ELSEVIER

Michael LESK

Understanding Digital Libraries

SECOND EDITION

MK MORGAN KAUFMANN

Understanding Digital Libraries

The Morgan Kaufmann Series in Multimedia Information and Systems

Series Editor, Edward A. Fox, Virginia Polytechnic University

Understanding Digital Libraries, Second Edition
Michael Lesk

Bioinformatics: Managing Scientific Data
Zoe Lacroix and Terence Critchlow

How to Build a Digital Library
Ian H. Witten and David Bainbridge

Digital Watermarking
Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom

Readings in Multimedia Computing and Networking
Edited by Kevin Jeffay and HongJiang Zhang

Introduction to Data Compression, Second Edition
Khalid Sayood

Multimedia Servers: Applications, Environments, and Design
Dinkar Sitaram and Asit Dan

Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition
Ian H. Witten, Alistair Moffat, and Timothy C. Bell

Digital Compression for Multimedia: Principles and Standards
Jerry D. Gibson, Toby Berger, Tom Lookabaugh, Dave Lindbergh, and Richard L. Baker

Readings in Information Retrieval
Edited by Karen Sparck Jones and Peter Willett

Understanding Digital Libraries

Second Edition

Michael Lesk

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
MORGAN KAUFMANN PUBLISHERS IS AN IMPRINT OF ELSEVIER

ELSEVIER

MORGAN KAUFMANN PUBLISHERS

Publishing Director	Diane D. Cerra
Senior Editor	Rick Adams
Acquisitions Editor	Lothlórien Homet
Publishing Services Manager	Simon Crump
Project Manager	Dan Stone
Developmental Editor	Corina Derman
Editorial Coordinator	Mona Buehler
Editorial Assistant	Asma Stephan
Cover Design	Frances Baca
Composition	CEPHA Imaging Pvt. Ltd.
Technical Illustration	Dartmouth Publishing Inc.
Copyeditor	Yonie Overton
Proofreader	Daniel Easler
Indexer	Kevin Broccoli
Interior printer	The Maple-Vail Book Manufacturing Group
Cover printer	Phoenix Color

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
APPLICATION SUBMITTED

ISBN: 1-55860-924-5

For information on all Morgan Kaufmann publications,
visit our Web site at www.mkp.com or www.books.elsevier.com

Printed in the United States of America
05 06 07 08 09 5 4 3 2 1

Again, this book is dedicated to the late Gerard Salton (1927–1995), a great information retrieval researcher who introduced many algorithms in the 1960s that are only now becoming commercially recognized, whose students have taught and studied in many universities, and who in 1961 first introduced me to computers, programming, and information retrieval.

This Page Intentionally Left Blank

Contents

List of Figures xiii

List of Tables xix

Figure Credits xxi

Preface xxvii

1. Evolution of Libraries 1

1.1 Why Digital Libraries? 1
1.2 History of Libraries 3
1.3 Vannevar Bush 13
1.4 Computer Technology 16
1.5 Early Language Processing 23
1.6 The Internet and the Web 25
1.7 Summary 29

2. Text Documents 31

2.1 Computer Typesetting 32
2.2 Text Formats 35
2.3 Ways of Searching 39
2.4 Web Searching 51

2.5 Thesauri 52
2.6 Statistical Language Processing 53
2.7 Document Conversion 55
2.8 Summary 59

3. Images of Pages 61
3.1 Scanning 62
3.2 Image Formats 68
3.3 Display Requirements 69
3.4 Indexing Images of Pages 73
3.5 Shared Text/Image Systems 77
3.6 Image Storage vs. Book Storage 80
3.7 Large Scale Projects 87
3.8 Summary 88

4. Multimedia Storage and Retrieval 91
4.1 Sound Formats: Vinyl, Tape, and CD 91
4.2 Pictures: GIF and JPEG 95
4.3 Color 98
4.4 Image Search 100
4.5 Automatic Speech Recognition 105
4.6 Moving Images 107
4.7 Summary 115

5. Knowledge Representation Schemes 117
5.1 Library Classifications 119
5.2 Indexing: Words and Thesauri 127
5.3 Metadata 132
5.4 Knowledge Structures 134
5.5 Hypertext 137
5.6 Vector Models 139
5.7 XML and the Semantic Web 143
5.8 User-Provided Links 148
5.9 Summary 150

