier and half a world away, the same author was giving a presentation about a digital library software system at an international digital library conference in Virginia, when a colleague in the audience noticed someone in the next row who, instead of paying attention to the talk, downloaded that very software over a wireless link, installed it on his laptop, checked the documentation, and built a digital library collection of his e-mail files—all within the presentation's 20-minute time slot.

These little cameos illustrate the extremes. Digital libraries?—colossal investments, which like today's national libraries will grow over decades and centuries, daunting in complexity. Conversely: digital libraries?—off-the-shelf technology; just add documents and stir. Of course, we are talking about very different things: a personal library of ephemeral notes hardly compares with a national treasure-house of information. But don't sneer at the "library" of e-mail: this collection gives its user valued searching and browsing facilities, and with half a week's (rather than half an hour's) work, one could create a prototype document-management system that organizes documents for a large multinational corporation.

Digital libraries are organized collections of information. Our experience of the World Wide Web—vibrant yet haphazard, uncontrolled and uncontrollable—daily reinforces the impotence of information without organization. Likewise, experience of using online public-access library catalogs from the desktop—impeccably but stiffly organized, and distressingly remote from the actual documents themselves—reinforces the frustrations engendered by organizations without fingertip-accessible information. Can we not have it both ways? Enter digital libraries.

Whereas physical libraries have been around for 25 centuries, digital libraries span only 15 years. Yet, in today's information society, with its Siamese twin the knowledge economy, digital libraries will surely figure among the most important and influential institutions of the new century. The information revolution not only supplies the technological horsepower that drives digital libraries, but also fuels an unprecedented demand for storing, organizing, and accessing information. If information is the currency of the knowledge economy, digital libraries are the banks where it is invested.

We do not believe that digital libraries are supplanting existing bricks-and-mortar libraries—not in the near- and medium-term future that this book is about. And we certainly don't think you should be burning your books in favor of flat-panel displays! Digital libraries are new tools for achieving human goals by changing the way that information is used in the world. We are talking about new ways of dealing with knowledge, not about replacing existing institutions.

What is a digital library? What does it look like? Where does the information come from? How do you put it together? Where to start? The aim of this book is to answer these questions in a plain and straightforward manner, with a strong practical "how to" flavor.

We define digital libraries as

> focused collections of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance.

To keep things concrete, we show examples of digital library collections in an eclectic range of areas, with an emphasis on cultural, historical, humanitarian, and musical applications, as well as technical ones. These collections are formed from different kinds of material, organized in different ways, presented in different media and different languages. We think they will help you see how digital libraries can be applied to real problems. Then we show you how to build your own.

The Greenstone Software

The Greenstone Digital Library Software is a comprehensive software resource that illustrates the ideas in the book and could form a basis for your own digital library. It is freely available as open source software on the World Wide Web (at www.greenstone.org) and comes precompiled for all popular platforms. A fully operational, flexible, extensible system for constructing easy-to-use digital libraries, Greenstone is widely deployed internationally and is being used (for example) by United Nations agencies and related organizations to deliver humanitarian information in developing countries. The ability to build new digital library collections, particularly in developing countries, is promoted by UNESCO's *Information for All Programme*. Through this initiative, the intergovernmental agency provides support for Greenstone in the form of testing, translating, and distributing the software.

In this second edition of *How to Build a Digital Library* we have decoupled Greenstone from the more general material in the book by packing everything related to the software into a separate Part II, which now serves as a comprehensive tutorial guide to Greenstone. All the material in Part I has broad application and is not tied to any particular software infrastructure for digital libraries.

Since the first edition was published, Greenstone has acquired a new interactive interface for librarians that makes it far easier to build and serve collections. In addition, the software has grown enormously and now includes an unparalleled range of facilities for textual and multimedia documents, handling and extracting metadata in various formats, configuring collections, and interoperation with other standards and protocols. In fact, almost everything described in Part I can be accomplished within the Greenstone software: it is a complete industrial-strength implementation of essentially all the techniques covered in this book.

Updated and Revised Content

We finished writing the first edition of this book in late 2002 and now, in August 2009, are just polishing this second edition. The field of digital libraries has changed radically in the intervening years. Although much of the core material remains the same, we have made the most of our opportunity to update it to reflect the changes that have taken place over seven years. We have thoroughly revised and edited everything: in addition, we have a new co-author, David Nichols. The most enjoyable part has been adding new material—here are the highlights.

