From Paper to Web

How to Make Information Instantly Accessible

Tony McKinley

Table of Contents (mini)

Table of Contents (actual)

Index

Introduction

Dedication

About the Author

ADOBE PRESS

Adobe Press books are published by Macmillan Computer Publishing USA.

From Paper to Web, the paper edition, is available at bookstores, both physical and virtual. ISBN: 1-56830-345-9. $45 US.

© 1997 Adobe Systems Incorporated. All rights reserved.

Ch. 1: The Past - A Paper Legacy
Ch. 2: The Present: Making Existing Digital Documents Accessible
Ch. 3: The Future: Digital Documents
Ch. 4: Acrobat Exchange: An Architecture for Instant Access
Ch. 5: Acrobat Catalog: Creating the Keys to Instant Access
Ch. 6: Decisions in Indexing
Ch. 7: Acrobat Search
Ch. 8: Enhanced PDF Collections on the Web
Ch. 9: Advanced Navigation for Superior Information Access
Ch. 10: Organizing Digital Documents
Ch. 11: HTML Documents: Creation, Editing, Management
Ch. 12: Advanced Searching Techniques
Ch. 13: Tapping the Web: A Fountain of Information
Ch. 14: Document Management: Information as a Corporate Asset
Ch. 15: Publishing on CD

