

| | |
|--|-------|
| <i>Preface</i> | 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 |
| <i>Part I Principles and Practices</i> | 1 |
| <i>Chapter 1 Orientation: The world of digital libraries</i> | 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 |
| <i>Chapter 4: Textual documents: The raw material.....</i> | | <i>127</i> |
| 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 |
| <i>Chapter 5: Multimedia: More raw material.....</i> | <i>215</i> |
| 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 |
| <i>Chapter 6: Metadata: Elements of organization.....</i> | | <i>285</i> |
| 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 |
| <i>Chapter 7: Interoperability: Protocols and services.....</i> | <i>343</i> |
| 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 |
| <i>Chapter 8: Internationalization: The global challenge.....</i> | <i>371</i> |
| 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 Indie Scripts..... | 389 |
| ISCII: Indian Script Code for Information Interchange..... | 389 |
| Unicode for Indie 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 |
| <i>Chapter 9: Visions: Future, past, and present.....</i> | <i>401</i> |
| 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 |
| <i>Part II Greenstone Digital Library Software</i> | 433 |
| <i>Chapter 10: Building collections</i> | 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 |
| <i>Chapter 11: Operating and interoperating</i> | | 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 |
| <i>Chapter 12: Design patterns for advanced user interfaces</i> | 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 |
| <i>Glossary</i> | 589 |
| <i>References</i> | 597 |
| <i>Index</i> | 607 |