6. Distribution 153
6.1 Books, CD-ROMs, and DVDs 153
6.2 Computer Networks 157

6.3 Information on the Internet 165
6.4 Grid Computing 168
6.5 Open Source and Proprietary Systems 170
6.6 Handheld Devices 172
6.7 Security and Cryptography 174
6.8 Privacy 181
6.9 Summary 184

7. Usability and Retrieval Evaluation 185
7.1 General Human Factors Considerations 186
7.2 Text Displays: Fonts and Highlighting 189
7.3 Image Displays and Compression Systems 191
7.4 Web Page Graphics 193
7.5 Interface Controls: Menus and Keywords 195
7.6 Access Methods 203
7.7 Retrieval Evaluation 209
7.8 Page Ranking and Google 213
7.9 Summary 215

8. User Needs 217
8.1 Overview 217
8.2 User Services 219
8.3 Acceptance in the Past: Why Not Microfilm? 220
8.4 Finding Library Materials 221
8.5 Web Searching 223
8.6 Performance Measurement 226
8.7 Need for Quality 228
8.8 Summary 231

9. Collections And Preservations 233
9.1 Traditional Paper Collections 233
9.2 Traditional Preservation Problems: Acid Paper and Perfect Binding 236
9.3 Digitizing Special Collections and Archives 241
9.4 Preservation Through Sharing and Distribution 245
9.5 New Materials and Their Durability 248
9.6 Emulation as a Preservation Technique 259
9.7 Summary 260

10. **Economics 261**

10.1 Library Benefits 263
10.2 Traditional Economics and Economies of Scale 266
10.3 Scholarly Publishing Today 268
10.4 Models for Library Funding 271
10.5 Access vs. Ownership 278
10.6 Administrative Costs 279
10.7 Electronic Commerce 286
10.8 The “dot-com” Boom and Bust 289
10.9 The Future of Quality Information 289
10.10 Summary 291

11. **Intellectual Property Rights 293**

11.1 History of Copyright Law 295
11.2 History of Patent Law 301
11.3 Access Regulation and Control 305
11.4 Technology for Intellectual Property Protection 307
11.5 Peer to Peer Systems 316
11.6 Summary and Future Research 319

12. **A World Tour of Digital Libraries 321**

12.1 Information Policy, Not Industrial Policy 321
12.2 Multilingual Issues 324
12.3 Multicultural Issues 326
12.4 Text-oriented Projects 328
12.5 Images 336
12.6 Sound and Music 342
12.7 Video 343
12.8 3-D Images 346
12.9 Scholarship in Digital Libraries 349
12.10 Scientific Data 353
12.11 Access and Infrastructure 357
12.12 Summary 358

13. **Scope of Digital Libraries 361**

13.1 Readers 361
13.2 Authors 363
13.3 Flow Control 365

13.4 Education 367
13.5 Science 368
13.6 Technology and Creativity 369
13.7 How Big Can They Get? 371
13.8 Summary 372

14. Future: Ubiquity, Diversity, Creativity, and Public Policy 375

14.1 Dream to be Realized 375
14.2 Future Roles in Information Handling 378
14.3 Effect of Digital Technology on Universities 380
14.4 Digital Libraries and Society 381
14.5 Digital Libraries, Society, and Creativity 382
14.6 Public Policy Questions 384
14.7 Projections 385

References 387

Index 413

This Page Intentionally Left Blank

Figures

Figure 1.1 The Allston neighborhood of Boston: (a) 1903 map; (b) modern map; (c) digital orthophotoquad (from aerial photography); (d) Landsat space photograph. 6

Figure 1.2 Rate of growth in university libraries for the past 100 years. 10

Figure 1.3 Rate of growth of scientific journals. 11

Figure 1.4 Recent rate of growth of periodicals. 11

Figure 1.5 *Life Magazine's* impression of the Memex. 15

Figure 1.6 Colossus—the first electronic computer. 17

Figure 1.7 Transistors per chip. 18

Figure 1.8 Changes in the price of memory. 20

Figure 1.9 Computational progress. 22

Figure 1.10 Growth in the size of the Web. 26

Figure 1.11 Growth of Web pages as reported by Google. 26

Figure 2.1 Growth of online databases. 34

Figure 2.2 OCR results of newspaper text. 56

Figure 2.3 OCR of journal passage. 57

Figure 3.1 A comparison of representations of the word *cat.* 62

Figure 3.2 Suspending a flatbed scanner to get "look-down" operation. 64

Figure 3.3 Page-turning and scanning machines: (a) Digitizing Line; (b) Kirtas Bookscan. 65