Responding to popular demand, we include a new chapter on people in digital libraries. We entered the digital library field with a perspective that was strongly colored by our technical background in computer science. However, from our experience in giving scores of Greenstone courses internationally and interacting with the user base on the mailing list, we now have extensive experience of the way digital libraries operate. Chapter 2 introduces the roles that people play in digital libraries and discusses issues of identity and anonymity, help and support services, individual usage and group collaboration, and the growing area of user contributions to digital collections.

Chapter 5 on multimedia is also completely new. The pace and scope of change in global multimedia are staggering. YouTube, iTunes, and Flickr are just a few examples of how people around the world create and share material in newly connected and digitally complex ways. Low-cost portable devices are helping to shape a dynamic multimedia-driven approach to communications, arts, business, and research. Multimedia content management presents cross-disciplinary challenges in integrating and organizing data from a plethora of media types that, to date, have remained stratified (e.g., images of various kinds, music and songs, commentary, video, and text in various guises and formats). Few technologies today allow high-quality, user-friendly, and graceful mixing of multiple media types into rich collections of accessible information. This is a gap that digital libraries must fill.

Chapter 6 on metadata includes extensive new material on metadata for audio, video, multimedia, and compound objects, as well as a new discussion of metadata quality.

Internationalization is another burgeoning area. The first edition of *How to Build a Digital Library* contained plenty of information on the subject, but recognizing that some readers with an English-only perspective may find such material unnecessary and distracting, we have consolidated it into a single new chapter, Chapter 8. Greenstone itself is fully internationalized, with interfaces in fifty different languages.

There are countless other changes in the book that reflect the way the field is developing. We have updated many of the examples to reflect today's larger and more comprehensive digital collections. We include more material on planning a digital library, document surrogates, and faceted browsing. Many new formats are described, including the Open Document Format for Office Applications and Microsoft's Office Open XML, as well as other document types, such as e-mail, spreadsheets, and presentations. We have focused our metadata chapter on external metadata and separated the issue of markup (internal metadata); we also reduced the level of detail in which the HTML format is described. This edition includes much new information on interoperability, and separates it from the issue of Greenstone support—which, as stated above, is consigned to Part II.

How the Book Is Organized

The gulf between the general and the particular has presented interesting challenges in organizing this book. As the title says, our aim is to show you how to build a digital library, and we really do want you to build your own collections (it doesn't have to take long, as the conference attendee mentioned in the first paragraph discovered). But to work within a proper context, you need to learn something about libraries and information organization in general. And if your practical work is to proceed beyond a proof-of-concept prototype, you will need to come to grips with countless nitty-gritty details.

We have tried to present what you need to know in a logical sequence, introducing new ideas where they belong and developing them fully at that point. However, we also want the chapters to function as independent entities that can be read in different ways. We are well aware that books like this are seldom read through from cover to cover! The result is, inevitably, that some topics are scattered throughout the book.

We cover three different themes: the intellectual challenges of libraries and digital libraries, the practical standards involved in representing documents digitally, and how to use Greenstone to build your own collections. Many academic readers will want a textbook, some a general text on digital libraries, others a book with a strong practical component that supports student projects.

For a general introduction to digital libraries, read Chapters 1 and 2 to learn about libraries and library organization, then Chapter 3 to find out about what digital libraries look like from a user's point of view, and then skip straight to Chapter 9 to see what the future holds.

To learn about the ways that documents are represented digitally, skim Chapter 1, read Chapters 4, 6, and 7 to learn about the standards, and then look at Chapter 3 to see how they can be used to support interfaces for searching and browsing. If you are interested in multimedia, read Chapter 5 as well—or instead, because it largely stands alone.

To learn how to build a digital library as quickly as possible, skim Chapter 1 (but check Sections 1.6 and 1.7) and turn straight to Part II. If you run into things you need to know about library organization, different kinds of interfaces, document formats, or metadata formats, you can return to the intervening material.

For a textbook on digital libraries without any commitment to specific software, use Part I of the book in sequence. For a course with a strong practical component, read all chapters—and, in parallel, turn your students loose on Part II!

What the Book Covers

We open with four scenarios intended to dispel any ideas that digital libraries are no more than a routine development of traditional libraries with bytes instead of books. Then we discuss the concept of a *digital library* and set it in the historical context of library evolution over the ages. We go on to exemplify features of a large-scale, real-world digital library by looking at a usage scenario. One thread that runs through the book is internationalization and the role of digital libraries in developing

countries—for whom we believe that digital libraries represent a "killer app" of computer technology. There follows a discussion of issues involved in copyright and "harvesting" material from the Web. Finally, we close by discussing the planning of a digital library and by briefly introducing the Greenstone software, which is fully described in Part II.