table of contents

Part 1: Creating Digital Content

Chapter 1: The Past - A Paper Legacy 3

Paper and Microfilm: Imperfect Mediums 4

Digital vs. Paper Documents 5

Tools for Global Distribution 6

Problems on the Digital Road 11

Competing with the Comfort of Paper 15

Working with Digital Documents 16

Scanned Document Management 24

Getting from Paper to Digital Format 25

Document Preparation 27

Scanning Techniques 33

Scanner Specifications and Settings 35

Fine Tuning OCR 40

Chapter 2 - The Present: Making Existing Digital Documents Accessible 43

Why use a Universal Format? 45

Platform Independence Provides Easy Viewing 46

Printer Independence to Retain Appearance 46

Integrate PDF with Text Databases 47

The Common Archival Solution 47

Benefits of Independence 48

Sacrifices 49

Decisions to Make 53

Working with Applications 54

Chapter 3 - The Future: Digital Documents 59

The Fantastic Pace of the Web 60

Surfing in Waves of Information 62

Intelligent Agents 63

Digital Information in Perpetual Action 66

Making Digital Documents Better than Paper 72

Part 2: Managing Digital Content

Chapter 4: Acrobat Exchange:
An Architecture for Instant Access 77

Finding Aids: Author, Title, Subject, Keywords 79

Acrobat General Info Fields 80

PDF Document Info 81

Additional Fields Customize Document Appearance 82

Superior Navigation 85

Benefits of Delivery of PDF 88

Precision: Retrieve ONLY Relevant Docs 90

Recall: Retrieve ALL Relevant Docs 90

Sacrifices 93

The Dreaded Learning Curve 94

Decisions to Make 96

Process: Secondary Publishing (BIOSIS) 100

Chapter 5: Acrobat Catalog: Creating
the Keys to Instant Access 103

Transforming Vast Collections into Retrievable Information 104

Document Management #101 106

Full Powered Boolean Search 110

Field / Full Text Search Combined 111

Optimizing Build Options 113

Chapter 6: Decisions in Indexing 117

How Will Future Users Access the Collection 118

Dynamic or Static Collection — How to Manage Updates 120

Naming Conventions 121

The Process 124

Gather PDF Files for Your Collection 124

Finalize Doc Info and Navigation Enhancements 125

Help Users Grasp the Collection 126

Building the Index 127

Serve Your Index 129

Part 3: Searching Digital Content

Chapter 7: Acrobat Search 131

Why Use Search? 132

Basic Text Searching 133

Building More Extensive Queries 136

Variables to Restrict Queries 142

Narrowing Your Search via Doc Info 143

Beyond Doc Info Fields: Expert Searching 146

Range Searches 147

Relevancy Ranking in Results List 148

Building on Previous Retrievals 151

Limitations of Acrobat Searching 153

Chapter 8: Enhanced PDF Collections on the Web 159

Full Power Searching on the Web 161

Basic Boolean Operators 162

Evidence Operators 164

Proximity Operators 164

Relational Operators 165

Concept Operators 165

Modifiers 167

Simple Search 169

Standard Search 170

Power Search 171

Web Servers and SearchPDF 172

Search Results in HTML 172

Multi-Platform Access 175

Chapter 9: Advanced Navigation for Superior Information Access 179

Acrobat 3 Plug-Ins 180

Re:Mark Tools 187

Compose Tools 190

Aerial Tools 193

Sys-Print 197

RAELS 200

PDF Reference Manual 201

Part 4: Using Digital Content

Chapter 10: Organizing Digital Documents 203

Using HTML Documents to Organize Files 205

Building Organization into New Documents 215

Chapter 11: HTML Documents: Creation, Editing, Management 221

PageMill: The Web Page Builder 222

Management with SiteMill 229

Part 5: Advanced Resource Guide

Chapter 12: Advanced Searching Techniques 233

Methods for Advanced Searching 235

Fuzzy Logic 239

Semantic Word Relationships 240

Intelligent Agents 241

Relevance Ranking and Term Weighting 242

Advanced Text Searching in Action 243

Exploiting Verity Search: A Boolean Primer 244

Chapter 13: Tapping the Web: A Fountain of Information 249

Searching the WWW: Tools to Target Your Information Hunt 250

Getting Noticed: Attracting Humans and Electronic Spiders 260

Chapter 14: Document Management: Information as a Corporate Asset 267

Document Management Principles 269

Classic Document Management Fields 269

Assembling Virtual Documents on the Fly 272

Workflow: Assembly Line Style 273

Complex Document Management via RDBMS 272

Chapter 15: Publishing on CD 277

Archiving to CD 281

Write It Once 282

Acrobat Catalog and Search on CD 285

Production via CD-Recordable 286

Mass Production 286

Web Links 287

index

- **3-D Virtual Reality** **10**
- **3-D visualization software** **97**
- **8.3 convention** **13**
- **Abbott, Dave** **122**
- **abstract** **22**
- **access to information** **20**
  - *HTML links* *205*
  - *improved access* *114*
  - *instant access* *20, 113, 118*
  - *Intranets* *205*
  - *Rapid Access Electronic Library* *200*
  - *speed* *121*
  - *why digital docs are superior* *180*
- **accession #** **23**
- **accuracy rates** **19**
- **Acrobat** **8, 24, 52**
  - *Acrobat 3 with Verity SearchPDF* *160*
  - *Acrobat 3 benefits* *24*
  - *Acrobat 3 PDF links actions* *92*
  - *Acrobat 3 Reader* *9, 129, 160, 185, 198, 209, 285*
  - *Acrobat Software Development Kit* *155*
  - *adding sound AIFF or WAV format* *92*
  - *add-ons, see Ambia*
  - *capabilities mirroring HTML* *92*
  - *create link tool* *91*
  - *custom forms built online* *182*
  - *document management using Doc Info fields* *272*
  - *frequently asked questions, FAQ* *57*
  - *General Info fields* *80*
  - *ini file* *127*
  - *on a LAN* *129*
  - *on the Internet* *209*
  - *RAELS PDF Links* *200*
  - *Reader download* *93*
- **Acrobat Capture** **8**
  (see also OCR, scanning)
  - *Acrobat 3 Capture Plug-in* *180*
  - *clean-up* *198-199*
  - *editing* *186*
  - *Emerge case study* *198-199*
  - *formatting* *199*
  - *graphics* *199*
  - *OTA case study* *12-13*
  - *PDF Image & Text* *181, 198*
  - *PDF Normal* *181, 198*
  - *samples of Capture processing at* From Paper to Web *site* *8*
  - *TouchUp tool* *186*
- **Acrobat Capture Service Bureau** **198**
  - *process* *198-199*
- **Acrobat Catalog** **104**
  - *collections* *106, 113*
  - *customize fields* *126*
  - *dynamic collection* *113*
  - *index and full text search* *115*
  - *index for CD* *198, 285*
  - *index process* *104-105, 113-115, 118, 124-125, 134*
  - *index subdirectories -assists, morgue, parts, pdd, style, temp, topicdx, trans, work* *113, 124*
  - *indexing options* *283*
  - *online guide* *127, 128*
  - *process for creating and maintaining* *113-115*
  - *purge option* *128*
  - *searchable database* *104, 109*
  - *stopwords* *119, 126*

- *term expansion using evidence operators* 164
- *use conventional directory tree* 124
- *uses Verity text-search* 104

Acrobat Exchange

- *action buttons* 135
- *annotate* 91
- *find results containing text* 136, 139
- *FTR and index search* 109
- *index features and choices* 136
- *punctuation* 139
- *search screen* 135
- *term expansion* 139
- *text search* 81, 133
- *Verity Search* 245
- *wildcards* 139

Acrobat plug-ins 172, 180

- *AutoClose* 180
- *AutoIndex* 180
- *Capture* 180
- *EPS Links* 181
- *ExecMenu* 181
- *finding plug-ins on the Web* 201
- *Forms* 182
- *Import Image* 182
- *Movie* 183
- *OLE Server* 183
- *Optimizer* 183
- *PDF Type Utility* 184
- *Scan* 184
- *Search* 185
- *SuperCrop* 185
- *SuperPrefs* 185
- *TouchUp* 186
- *WebLink* 186

Acrobat Search 90, 104, 132

- *advantages over free text retrieval* 90
- *features* 114, 135
- *free Reader* 132
- *limits* 153
- *rank by option* 242
- *Search for CD* 134, 285

ActiveX 145, 147

- *citation applet* 147

Adobe Distiller 283

Adobe Illustrator 227

Adobe magazine

- *adobemag.com online help and articles* 86

Adobe PageMill 222

- *animated GIF* 227
- *background* 227
- *connection speeds* 228
- *download statistics tool* 228
- *drag and drop copying in PageMill* 222
- *GIF images* 228
- *GUI* 222
- *guide* 227
- *HTML* 222
- *HTML imported into PageMill* 222
- *JavaScripts* 229
- *limiting images* 227
- *navigation bars* 226
- *placeholder* 229
- *plug-ins* 222, 228
- *text enhancements* 222
- *Web authoring tools* 222

Adobe PDF Writer 283

(see also Acrobat Capture)

Adobe Photoshop 227

Adobe Premier 187

Adobe SiteMill 229-230

- *broken links* 229
- *error messages* 230-231
- *managing Web sites* 231
- *site view* 229

tracking URL addresses 230

Webmaster 229

Adobe Systems 6, 46, 74

advanced text search 234-246

(see also search)

advanced browsing 238

automate thinking 235

automatic summary 238

concept searching 235

dictionaries 240

fuzzy logic 235

goals of text searching 235

human-like accessibility 234

intelligent agents 235, 241

natural query language 235

query expansion 235

relevance ranking 242

search for ideas 234

semantic searching 240

term weighting 242

thesauri 240

Aerial from Ambia 193

alphabets — see documents with foreign language

alpha-numeric strings 39

AltaVista

advanced search 251

screen for Boolean queries term weighting 251

Ambia, Inc. 100, 187

Automatic bookmarks and TOC 191

Compose 190, 198

Re:Mark 187

review 190

using Compose at a service bureau 198

American Library Project 71

American Memory collection 72

Ameritech 72

Ancient Scriptorium 56

annotated lists 212

annotation 4

(see also Exchange, Ambia Re:Mark)

ability to color, label, customize fonts 91

annotated 187

colors 187-189

highlighted 187

like Post-It notes 91

sharing and annotating files using Exchange 91

tools 187

updated 187

appearance 16

Appendices 219

Apple Laser Writer 50

Apple QuickTime Viewer 92, 183

Archie 70

Architext 252

archive 23, 47, 55

to CD 281

article 15, 84

Artificial Intelligence (AI) 61, 73

As We May Think 63, 81, 265

AS/400 64

ASCII 25, 45, 67, 78

80 column ASCII 78

Association of Research Libraries 213

AT & T 63

author 22

Author 78, 80

normal or reversed order 81

author's guide

options that affect index size, speed, and search features 114

Automatic bookmarks and TOC 191

Automatic summaries **253**

AVI video file **183**

background

unwanted *35*

base font **199**

batches **12**

(see also scanning)

batch management *26*

optimize *184*

scanning and indexing with keyword fields *100*

sizes *30*

Bell Telephone **65**

Bibliography

bibliographic index *22*

binary **36**

binary files **45**

Binary Large Object (BLOB) **27**

bindings **28**

Biological Abstracts **22**

Biological Sciences Information Services (BIOSIS) **96,100, 259**

bitmap **18**

BLOB **27**

(see also Binary Large Object)