Figure 3.4 Comparison of scanning resolutions: (a) 300 dpi; (b) 150 dpi; (c) 100 dpi. 67
Figure 3.5 An example of antialiasing. 70
Figure 3.6 Examples of antialiased text: (a) 300 dpi; (b) 100 dpi bitonal; (c) 100 dpi antialiased. 70
Figure 3.7 Example of background, foreground, and a full page. 72
Figure 3.8 A scanned newspaper page from the eighteenth century. 75
Figure 3.9 Scanned heads of newspaper stories. 76
Figure 3.10 Detailed story. 76
Figure 3.11 An example of a journal column. 78
Figure 3.12 Density function. 79
Figure 3.13 Correlation function. 81
Figure 3.14 Classified page. 82
Figure 3.15 (a) Harvard depository; (b) Bibliothèque Nationale de France. 83
Figure 4.2 (a) Thresholding vs. (b) Dithering. 98
Figure 4.4 Gamma representation. 100
Figure 4.5 Image retrieval of "horses." 102
Figure 4.6 Recognizing and labelling parts of an image. 102
Figure 4.7 Assigning keywords to images. 103
Figure 4.8 Recognizing the same window in two pictures. 104
Figure 4.9 Recognizing the same building in two pictures. 104
Figure 4.10 Video segmentation. 111
Figure 4.11 Density distributions for speaker and viewgraphs. 112
Figure 5.1 (a) Frequency of books in Dewey classes; (b) Frequency of books in Library of Congress classes. 123
Figure 5.2 Top level classification on Yahoo. 124
Figure 5.3 Screen display of news stories. 143
Figure 5.4 Structured and diagrammed knowledge representation (XML and tree). 147
Figure 6.1 Abilene network links in the United States. 161
Figure 6.2 Structure of Internet address names. 162
Figure 6.3 Growth of Internet hosts. 163
Figure 7.1 SuperBook interface, 1989. 190
Figure 7.4 Treemap. 205
Figure 7.5 An example of Tilebars. 206
Figure 7.6 Envision retrieval list. 207
Figure 7.7 Kohonen map. 207
Figure 7.8 Recall-Precision curves from the Aslib/Cranfield study. 211
Figure 9.1 A 1958 railroad timetable printed on newsprint. 238

Figure 9.2 (a) Fifth design for the US Capitol, principal floor, by Stephen Hallet, 1793 (Library of Congress); (b) Letter from Thomas Jefferson to Charles William Frederic Dumas, Sept. 1, 1788 (Library of Congress); (c) Notebook page, draft of application to Carnegie Foundation by Charles Sanders Peirce, 1902 (Harvard University Library); (d) Gen. William Tecumseh Sherman, photograph by Matthew Brady, 1865 (Library of Congress). 242

Figure 9.3 Current New England state boundaries superimposed over New England in 1771. 250

Figure 9.4 Map of Orange, New Jersey, from the early 1900s. 252

Figure 9.5 Vector map of Orange, New Jersey, from the TIGER map file. 252

Figure 9.6 Diskettes (8-inch, 5.25-inch, and 3.5-inch). 253

Figure 9.7 Magnetic tape formats. 254

Figure 9.8 Punch card. 254

Figure 9.9 Tape life expectancy. 255

Figure 9.10 Tape storage recommendations. 256

Figure 10.1 Proportion of readings that achieve savings vs. time spent reading. 264

Figure 10.2 Library economics. 265

Figure 10.3 Supply and demand curves. 266

Figure 10.4 Supply-demand curve for Journal reading. 267

Figure 10.5 Book purchases vs. published titles. 269

Figure 10.6 Comparison of journal prices and cost per issue (CPI). 270

Figure 10.7 Numbers of ACM members and ACM journals. 270

Figure 10.8 Fraction of published papers not cited in 10 years (for papers originally published in 1984). 283

Figure 10.9 Predictions of online advertising revenue vs. reality. 288

Figure 10.10 Growth of online subscriptions to *The Wall Street Journal* and *Consumer Reports.* 288