Chapter 2 is about the people in digital libraries. As noted above, it covers the many roles played by people and discusses issues of identity and anonymity, help and support services, individual usage and group collaboration, and the growing area of user contributions to digital collections—and what all this implies in terms of software support.

As the definition of *digital library* given earlier implies, digital libraries involve two communities: end users who are interested in access and retrieval, and librarians who select, organize, and maintain information collections. Chapter 3 takes the user's point of view. Of course, digital libraries would be a complete failure if you had to study a book in order to learn how to use them—they are supposed to be easy to use!—and this book is really directed at the library builder, not the library user. Nevertheless, it is useful to survey what different digital libraries look like. Examples are taken from domains ranging from human development to culture, with audiences ranging from children to library professionals. Document contents range from text to newspaper images, and multimedia material ranges from musical query-by-humming to browsing pictures according to their content. (International examples are reserved for Chapter 8.) We show many examples of browsing structures, from simple lists to hierarchies, date displays, and faceted structures, and close with a description of the use of a popular institutional repository system.

Next we turn to documents, the digital library's raw material. Chapter 4 begins with character representation, in particular Unicode, which is a way of representing all the characters used in all the world's languages (although again international aspects are covered in Chapter 8). Plain text formats introduce some issues that you need to know about. Here we take the opportunity to describe full-text indexing, the basic technology for searching text, and we also introduce the issue of word segmentation. We then describe the process of optical character recognition (OCR), including typical costs and an example OCR project. Next we look at documents on the Web: HTML and XML, including style sheets and the presentation of Web documents. Next we study popular formats for document representation, beginning with the page description languages PostScript and PDF (Portable Document Format) and continuing with the word-processor formats RTF (Rich Text Format), the Open Document Format for Office Applications, and LaTeX, which is commonly used for mathematical and scientific documents. Finally, we look at other document types: e-mail, spreadsheets, and presentation files.

Chapter 5 gives a comprehensive account of multimedia, beginning with a brief introduction to compression and transforms. Then we describe audio, image, and video formats; we go on to introduce so-called "rich media," a term designed to emphasize interaction with multimedia composed of potentially different types of media; and we end with a discussion of music and digital music libraries. A plethora of coding techniques and formats are covered: PCM, WAV, AIFF, AU, MP3, AAC, Ogg Vorbis, and FLAC for audio; GIF, PNG, JPEG, JPEG 2000, and TIFF for images; MPEG-1, MPEG-2, MPEG-4, and Ogg Theora for video, along with a discussion of streaming and proprietary formats; SMIL and Adobe Flash for rich media; and MIDI for music.

Besides textual and multimedia documents, there is another kind of raw material for digital libraries: metadata. Often characterized as "data about data," metadata figures prominently in this book because it forms the basis for organizing both digital and traditional libraries. Chapter 6 covers metadata and explains how it is expressed in traditional library catalogs and in digital libraries. Like the previous two chapters, Chapter 6 covers many different formats, including MARC, MARCXML, Dublin Core, qualified Dublin Core, MODS, BibTeX, and EndNote for documents; TIFF, EXIF, XMP, IPTC, and MIX for images; MPEG-7 and MPEG-21 for multimedia; RDF, METS, OAI-ORE, LOM, and SCORM for compound objects. We introduce issues of metadata quality, and the idea of extracting metadata from the raw text of the documents themselves, giving examples of what can be extracted.

Chapter 7 reaches out to look at other standards and protocols that allow digital libraries to interoperate with one another and with related technologies. We describe the Z39.50 protocol used by current library automation systems, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), methods for persistent object identification, including various kinds of digital object identifier (Handles, DOIs, and OpenURLs), Web services, and the Search/Retrieve URL Service (SRU). We also discuss different standard ways of authenticating users: LDAP, OpenID, and Shibboleth. We close by looking at two prominent open source digital library software systems: DSpace and Fedora.

Chapter 8 also reaches out, this time internationally. We begin with some examples of multilingual interfaces to digital libraries and examples of collections of documents in different languages. Then we give a comprehensive account of the Unicode standard for representing the characters used in all the world's languages. Hindi and Indic scripts present interesting problems of character coding that have not yet been entirely solved in Unicode-compliant applications, and we give a glimpse of the complexities involved. Chinese and some other Asian languages involve other issues because they are written without word spaces. Also, "alphabetical order" is moot in non-alphabetic ideographic languages, and this has implications for browsing digital libraries.

Part I closes with visions of the future of digital libraries and mentions some important related topics that we have not been able to develop fully. We hope that this book will help you learn the strengths and pitfalls of digital libraries, gain an understanding of the principles behind the practical organization of information, and come to grips with the tradeoffs that arise when implementing digital libraries.