Blueworld **55**

bold **16**

book scanners **28**

adjustable bed *28*

bookmark by example **191**

recognized by combination of font, size, format *191*

bookmarks **20, 46, 82, 157**

auto-generated using FrameMaker, Pagemaker 86

automated bookmarks *200*

Bookmark by example in Ambia's Compose *191, 198*

like thumb tabs *86*

multiple level bookmarks *191*

organizer *209*

use with chapters, section, subsections, charts, illustrations *86*

viewed in vertical window *86*

books **5,15, 97**

bound documents *28*

Boolean

Boolean Query Translation project *62, 247*

concept search software *247*

double quotes to define a phrase *246*

expanded search *138*

expressions *144*

logic in query terms *110, 137*

operators - And, Or, Not *110, 137-138,162*

query example *138*

search *110*

stopword *246*

text searching *156*

text-search operation *244*

within Document Info fields *244*

border removal **36**

bound documents **26, 28**

brightness **35**

broken characters **40**

browser **46, 222**

(see also Netscape, Microsoft)

across platforms *88*

configurability *46*

LYNX *160*

Mosaic *160*

reading from local drive or network *88*

browsing **87**

advanced browsing *238*

build **113**

include, exclude *113*

  - *options* 113, 120-121
  - *query expansion* 113
- **bulletin board systems BBS 279**
- **Bush, Vannevar 7, 63, 97, 265**
  - *“As We May Think”* 63
- **byte-serving 183, 209**
- **Capture 184**
  - **(see also Acrobat Capture, OCR, scanning)**
- **card catalog 52, 79, 216, 218**
  - *online catalog vs. searching physical library* 218
- **card stock 30**
- **cascading stylesheets 46**
- **case sensitive 114, 119**
- **Catalog**
  - **(see Acrobat Catalog)**
- **categorical information 23**
- **categorization 63**
- **categorizing 21**
- **CCD arrays 32,35**
- **CCITT Group IV 38**
- **CD-ROM 13, 88**
  - *Acrobat Web links* 287-289
  - *audio* 285
  - *bookmarks* 128
  - *capacity* 128
  - *CD-recordable* 286
  - *distribution* 289
  - *ganged array* 286
  - *hybrid* 198
  - *index using Catalog* 198
  - *ISO 9660* 121, 124, 198
  - *jukebox* 286
  - *mass production* 286
  - *maximum number of files indexed on CD* 128
  - *navigation* 128
  - *optimize option* 128
  - *SearchCD* 129, 134, 285
  - *speeds* 283
- **CD-ROM publishing**
  - *archiving to CD* 281
  - *cost of mass-produced CD-ROM* 278
  - *with web* 87
- **centralized printing 67**
- **CERN 50**
- **character 18**
- **Character styles**
  - *logical styles* 223
  - *PageMill* 223
  - *physical styles-plain, bold, italic* 223
- **character-based interfaces 46**
- **Charge-Coupled Device (CCD) 32**
- **Chemical Abstracts 22**
- **citation 144, 146**
- **clean up 26, 199**
- **collections see also Acrobat Catalog**
  - *customized fields* 126
  - *design rules* 126
  - *dynamic or static* 120
  - *indexing* 118
  - *large collections* 177, 278
  - *linked by HTML* 204
  - *managing* 156
  - *meta-information* 146
  - *mingled lists* 211
  - *online help on home pages* 174
  - *preparation using relational operators* 164
  - *RDBMS* 146
  - *sharing common collections* 268
  - *simple lists* 210
  - *size* 118, 127
  - *subsets of* 106
  - *updates* 120
  - *value* 180

- color-coded annotations 190
- comment 189
- commercial printers 72
- commercial publishers 98
- communications 63
- Compose from Ambia 190
- compression 4, 37-38, 44
  - *compression schemes 44*
  - *compression setting 51*
  - *high-power optics 4*
  - *physical compression 5*
- computer display 17
- computer memory 60
- computerized information 108
- Concept Searching
  - *wildcards 237*
  - *word stemming 237*
- content info 79
- contrast 35
- conversion 4, 38, 97
- copiers 30
- Cornerstone Imaging 184
- corporate memory 132
- costs 11
  - *advertising and mailing 99*
  - *conversion 4*
  - *microfilm vs. paper 4*
  - *of change 94*
  - *of information acquisition 100*
  - *paper and toner 98*
  - *printing 98*
  - *scanning 19*
  - *shipping vs. transmitting information (see table) 10, 279*
  - *shipping paper vs. CD 6, 278-280*
  - *staffing 11*
  - *training 11*
- Courtot, Phillippe 132, 235 (see also Verity)
- Cover, Robin 47
- Create Link Tool 91
- created date in PDF 82
- creator in PDF 81
- cross-architecture with PostScript 52
- cross-platform with PostScript 52
- Current Contents 22
- customize doc appearance in PDF 82
- data 18
- databases 44
- data-compression 24
- data-mining 197, 204
- Dataware 240
- date 22
- de-bind 26
- decremented page count 29
- density of information 18
- Designing Web Graphics
  - *216 color palette 227*
  - *Weinman, Lynda 227*
- de-skew 26
- de-speckle 26
- dial-up service 74
- digital document 18
  - *superior to paper 180*
- digital libraries 68-69, 72, 135, 146, 156
- Digital Libraries Research and Development 65
- Digital Library Project 62, 67
- digitization 11
- Distiller 51, 53, 78
  - *Distiller Assistant 53*
- document information fields 79
  - *Boolean functions within fields 244*