Figure 11.1 Global trends in software piracy. 296

Figure 11.2 Examples of hardware locks. 309

Figure 11.3 Bell Labs document marking. Examples (a) and (b) show word-shift coding. In (a) the top text line has the same spacing after the "for." In (b) these same text lines are shown again without the vertical lines to demonstrate that either spacing appears natural. Examples (c)–(e) show feature coding performed on a portion of text from a journal table of contents. In (c) no coding has been applied. In (d) feature coding has been applied to select characters. In (e) the feature coding has been exaggerated to show feature alterations. 311
Figure 11.4 Flickering for protection. 314
Figure 11.5 Flickering, lower density. 314
Figure 11.6 Flickering letters. 315
Figure 11.7 Fraction of GDP spent on music. 318
Figure 12.1 US imports and exports of databases and information services. 323
Figure 12.2 International Children's Digital Library Web page. 327
Figure 12.3 A scan of an original page of *Beowulf.* 331
Figure 12.4 A scan of an original page of a Gutenberg Bible. 332
Figure 12.5 Scanned pages from *Yatsu Yagi*, or "Eight Goats," a Japanese children's book. 333
Figure 12.6 Storage of the wooden printing blocks of Tripitaka Koreana. 333
Figure 12.7 Printing block and print of Tripitaka Koreana. 334
Figure 12.8 Map from the French national collection. 336
Figure 12.9 Aztec manuscript from the Vatican library. 337
Figure 12.10 Historical photograph from the State Library of Victoria. 337
Figure 12.11 Bitmap of a letter written by George Washington. 339
Figure 12.12 Partitioning of aerial photography into areas of similar land use. 340
Figure 12.13 Chest radiograph from UCLA. 340
Figure 12.14 Images from a dolphin fin digital library at Eckerd College. 341
Figure 12.15 Papyrus. 342
Figure 12.16 Score from Aida in Johns Hopkins University music library project. 344
Figure 12.17 An example of parsing and searching music. 344
Figure 12.18 Search results of "bin laden couriers" in CMU's video digital library. 346
Figure 12.19 Three-dimensional cuneiform tablet from UCLA's 3-D digital library project. 347

Figure 12.20 One side of a 3-D scan of a piece of chiseled stone originally belonging to an ancient Roman map chiseled in marble. 348
Figure 12.21 Digital representation of Beauvais Cathedral. 348
Figure 12.22 (a) Historic photograph of the synagogue in Wiesbaden; (b) Computer visualization of a 3-D model of the synagogue. 350
Figure 12.23 CT of crocodile skulls from the University of Texas digital library of fossil vertebrates. 351
Figure 12.24 Analysis of human movement. 351
Figure 12.25 Searching an experimental 3-D digital library for chairs. 352
Figure 12.26 Page from George Washington's diary from the American Memory site. 354
Figure 12.27 Growth of the Protein Data Bank. 355
Figure 12.28 Image from the Sloan Digital Sky Survey. 355
Figure 12.29 Earthquakes around the world from the IRIS consortium's seismic database. 356
Figure 12.30 Cave paintings from Dunhuang, China, as digitized by the International Dunhuang Project. 359
Figure 13.1 Growth of number of articles read by a typical scientist in a typical year. 363
Figure 13.2 Growth of purchase of electronic journals by ARL libraries. 373

This Page Intentionally Left Blank

Tables

Table 1.1 Number of volumes held by major US libraries. 9
Table 1.2 Number of volumes held by major global libraries. 9
Table 1.3 Memory sizes. 19
Table 2.1 Operators and their meanings. 40
Table 2.2 Revenues of text retrieval software companies. 50
Table 3.1 An example of bytes in different compression formats. 69
Table 3.2 Costs of new academic library buildings. 85
Table 3.3 Comparison of book storage costs. 85
Table 4.1 Size of GIF image depending on number of colors. 97
Table 4.2 Improvements in speech recognition. 106
Table 4.3 Disk space for one minute. 109
Table 5.1 Classification systems used by the Library of Congress. 120
Table 5.2 British Museum Library classification system prior to 1808. 121
Table 5.3 Top levels of Antonio Panizzi's classification system for the British Museum Library. 121
Table 5.4 Dewey Decimal system and current Library of Congress classification system. 122
Table 5.5 Top level organization of Internet portals. 125
Table 5.6 Effectiveness of recommending movies. 149
Table 6.1 Cost breakdown for producing a journal. 154
Table 6.2 Transmission speeds. 160
Table 6.3 Comparision of international Internet broadband growth. 164
Table 7.1 Comparison of search results. 191