Part II is about the Greenstone Digital Library Software. Switching to a tutorial style of presentation, we present exercises that develop collections that demonstrate particular capabilities discussed in Part I. Chapter 10 commences by building a diverse range of digital library collections using the Librarian Interface, collections that include a representative selection of document and multimedia types from Chapters 4 and 5: documents in Word, HTML, and other formats; photographic and facsimile images; audio and video. They also utilize several bibliographic formats from Chapter 6. In some cases, bibliography files constitute digital library documents in their own right; in others, they are combined with the source documents that the bibliographic entries describe.

Chapter 11 describes the internal structure of Greenstone—the files and folders that support its operation—and shows how collections can be built outside the Librarian Interface using command-

line scripts. This provides greater control and is often appropriate for very large collections. The chapter also presents practical examples of interoperability, again mirroring the topics covered in Part I's Chapter 7, and it closes with a discussion of very-large-scale usage of Greenstone. Finally, Chapter 12 explores the possibilities afforded by Greenstone's presentation layer in more depth to illustrate advanced user interface techniques.

We hate acronyms and shun them wherever possible—but in this area you just can't escape them. A glossary of terms is included near the end of the book to help you through the swamp.

Having worked your way through Parts I and II, you will be well prepared to develop large-scale production-level digital library systems like those illustrated in the book. The rest is up to you. Our aim will have been achieved if you actually *build a digital library*!

About the Web Site

A great deal of supplementary material is available on the *How to Build a Digital Library* Web site at www.greenstone.org/howto. There you will find a novel full-text index of the book at the sentence level in which you can locate sentences containing any word combination and find their page numbers in the printed book. You can view all the figures in full color and browse a list of acronyms, a hierarchical structure of phrases, and a collage of the images. You can also download machine-readable versions of all the XML examples, and sample files for the exercises in Part II. The Web site also contains an appendix that gives more information on markup and XML.

Acknowledgments

The best part of writing a book is reflecting on all the help you have had from your friends. This book is the outcome of a long-term research and development effort at the University of Waikato—the New Zealand Digital Library Project. Without the Greenstone software, the book would not exist, and we begin by thanking Rodger McNab, who charted our course years ago by making the major design decisions that underlie Greenstone. Rodger has long departed from our group, but the influence of his foresight remains—a legacy that this book exploits. Next comes Stefan Boddie, the man who kept Greenstone going for many years, navigating the shoals with a calm and steady hand on the tiller. Craig Nevill-Manning had the original inspiration for the expedition: he showed us what could be done, and left us to it. Today the ship is capably steered by Kathy Don, Oran Fry, Anna Huang, Anu Krishnan, and Max Rouast.

Every crew member, past and present, has helped with this book, and we thank them all. Most will have to remain anonymous, but we must mention a few striking contributions (in no particular order). Te Taka Keegan and Mark Apperley undertook the Māori Newspaper project described at the end of Section 4.2 and illustrated in Figure 8.1. Through Te Taka's efforts we receive inspiration every day from the magnificent *toki* that resides in our laboratory and can be seen in Figure 1.11, a gift from the Māori people of New Zealand that symbolizes our practical approach to building digital libraries. Lloyd Smith (along with Rodger and Craig) did the groundwork for the music collections that are illustrated here. Steve Jones builds many novel user interfaces, especially ones involving phrase

browsing. Sally Jo Cunningham is the resident expert on library organization and related matters. Stuart Yeates designed and built the acronym extraction module, while Dana McKay worked on such things as extracting date metadata, as well as drafting the Greenstone manuals. YingYing Wen was our chief source of information on the Chinese language and culture, while Malika Mahoui took care of the Arabic side. Matt Jones from time to time provided us with sage and well-founded advice. Kathy Don helped us get many technical details straight, and Anna Huang helped enormously with the figures.

Many others in the digital library lab at Waikato have made heroic technical contributions to Greenstone. John Thompson designed and implemented the original Librarian Interface. Michael Dewsnip made countless improvements to the software. Gordon Paynter built the phrase browsing interface, helped design the Greenstone communication protocol, and improved many aspects of metadata handling. Hong Chen, Trent Mankelow, John McPherson, David Milne, Todd Reed, Shaoqun Wu, and Xiaofeng Yu have all worked to improve the software. Geoff Holmes and Bill Rogers helped us with some very nasty low-level Windows problems. Eibe Frank and Olena Medelyan worked on key-phrase extraction, while Veronica Liesaputra built the realistic book illustrated in Figure 3.5. Annette Falconer worked on a Women's History collection that opened up new avenues of research. Rob Akscyn has been a continual source of inspiration, and his wonderful metaphors enliven this book. There are many others: we thank them all.