comparison operators *245*

fields as hyperlinks *214*

info fields in PDF

81-82, 90, 92, 106, 144-145, 157, 160, 268

queries for multiple fields *245*

search capability *246*

special operators list *245*

document management **4, 106-107**

access control *269*

and Acrobat fields *270*

audit tracking *268-269*

browser to handle MIME *271*

Catalog *115*

date field *106*

Documentum *269*

global contract *107*

law firms *268*

MIME Type *271*

Open Text *269*

organizing digital documents *204-218*

ownership *268*

parallel systems *274*

password-group *106*

PCDocs *269*

permissions *268, 270*

RAELS PDF Links, a proven time saver *200*

RDBMS *273*

remotely shared collections *275*

rules to define relations between fields *200*

Saros *269*

security *106*

version control *106, 268-269*

virtual docs *272*

Web-accessible *275*

workflow *273, 275*

document prep **26-32**

de-bind *28, 198*

fragile documents *31*

guillotine paper cutter *28, 29, 198*

unbound *28*

documents **48**

appearance and page layout with PDF *78, 81-86*

building collections *101*

building digital documents *201*

collections *104*

comments and emphasis using markup tools *187*

components *107*

encapsulated format *133*

encryption *84*

filenames *204*

finding digital documents *204*

flow *84*

foreign language *78*

format *127*

info fields in PDF *81-82, 90, 92, 106, 144-145, 157, 160*

info specific to PDF *81*

legal docs *97, 107*

linked by HTML *204*

monitor settings in PDF *83*

password-protected *84*

preservation *47*

readability *48*

reading large documents on a monitor *84*

searching with Acrobat Search *90*

security *84*

settings and controls in PDF *81-86, 90*

standards *98*

universal *98*

use PDF with formulas *78*

Web-centric *204*

dots per inch (dpi) **37**

(see also resolution)

downloading **24**

- **draw on 189**
- **driver 184**
- **drop-out colors 40**
- **duplex 32**
- **Dynabook 14, 60, 148**
- **edge detection 36**
- **editing 26**
- **Education Automation 68, 265**
- **electronic books 66**
- **electronic distribution 279**
- **electronic documents**
  - *superior to paper 10*
- **electronic publishing 44,53**
- **Electronic Workplace Technologies 41**
- **e-mail 44, 189**
  - *research by e-mail 55*
- **embedded fonts 51**
  - **(see also fonts)**
- **Emerge 55, 198**
  - *Acrobat Capture Service Bureau Case Study 198*
  - *Acrobat plug-ins 201*
- **encryption 84**
- **enhanced navigation 79**
- **error correction 27**
- **Excalibur Technologies 47, 73, 133, 240, 274**
- **Exchange 51**
  - **(see also Acrobat Exchange)**
  - *combines FTR with index fields 90*
  - *Horizontal Split 51*
  - *PDF info fields 90*
  - *Replace pages 51*
  - *using to share and annotate files 91*
- **Excite 64, 105, 264**
  - *concept searching 252*
  - *Excite and Magellan home page 252*
  - *for Web Servers 236*
  - *free search engine 105*
  - *Magellan editor's reviews 252*
  - *Web Search Engine 236*
- **excluded terms 114**
- **expert systems 39**
- **fast SCSI 37**
- **fax 18**
- **fax machines 32**
- **fax server 32**
- **file attachments 189**
- **file formats 44**
- **file server**
  - *local 207*
- **file size 16**
  - *aggregate file 16*
  - *estimate 50K per page image 16*
  - *Group IV compressed .tif image 92*
- **File Transfer Protocol FTP 207**
- **files 12**
  - *.RTF file 45*
  - *.SIT file 44*
  - *.ZIP file 44*
  - *8.3 convention 13*
  - *batches 12*
  - *binary files 45*
  - *compression 209*
  - *generated daily 3*
  - *lost 4*
  - *moving large files like video and sound 93*
  - *names 12, 118*
  - *naming PDF files 288*
  - *optimize PDF files 209*
  - *server 207*
  - *sharing and annotating using Exchange 91*
    - *(see also annotations)*
  - *sound files 48,92*
  - *tips for richness and efficiency 51*

- **filing** 4
  - *computer- assisted retrieval* 5
  - *conventional* 4
  - *file cabinet* 4
  - *file folder* 4
  - *misfiled* 4
  - *strictly defined field* 5
- **financial statements** 52
- **Finding Aids**
  - *Author, Keywords, Title, Subject* 78
- **finding aids** 5
- **fine mode fax** 38
- **floppy disks** 47
  - *3.5 inch format* 47
  - *5.25 inch format* 47
  - *8 inch format* 47
- **font**
  - *attributes* 16
  - *embedded* 51, 84
  - *relative font sizes* 224
  - *size changes* 223
  - *subsets* 51
  - *Type 1, True Type* 83
- **font size** 199
- **footnotes** 5, 20
- **format** 45
  - *BMP* 182
  - *GIF* 182
  - *HTML* 48
  - *Native file formats* 45
  - *old or unreadable* 47, 94
  - *PCX* 182
  - *PDF* 48, 182
  - *publication format* 49
  - *RTF* 45
  - *TIF* 182
  - *universal format* 45, 54, 94, 99
- **formatting** 199
- **FrameMaker** 54
- **Frames in PageMill** 224
  - *Java script buttons* 224
  - *no borders* 224
- **free software** 49
- **FTP** 13
- **FTR — see full text retrieval**
- **Fujitsu Scanners**
  - *3093* 34
  - *M3099* 34, 184
  - *ScanPartner 10C* 34, 184
- **Fulcrum Technologies** 47, 73, 133, 237
  - *Vector Space Similarity Searching* 237
- **full screen** 82
- **full text retrieval FTR** 108, 110
- **Fuller, Bucky** 64, 68, 97, 264
  - *"Education Automation"* 6
- **fuzzy logic**
  - *Expanded Or search* 239
  - *OCR errors* 239
  - *to recognize errors or variables in words* 239
- **garbage** 45
- **Gates, Bill** 64, 145, 281
- **general info** 21
- **Gernert, Glenn** 198-199
- **Geschke, Chuck** 52, 88
- **GIF** 24, 33, 264
- **GIFbuilder** 227
- **global library** 265
- **glossary** 5, 219
- **Gopher** 70, 207
- **Government Printing Office** 12
- **gradient** 35
- **graphical user interface** 6
  (see also GUI)
- **graphics** 33, 199

