

Contents

List of figures	xiii
List of tables	xix
Foreword	
by Edward A. Fox	xxi
Preface	xxv

1. Orientation: The world of digital libraries 1

Example One: Supporting human development	1
Example Two: Pushing on the frontiers of science	2
Example Three: Preserving a traditional culture	3
Example Four: Exploring popular music	4
The scope of digital libraries	5
1.1 Libraries and digital libraries	5
1.2 The changing face of libraries	8
In the beginning	10
The information explosion	11
The Alexandrian principle	14
Early technodreams	15
The library catalog	16
The changing nature of books	17

- 1.3 **Digital libraries in developing countries** 20
 - Disseminating humanitarian information 21
 - Disaster relief 21
 - Preserving indigenous culture 22
 - Locally produced information 22
 - The technological infrastructure 23
- 1.4 **The Greenstone software** 24
- 1.5 **The pen is mighty: Wield it wisely** 28
 - Copyright 29
 - Collecting from the Web 31
 - Illegal and harmful material 34
 - Cultural sensitivity 34
- 1.6 **Notes and sources** 35

- 2. Preliminaries: Sorting out the ingredients** 39
 - 2.1 **Sources of material** 40
 - Ideology 41
 - Converting an existing library 42
 - Building a new collection 43
 - Virtual libraries 44
 - 2.2 **Bibliographic organization** 46
 - Objectives of a bibliographic system 47
 - Bibliographic entities 48
 - 2.3 **Modes of access** 55
 - 2.4 **Digitizing documents** 58
 - Scanning 59
 - Optical character recognition 61
 - Interactive OCR 62
 - Page handling 67
 - Planning an image digitization project 68
 - Inside an OCR shop 69
 - An example project 70
 - 2.5 **Notes and sources** 73

- 3. Presentation: User interfaces** 77
 - 3.1 **Presenting documents** 81
 - Hierarchically structured documents 81
 - Plain, unstructured text documents 83

- Page images 86
- Page images and extracted text 88
- Audio and photographic images 89
- Video 91
- Music 92
- Foreign languages 93
- 3.2 Presenting metadata 96
- 3.3 Searching 99
 - Types of query 100
 - Case-folding and stemming 104
 - Phrase searching 106
 - Different query interfaces 108
- 3.4 Browsing 112
 - Browsing alphabetical lists 113
 - Ordering lists of words in Chinese 114
 - Browsing by date 116
 - Hierarchical classification structures 116
- 3.5 Phrase browsing 119
 - A phrase browsing interface 119
 - Key phrases 122
- 3.6 Browsing using extracted metadata 124
 - Acronyms 125
 - Language identification 126
- 3.7 Notes and sources 126
 - Collections 126
 - Metadata 127
 - Searching 127
 - Browsing 128
- 4. Documents: The raw material 131**
 - 4.1 Representing characters 134
 - Unicode 137
 - The Unicode character set 138
 - Composite and combining characters 143
 - Unicode character encodings 146
 - Hindi and related scripts 149
 - Using Unicode in a digital library 154
 - 4.2 Representing documents 155
 - Plain text 156

- Indexing 157
- Word segmentation 160
- 4.3 Page description languages: PostScript and PDF 163
 - PostScript 164
 - Fonts 170
 - Text extraction 173
 - Using PostScript in a digital library 178
 - Portable Document Format: PDF 179
 - PDF and PostScript 183
- 4.4 Word-processor documents 184
 - Rich Text Format 185
 - Native Word formats 191
 - LaTeX format 191
- 4.5 Representing images 194
 - Lossless image compression: GIF and PNG 195
 - Lossy image compression: JPEG 197
 - Progressive refinement 203
- 4.6 Representing audio and video 206
 - Multimedia compression: MPEG 207
 - MPEG video 210
 - MPEG audio 211
 - Mixing media 212
 - Other multimedia formats 214
 - Using multimedia in a digital library 215
- 4.7 Notes and sources 216