Tucked away as we are in a remote (but very pretty) corner of the Southern Hemisphere, visitors to our department play a crucial role: they act as sounding boards and help us develop our thinking in diverse ways. Some deserve special mention. George Buchanan has been a frequent visitor from the United Kingdom; he helped develop the OAI-PMH server and built the CD-ROM writing module, and he continues to work with our team. Elke Duncker, also from the United Kingdom, advised us on cultural and ethical issues, while Stefan Rüger helped with multimedia digital libraries. The influence of Carl Gutwin from Saskatchewan is particularly visible in the phrase browsing and key-phrase extraction areas; Wendy Osborn from Alberta developed the scheduled rebuilding capability. Gary Marsden from Cape Town also made significant contributions. Dan Camarzan, Manuel Ursu, and their team of collaborators in Brasov, Romania, have worked hard to improve Greenstone and put it into the field. Alistair Moffat from Melbourne, Australia, along with many of his associates, was responsible for MG, the full-text searching component, and he and Tim Bell of Christchurch, New Zealand, have been instrumental in helping us develop the ideas expressed in this book.

Special thanks are due to Michel Loots of Human Info in Antwerp and John Rose, formerly of UNESCO in Paris. Informed by their great wealth of experience, both have encouraged, cajoled (and occasionally bullied) us into making our software available in a form designed to be most useful to people in developing countries. We are particularly grateful to them for opening up this new world to us; it has given us immense personal satisfaction and the knowledge that our technological efforts are materially helping people in need. John now works for our project and is instrumental in our outreach activities in developing countries.

We acknowledge the support of Maria Trujillo of Colombia, and Chico Fernandez-Perez of the FAO in Rome. Until he was so sadly and unexpectedly snatched away from us, we derived great benefit from the boundless enthusiasm of Ferrers Clark at the Canadian national science and technology

library. We have learned much from conversations with Dieter Fellner of Braunschweig, particularly with respect to generalized documents, and from Richard Wright at the BBC in London. Last but by no means least, Harold Thimbleby in the United Kingdom has been a constant source of moral support.

We would like to acknowledge all who have translated the Greenstone interface—at the time of writing we have interfaces in 50 different languages. We are grateful to Jojan Varghese and his team from Vergis Electronic Publishing, Mumbai, India, for taking the time to explain the intricacies of Hindi and related scripts. We also thank everyone who has contributed to the packages included in the Greenstone distribution. And we thank all our Greenstone users, from whom we have learned so much.

The Department of Computer Science at the University of Waikato has supported us generously in all sorts of ways, and we owe a particular debt of gratitude to Mark Apperley for his enlightened leadership, warm encouragement, and financial help. In the early days we were funded by the New Zealand Lotteries Board and the New Zealand Foundation for Research, Science and Technology, which got the project off the ground. We have also received support from the Ministry of Education, while the Royal Society of New Zealand Marsden Fund has supported closely related work on text mining and computer music. The Alexander Turnbull Library has given us access to source material for the Māori Niupepa project, along with highly valued encouragement.

We thank Rick Adams and Heather Scherer of Morgan Kaufmann, and we gratefully acknowledge the efforts of the reviewers Bonnie Wilson and J. Stephen Downie, who carefully read the manuscript and made pertinent and constructive comments that helped us improve this book significantly.

Much of this book was written in people's homes while the authors were traveling around the world. We visited an extraordinary variety of delightful little villages—Killinchy in Ireland, Great Bookham and Welwyn North in England, Pampelonne in France, Mascherode in Germany, Canmore in Canada—as well as cities such as London, Siena, Heidelberg, Paris, Tokyo, Calgary, New Orleans, Austin, and San Francisco. To our hosts: you all know who you are—thanks! Numerous institutions helped with facilities, including Middlesex University in London, Braunschweig Technical University and the European Media Lab in Germany, the University of Calgary in Canada, the University of Texas at Austin, the University of North Carolina at Chapel Hill, the Payson Center for International Development and Technology Transfer in New Orleans, and the University of Siena in Italy. The generous hospitality of Google during a two-month stay is gratefully acknowledged: this proved to be a very stimulating environment in which to think about large-scale digital libraries.

All our traveling has helped spin the threads of internationalization and human development that are woven into the pages that follow. Our families—Annette, Pam, Anna, Elizabeth, Natasha, and Nikki—have supported us in countless ways, sometimes journeying with us, sometimes keeping the fire burning at home in New Zealand. They have had to live with this book, and we are deeply grateful for their sustained support, encouragement, and love.