- **gray 32**
- **grayscale 33**
- **groupware 4**
- **GUI 206**
- **guillotine 29**
- **Gutenberg Press 15, 66**
- **halftones 36**
- **handwriting 65**
- **headlines 15**
- **High Sierra Standard 284**
- **highlight 7, 21, 187-188**
- **highlighters 40, 188**
- **high-speed communications 74**
- **hit lists 73, 149**
  - **(see also relevance ranking, results list)**
  - *all query terms appear 242*
  - *automatic summary for contents of hits 238*
  - *highlight hits 151*
  - *hit clustering 242*
  - *hit density 149, 168,242*
  - *occurrences 149*
  - *proximity 149*
  - *reconfigurable hit lists 73*
  - *relevance ranking 242*
  - *search results 149*
  - *term weighting 242*
  - *total hits 149*
- **Horizontal Split 51**
- **HotBot 254**
- **HotWired 254**
- **HP Accupage, programs for 33**
- **HP NetServer 37**
- **HP ScanJet IIIC 34**
- **HTML 45-46,48, 51-53, 56, 67, 264**
  - *across platforms 147*
  - *anchors 206*
  - *create content not appearance 222*
  - *custom index fields 262*
  - *editing 205*
  - *Frequently Asked Questions (FAQs) 208*
  - *hypertext reference 206*
  - *IETF HTML 2.0 spec 205*
  - *linked documents and collections 204*
  - *locally available fonts and graphics 222*
  - *Meta tag fields 99, 215, 261, 262*
  - *print and HTML design 208*
  - *view source code 215*
  - *web pages 46*
  - *WebLink 186*
- **HTTP 46, 261**
  - **(see Hypertext Transport Protocol)**
  - *access method 206*
- **HyperDOC**
  - *a service of U.S. National Lib. Of Medicine 67*
- **hyperlinks**
  - *hot text 205*
  - *hyperlink copying in Ambia Compose 191*
  - *RAELS PDF Links in very large documents 200*
  - *underline or highlight represents link 205*
- **hypertext links 7, 20, 85, 157, 204, 261**
  - *contents 208*
  - *search 208*
  - *techniques 208*
- **HyperText Markup Language 7**
  - **(see also HTML)**
- **Hypertext REFerence (HREF) 206**
- **HyperText Transport Protocol 7, 46**
  - **(see also HTTP)**
- **illustrations 33**
- **Illustrator 51, 55, 215**
- **image maps 207, 215**
  - *view source code 215*
- **image(s) 16,18**
  - *downsampling 228*

- *image database* 24
- *image enhancement* 26
- *image files* 24
- *image processing* 30
- *in PDF* 182
- *interlaced images* 228
- *label* 228
- *resizing* 228
- *scanned images* 16
- *schematic drawings* 215
- *treatments for online display* 227

impact printers 6

import 45

incremental numbering 11, 12

incremented page count 29

independent platforms 46

index 4, 5, 84, 219

- *basic index fields* 13, 99, 125
- *bibliographic* 22
- *build options* 113, 127
- *case study, what to index* 96
- *Catalog nine subdirectories* 113, 128
  - *(see also Acrobat Catalog)*
- *citation* 144, 146
- *collections* 113-114
- *combined text and index queries* 115
- *continuously option* 120
- *creation process* 113-114, 124
- *design decisions* 96, 114, 119
- *every option* 121
- *how to* 112, 118-129
- *index criteria* 85, 118
- *index data* 24
- *index field* 4
- *index information* 21
- *index scheme* 4
- *indexing info* 23
- *limit of indexing* 22
- *Meta Tags* 99
- *navigation with Catalog* 115, 125
- *options* 119
  - *(see also index options)*
- *purge* 113
- *retrieval through doc info fields* 115
- *searching* 109
- *size recommendations* 127
- *stopwords* 119
- *value* 99

index options 119, 120, 124

- *build options* 119
- *bus connections* 127
- *case sensitive* 120
- *effects of options on index search speed* 120
- *effects of options on index size* 119
- *info fields* 125
- *numbers* 120
- *random access memory RAM* 127
- *sounds like* 120
- *stemming* 120
- *stopwords* 119, 126

Indices 51

info fields 79

- *four classes of data* 79
- *in Acrobat PDF - author, keywords, subject, title* 79 - 80
- *system info, doc info, content info, enhanced navigation* 79

Info-fill in Ambia Compose 193

- *populate the Info Fields with data from an ASCII file* 193

info-gathering robot 62

(see also Intelligent Agent)

information 18

- *density* 18,53

- *equal rights to* 65
- *paper-based* 25
- *speed of transmission* 183
- *strategic use* 65
- *value* 18,19

information retrieval 47, 73
(see also Salton, Dr. Gerard)
- *competition* 152
- *Exchange* 90
- *measuring* 90
- *precision and recall* 90, 161

Informix 274
- *Datablades* 274

InfoSeek 256, 263
- *five secrets to better searching* 257

infrastructure 60
- *communication* 60
- *computation* 60
- *memory* 60

ink-drawing tool 189

Inktomi 254

Input Solutions 36

intellectual density 98

Intelligent Agent software 14, 61- 63, 166
- *advanced text searching* 235
- *automated clipping service* 241
- *combining index fields with text in PDF* 88
- *constantly updated info* 241
- *continuously searching the Web* 241
- *overcoming non-productive surfing* 88
- *persistent searches* 241
- *searching and following links* 88
- *spiders* 241
- *too busy or don't want to surf* 88, 241
- *Web robots* 241