Table 7.2 Percentage of keyword searches in sessions. 199
Table 8.1 Time spent per URL. 223
Table 8.2 Frequency of visits by type of query. 224
Table 8.3 Search results from ArtAbstracts and Google for "paleography." 229
Table 8.4 Search results from ArtAbstracts and Google for "Raphael, fresco." 230
Table 9.1 Expected media lifetime. 255
Table 9.2 Word processing software advertised. 257
Table 10.1 Cost comparison of company library to no company library. 263
Table 10.2 Cost comparison of acquisition of item with savings from having it. 264
Table 10.3 Market value in October 1999 and April 2003. 289
Table 11.1 Costs of software piracy. 296
Table 11.2 Patents applied for by hardware and software companies. 302

Figure Credits

Figure 1.2 Based on data from Leach, S. 1986. “The Growth Rate of Major Academic Libraries: Rider and Purdue Reviewed.” College and Research Libraries 37 (Nov) and from Rider, F. 1944. *The Scholar and the Future of the Research Library*. New York: Hadham Press.

Figure 1.3 Based on data from Meadows, J (1993) “Too Much of a Good Thing?” in H. Woodward and S. Pilling, eds. *The International Serials Industry*, Aldershot, Hampsire: Gower Publishing.

Figure 1.5 Courtesy of Mrs. Joan Adria D’Amico, the executor of the estate of Mary Crimi, wife of Alfred D. Crimi.

Figure 1.6 Courtesy of the Public Records Office, London FD 850/234.

Figure 1.7 Based on data from several sources including “Birth of a Chip.” 1996. *Byte* (Dec) and “The Personal Computing Industry by the Numbers.” 1997 *PC Magazine* (March).

Figure 1.8 Based on data from several sources including Pugh, E. 1995. *Building IBM: Shaping an Industry and Its Technology*. Cambridge, MA: MIT Press.

Figure 1.9 Adapted from Hammann, D. 1983. “Computers in Physics: An Overview.” *Physics Today* 36(5).

Figure 2.2 Courtesy of the *London Daily Mail*.

Figure 2.3	Courtesy of the *American Chemical Society*.
Figure 3.1	Courtesy Ann B. Lesk.
Figure 3.2	'Flying' scanner by Bob Kobres, University of Georgia Libraries. Reprinted with permission.
Figure 3.3	(a) 4DigitalBooks.com, (b) www.kirtas-tech.com. Reprinted with permission.
Figure 3.7	From DjVuZone.com, Leon Bottou. Reprinted with permission.
Figure 3.8	The World. *Burney London Daily Journal*. 8.3.1728, PB.MIC.BUR.823B. Used by permission of the British Library.
Figure 3.9	Screenshot of heads. Used by permission of the British Library.
Figure 3.10	Screenshot of heads. Used by permission of the British Library.
Figure 3.11	*The Journal of Physical Chemistry*, Vol.92 No.26, p.7162 (Fig.2). Copyright 1988 American Chemical Society. Reprinted with permission.
Figure 3.13	*The Journal of Analytical Chemistry*, 63(17), 1697-1702. © Copyright 1991 American Chemical Society. Reprinted with permission.
Figure 3.14	*The Journal of Organic Chemistry*, Vol. 56, No. 26, p.7270 (Fig.1) © Copyright 1991 American Chemical Society. Reprinted with permission.
Figure 3.15	(a) Courtesy of the Harvard University Library, (b) Courtesy of the French National Library, Francois Mitterand, Dominique Perrault, Architect, Alain Goustard, Photographer.
Figure 4.2	(a) Courtesy of Anne B. Lesk.
Figure 4.4	Courtesy of www.bberger.net/rwb/gamma.html.
Figure 4.5	Courtesy of EE/CS University of California, Berkeley.
Figure 4.6	Courtesy of EE/CS University of California, Berkeley.
Figure 4.7	"Automatic Image Annotation and Retrieval using CrossMedia Relevance Models by Jeon, Lavrenko and Manmatha," *SIGIR'03*. Reprinted with permission.
Figure 4.8	Courtesy of the Department of Engineering Science, University of Oxford.
Figure 4.9	Courtesy of Department of Engineering Science, University of Oxford.