- 5. Markup and metadata: Elements of organization 221**
- 5.1 Hypertext markup language: HTML 224
 - Basic HTML 225
 - Using HTML in a digital library 228
- 5.2 Extensible markup language: XML 229
 - Development of markup and stylesheet languages 230
 - The XML metalanguage 232
 - Parsing XML 235
 - Using XML in a digital library 236
- 5.3 Presenting marked-up documents 237
 - Cascading style sheets: CSS 237
 - Extensible stylesheet language: XSL 245

- 5.4 Bibliographic metadata 253
 - MARC 254
 - Dublin Core 257
 - BibTeX 258
 - Refer 260
 - 5.5 Metadata for images and multimedia 261
 - Image metadata: TIFF 262
 - Multimedia metadata: MPEG-7 263
 - 5.6 Extracting metadata 266
 - Extracting document metadata 267
 - Generic entity extraction 268
 - Bibliographic references 270
 - Language identification 270
 - Acronym extraction 271
 - Key-phrase extraction 273
 - Phrase hierarchies 277
 - 5.7 Notes and sources 280
-
- 6. Construction: Building collections with Greenstone 283**
 - 6.1 Why Greenstone? 285
 - What it does 285
 - How to use it 288
 - 6.2 Using the Collector 292
 - Creating a new collection 293
 - Working with existing collections 300
 - Document formats 301
 - 6.3 Building collections manually: A walkthrough 302
 - Getting started 303
 - Making a framework for the collection 304
 - Importing the documents 305
 - Building the indexes 307
 - Installing the collection 308
 - 6.4 Importing and building 309
 - Files and directories 310
 - Object identifiers 312
 - Plug-ins 313
 - The import process 314
 - The build process 317

- 6.5 Greenstone archive documents 319
 - Document metadata 320
 - Inside the documents 322
- 6.6 Collection configuration file 323
 - Default configuration file 324
 - Subcollections and supercollections 325
- 6.7 Getting the most out of your documents 327
 - Plug-ins 327
 - Classifiers 336
 - Format statements 342
- 6.8 Building collections graphically 349
- 6.9 Notes and sources 353

7. Delivery: How Greenstone works 355

- 7.1 Processes and protocols 356
 - Processes 357
 - The null protocol implementation 357
 - The Corba protocol implementation 359
- 7.2 Preliminaries 360
 - The macro language 360
 - The collection information database 369
- 7.3 Responding to user requests 372
 - Performing a search 375
 - Retrieving a document 376
 - Browsing a hierarchical classifier 377
 - Generating the home page 378
 - Using the protocol 378
 - Actions 384
- 7.4 Operational aspects 385
 - Configuring the receptionist 386
 - Configuring the site 391
- 7.5 Notes and sources 392

8. Interoperability: Standards and protocols 393

- 8.1 More markup 395
 - Names 395

- Links 397
- Types 402
- 8.2 Resource description 408
 - Collection-level metadata 410
- 8.3 Document exchange 413
 - Open eBook 414
- 8.4 Query languages 419
 - Common command language 419
 - XML Query 422
- 8.5 Protocols 426
 - Z39.50 427
 - Supporting the Z39.50 protocol 429
 - The Open Archives Initiative 430
 - Supporting the OAI protocol 433
- 8.6 Research protocols 434
 - Dienst 435
 - Simple digital library interoperability protocol 436
 - Translating between protocols 437
 - Discussion 438
- 8.7 Notes and sources 440
- 9. Visions: Future, past, and present 443**
- 9.1 Libraries of the future 445
 - Today's visions 445
 - Tomorrow's visions 448
 - Working inside the digital library 451
- 9.2 Preserving the past 454
 - The problem of preservation 455
 - A tale of preservation in the digital era 456
 - The digital dark ages 457
 - Preservation strategies 459
- 9.3 Generalized documents: A challenge for the present 462
 - Digital libraries of music 462
 - Other media 466
 - Generalized documents in Greenstone 469
 - Digital libraries for oral cultures 471
- 9.4 Notes and sources 474

Appendix: Installing and operating Greenstone	477
Glossary	481
References	489
Index	499
About the authors	517