Intelligent Character Recognition (ICR) 61

interactive 3-D 181

Interleaf 54

Internet 6

Internet Publishing
- *Internet Publishing with Acrobat* 54

Internet Service Providers 44, 64

INTERNIC 218

Intranets 205

IP address 218

ISDN 63, 279

ISIS drivers 184

ISO 9660 121, 124, 283, 289
- *birth of ISO 9660* 284

ISP — see Internet Service providers

Java 73, 145, 147
- *citation applet* 147
- *definition* 250

job status 26

job tracking 26

Johnson, Kelly 64

joined characters 40

Karlin, Jon 240

Kay, Alan 14, 97
- *Dynabook* 14, 81

Kent, Gordon 54

key fields 20

key words 22, 38, 78, 157
- *registered* 81
- *undocumented* 81

Kofax 9275 Image Processor 37

Kurzweil Intelligent Scanning Systems 39

Kurzweil, Ray 60
- *Kurzweil's law* 60
- *The End of Handicaps* 61

labor 30

- **laser printers 6,17**
- **Lexical query expansion 235**
  - *word stemming, wildcards, fuzzy, pattern recognition 235*
- **Lexis 6**
- **librarian 23**
- **Library of Congress 68, 72**
- **links 18, 53, 86, 134**
  - **(see also web links)**
  - *Chain Linker and Link Auditor in Ambia Compose 192*
- **links to text 18**
  - **(see also links)**
- **list formats 223**
- **Lists 53**
  - *alphabetical 214*
  - *annotated 212*
  - *navigation with tools 214*
  - *speed dialer links 214*
- **Listservs 55, 78**
  - *Acrobat listserv 57*
  - *email discussion groups 55*
  - *PDF listserv 57*
- **live documents 45**
  - **(see also files, documents)**
- **Logical query expansion 235**
  - *thesaurus, dictionary, concept relations 235*
- **long-distance, telecomm market 75**
- **Los Alamos Research Library 212**
- **Lotus Notes 78, 83**
- **Lucent 63**
  - *56 Kb modem 74*
- **Lycos 73, 259, 262**
- **LYNX 46**
- **Mac 184**
- **magnifying glass 40**
- **mail to: function 207**
- **maps 50**
- **mark-up 41**
  - **(see also drop-out)**
  - *comment 189*
  - *draw on 189*
  - *file attachments 189*
  - *helping groupware applications 188*
  - *highlight 187-188*
  - *multi media attachments 189*
  - *strike-out 188*
  - *tools 187-188*
  - *underline 188*
- **Marsh, Bob 176**
- **Marx Law Library 69**
  - *Project Diana 69*
- **master copy 51**
- **Mead Data Central 6**
- **MedWeb 67**
- **megabyte 37**
- **memex 7, 63**
- **Meta tag fields 99, 215, 261, 262**
- **Meta-information 79, 106, 119, 146**
- **microfilm 4**
  - *viewer 5*
- **Microsoft 64, 75, 92**
  - *Access 155, 183*
  - *ActiveX 145, 147*
  - *Internet Explorer 52, 264*
  - *Internet Information Server 186*
  - *Microsoft Office 64*
  - *Microsoft Professional 64*
  - *Object Linking and Embedding OLE 145*
  - *Windows NT 254*
- **MIME 189**
- **Minolta book scanner 28**
- **Mirroring 9**

modems

1200/2400 baud 18

56 Kb technology from Lucent 74

downloading speeds 228

ISDN 63

modified date in PDF 82

moire patterns 36

Monastery of Christ in the Desert 56

monitor size 84

Moore, Andy 123

Moore's Imaging Dictionary 124

Moore's law 60

Mosaic 160, 185

Quarterdeck 186

Spyglass 186

multi-media and attachments 91, 189

Must Contain modifier 255

Myst 183

naming conventions 121

National Digital Library Project 68

National Science Foundation 67- 68

native file formats 45

natural language 152-153

search engines 252

navigation 5

Aerial enhancement package from Ambia 193

annotated lists 214

between documents 85

different from search 85

Help file contents 208

Help file search 208

index button 214

index criteria 85

moving within info 85

navigation buttons 46, 214

offer enhanced navigation 46

page forward, page back 214

page number navigation via Ambia Aerial 195

shortcuts 51

tag frequently referenced pages 195

TOC button 214

tools 219

using thumbnails 85-86

within a document 85

NCSA 50

NetCarta CyberPilot Pro 241

netiquette 55

Netscape 92, 172

default width of window 227

font size changes 223

Netscape Navigator

24, 52, 172, 186, 264, 299

network resources 24

neural computers 61

newsgroups 78

newsletters 51

Nexis 6

NT 64

object interactivity 97

OCR 19

(see Optical Character Recognition)

Office of Technology Assessment 12

online help

adobemag.com 86

open pdf settings 82

Open Text Corporation 47, 73, 133, 258

operating systems

client/server environment 175

Mac, Windows, UNIX, DOS 88

operator training 49

Optical Character Recognition (OCR)

19, 69

accuracy rates 19, 123

errors 40-41

  - *fuzzy search* 154, 239
  - *OCR errors* 40
  - *OCR process* 19
  - *OCR project* 19
  - *robot typist* 38
  - *user choices* 25
- **optimize** **183-184**
  - *byte-serving* 183, 209
- **orientation** **31**
- **OTA** **12**
- **output** **18**
  - *high quality with PostScript* 89
  - *image + text* 18
  - *image only* 18
  - *scanned documents* 18
  - *text only* 18
- **Output format** **25**
- **Output page** **25**
  - *reproduce page* 25
  - *single column plain text* 25
  - *text with graphics* 25
- **page composition** **7**
- **page layout** **16**
  - **(see also page)**
- **page magnification** **82**
- **page numbers** **82, 199**
- **page per minute (ppm)** **36**
- **page quality** **25**
  - *auto* 25
  - *dot matrix* 25
  - *normal* 25
- **page segmentation** **26**
- **page(s)** **5, 22**
  - *composition* 7
  - *count* 29
  - *double-sided* 29
  - *fragile* 31
  - *image page* 5
  - *layout and appearance retention* 8, 16, 82
  - *page sizes* 5
  - *single-side* 29
  - *size* 5
  - *text page* 5
  - *thickness* 30
- **Pagemaker** **50, 54**
- **PageMill — see Adobe PageMill**
- **Pantone color** **40**
- **paper** **30**
  - *carbon copies* 31
  - *glossy* 30
  - *onion-skin* 30
  - *stocks* 30
  - *thermal fax paper* 31
  - *thickness* 30
  - *to digital process* 26
  - *weights* 30
- **paper handling** **33**
- **paper weights** **30**
- **Paragraph Formats** **222**
  - *address format* 222
  - *heading sizes* 223
  - *paragraph alignment* 223
  - *pre-formatted* 222
  - *six levels of heading formats* 223
- **password-protected** **84**
- **PDF** **25, 46- 49, 51- 54, 67, 72, 78**
  - **(see also Portable Document Format)**
  - *and HTML* 52, 88-89
  - *and video* 92
  - *customize document appearance* 82
  - *doc info fields: creator, producer, created, modified* 81
  - *encapsulated document format* 133

  - *Images in PDF are universally accessible across platforms* 182
  - *info fields* 79
  - *navigation features* 51
  - *pdf links* 92, 181
  - *PDF normal* 154
  - *pdfmark reference manual* 82
  - *PDF-to-RTF converter in Ambia Re:Mark* 190
  - *presentation* 91
  - *presentation control* 224
  - *printing* 88-89
  - *progressive display* 209
  - *serve PDF files one page at a time* 209
  - *settings, views* 81-83
  - *summarizing contents of PDF docs* 91
  - *use PDF with formulas from simple math to complex equations* 78
  - *using PDF to share info online* 78
  - *Web Links* 87, 209
- **PDF Group** **72**
- **PDF Reference Manual** **201**
- **PDFWriter** **53, 78**
  - *as printer substitute* 54
- **Pentium Scan Station** **37**
- **performance**
  - *set preferences for peak performance* 127
- **Personal Library Systems (PLS)** **47, 133**
- **Persuasion** **55**
- **Photoshop** **45, 55, 215**
- **phrase searching** **254**
- **picture element** **32**
- **pixel** **32**
- **Pixel Translations** **184**
- **plain text** **78**
- **platforms** **17, 88**
  - *client server environment* 175
  - *hardware* 17
  - *HTML docs across platforms* 88
  - *independence* 46
  - *medium* 88
  - *Novell LAN, CD-ROM, WWW* 88
  - *multi-platform with PDF* 88
  - *platform specific files* 44
  - *software* 17
  - *web browsers for all platforms* 88
- **Plattner, Dr. Hasso** **64**
- **Popkin, Jaime** **41**
- **Portable Document Format** **8** **(see also PDF)**
- **PostScript** **8, 46, 49, 78, 89**
  - *as a standard* 17
  - *using on the web* 50-52
- **ppm** **36**
- **precision** **109, 153** **(see also search)**
- **preference** **21**
- **preservation** **47**
- **printers**
  - *dot-matrix printers* 17
  - *laser printers* 17
  - *substitute PDFWriter* 54
- **PrintGear** **88** **(see also printing)**
- **printing**
  - *centralized* 67
  - *comparing PDF with HTML* 88
  - *PostScript* 89
  - *printing selected page sections* 196
  - *using Adobe PrintGear* 88
- **printout** **48**
- **print-to-speech** **60**
- **priority change** **26**

- **process**
  - *paper to digital* 26
  - *secondary publishing* 100
- **producer**
  - *in PDF* 82
- **Project Diana** **69**
- **Project Open Book** **70**
- **proprietary** **44**
- **publication format** **49**
- **publication status** **51**
- **publishing**
  - *CD-ROM and the web* 87
  - *index design decisions* 96-97
  - *PostScript* 88
  - *with PrintGear* 88-89
- **punctuation** **15**
- **purge**
  - *deleting subdirectories before build* 113, 128
- **quality control** **26**
- **QuarkXPress** **50, 54**
- **Query language** **108**
- **Query**
  - **(see also search, precision, recall, advanced text search)**
  - *Boolean* 236
  - *Boolean search operations* 110
  - *build options* 113
  - *concept operators* 166
  - *concept searching* 237
  - *difficulty* 105-106
  - *examples* 138
  - *expansion* 111, 139-141, 164
  - *include/exclude feature* 151
  - *lexical term expansion* 235
  - *logical term expansion* 235
  - *match case* 142
  - *natural language query* 153, 236
  - *prefixes and suffixes* 140
  - *punctuation in queries* 139
  - *refine search* 151
  - *relational operators* 165
  - *relevancy rank* 142, 148-149
  - *restricting queries* 142
  - *root* 140
  - *searching for a phrase* 137
  - *sounds like* 114, 119, 140
  - *spelling* 139
  - *stopwords* 119, 137, 236
  - *terms* 135
  - *thesaurus* 141
  - *truncations* 140
  - *using Acrobat Search* 90
  - *weighting* 168, 171
  - *wildcards* 139-140, 236
  - *word assistant* 142
  - *word stemming* 140, 235-236
- **queue** **37**
- **Rapid Access Electronic Library (RAELS)** **200**
- **Re:Mark** **186**
  - *adds emphasis to original document* 189
  - *annotation tool* 186
- **Reader** **9**
  - *Acrobat Reader download* 93
  - *Reader EXE.* 9
- **reading machine** **61**
- **recall** **109**
  - **(see also search)**
- **receptivity** **35**
- **recognition** **25**
- **red drop-out** **41**
- **Reed Technology Information Services** **122**

references 20

Register-It! 265

rekeying 38

Relational Database Management System (RDBMS) 90

results lists 150

structured and unstructured 90

relevance ranking 148-149, 168, 171-172, 242, 255

remote user 28

Replace pages 51

repurposing 73

re-scan 26

resolution 33, 37-38

2400 dpi 34

600 dpi 35

72 dpi, 200 dpi, 300 dpi, 33

extended 34

optical resolution 34-38

results list 73, 148

clustered hits 150

density 168

dynamic views 150

highlight 150

in HTML 172

primary sort field 140

search preferences 150

sort 150

retrieval 90

(see also information retrieval IR, searching, full text retrieval FTR, query)

bandwidth 129

difficulty 90

Dr. Gerard Salton 90

FTR 109

large collections 273

precision 90,109

recall 90, 109

relevancy ranking 148, 242

results list 148

speed of 129

term weighting 242

using Exchange 90

review

Re:Mark tool 190

Rich Text Format (RTF) 25, 45, 67

richness 7

robots 146 , 260

(see also Intelligent Agent)

Eliza Program 146

Salton, Dr. Gerard 90

IR 244

measuring retrieval effectiveness 109

sans-serif 39

SAP 64

scanned images 16

(see also scan)

scanners 11

book scanners 28

desktop 11

DEST OCR scanner 123

dpi 34

Fujitsu 34

HP 34

resolution 34, 123

scanner controller software 11

settings 35-41

sheet feeder 28

scanning 25, 40, 69

accuracy 122

batch scanning 12, 198-199

binary 198

bound documents 198

case study *122, 198-199*
choices *18, 123*
configurations *34*
Exchange *198*
for the Government Patent Office *122*
grayscale *198*
guillotine *198*
high volume *122, 198-199*
images *12*
pages per minute ppm *34, 36*
procedures *27*
process *36, 122*
productivity *122*
speed *34, 36*
scientific papers 22
screen options in PDF 83
screen, split screen 85
scriptoria 56
scrolling
through thumbnails *85*
SCSI 34
Search 10, 20-21
(see also retrieval, index, query, advanced text searching)
advanced text search *234-246*
by author, title, subject *216*
date fields *147*
document info fields *151*
expanding a search *111*
expert searching *146*
field and full text search combined *111*
field values *143*
five basics of searching *257*
for ideas *235*
four basic stages *135*
full text search *10, 108-111, 204*
fuzzy search *154, 239*
grouping *138*
include/exclude feature *151*
match case *142*
modifiers *167*
Must Contain modifier *255*
narrow the search *143*
Near operator *256-257*
N-space term expansion *253*
phrase *257*
precision *109, 111, 161, 171*
preferences *150*
proximity operator *256-257*
proximity operators *142,164*
query examples *138*
range search *147*
recall *109, 111, 161, 171*
refining a search *111, 142, 151*
relevancy ranking *148, 149, 171, 242, 257*
results list *148-150*
searching for a phrase *137*
semantic term expansion *253*
simple, standard and power search *161-162, 169-171*
single and multiple field search *145*
sort *150*
sounds like *114, 119, 140*
stopwords *119, 137, 236*
term expansion using evidence operators *164*
term weighting *242, 257*
thesaurus *111*
topics *132*
use of parentheses *138*
user's knowledge *111*
wildcard *111, 146*
word assistant *142*
word search *10*
word stemming *111, 140*

search engine 21
(see also Acrobat Search, web search engines, Dataware, Excite, Excalibur, Fulcrum, Infoseek, Open Text, Personal Library Systems, ZyLab)
AltaVista *250-251*
automating Boolean Searches *110*
big search engines *250*
concept search engine *64,110*
global research on the web *250*
natural language *154*
topic AGENT Server *132*
Verity *132*
searchable archives 55
(see also archive)
SearchPDF 47
secondary publishers 22
security 84
serif 39
SGML 67, 69, 264
SGML Web Page 47
(see also Cover, Robin)
sheet feeder 28
SIGIR
ACM Special Interest Group on Information Retrieval *68*
signature 7,15
Silicon Graphics SGI 97
Cosmo *97*
SIT file 44
SiteMill — see Adobe SiteMill
size 16
skewed text 31
Solomon, John 36
Sony
8x CD drives *280*
Sony Electronic Publishing Systems
case study 176
CD image collection *176*
sort by confidence 253
sort by site 253
sorted 21
Sound to Acrobat files 92
sound file 10, 48
(see also files, sound)
AU, AIFF, AIFC, WAV *228*
sounds like 114, 119, 140
Source page 25
multi-column *25*
single column *25*
text with graphics *25*
spamdex 263
special handling 44
speck 35
spiders 146, 260
Spock 108
spreadsheets 44
SprintLink 64
stack 27
Standard Generalized Markup Language (SGML) 47
sticky notes 188
stopwords 119, 137, 236
storage systems 24
strike-out 188
Subject 78
(see also Doc Info fields)
Sun Microsystems 147
support 49
surfing the Web 62
automated by IA *63*
don’t want to surf *88*
symbols 15

Sys-Print for high-volume printing to PDF 197

system info 79

T1 and T3 lines 64

table of contents TOC 5

automatically generated via bookmark by example 191

hyperlinked to pages 191

instant page locator with links 208

Table/figure headings 199

tables in PageMill 225

border 226

caption 225

cells 225

columns 225

create table 226

imported from Excel 226

inspector 226

rows 225

spreadsheet 225-226

Tagged Image File Format (TIF, TIFF) 48

tagging 21

telco 64

telecommunications 44

(see also ISDN)

56KB 64

telecommunications service provider 64

teletype 18

term weighting 242, 254, 255

(see also relevance)

emphasis 242

positive and negative weighting 242

text 18

Text and Graphics together 199

text databases 47

text search 204

text searching 133

four basic stages 135

The Electronic Document Systems Assoc. 41

thesaurus 111, 157

thresholding 26

thumbnails 82, 86

generate automatically 157

jumping with thumbnails 85

miniaturized displays 85

scrolling through thumbnails 85

split screen 85

thumbnail views 46, 85

TIFF 24

Title 22, 78

(see also Document Information)

TOC 219

Touch up tool 49

trained 39

training 49

transmissibility 20

TWAIN 184

type styles 16

typeset page 16

typesetting 7

typewritten page 16

single-spaced 16

U.S. National Information Infrastructure Virtual Library 68

underline 16, 188

universal document format 44, 94, 97

(see also format)

Universal Resource Locator URL 87, 147

connecting LAN and WAN to Internet and Intranets 218

explanation of URL parts 206

home page 217

HTML information *205*
register your URL *254*
tracking with SiteMill *231*
University of Cincinnati 69
UNIX 7, 184

user interface options
in PDF *83*
value of information 19
Verity 47, 73, 109, 132, 235
online documentation *163*
PowerSearch Results page *173*
SEARCH'97 *132*
SearchPDF *109, 129, 160, 166, 172, 177*
simple, standard and power search
161-162, 169-171
three interface screens *161*
Topic *156, 166-167*
Veronica 70
versions 51
VGA monitor 33
video 189
video editing — see Adobe Premier
video input 65
videos 10, 92
viewer 45
viewing characteristics 185
virtual agency 13
Virtual Reality Modeling Language
VRML 97, 181
Parisi, Tony *97*
Pesce, Mark *97*
virtual working models *181*
VR browser *97*
VR-assisted *181*
voice input 11
WAIS 70
Warnock, John 46,52
Watermark 184
wavelength 35
Web
browsers *88*
Intelligent Agents *241*
search *62*
web of trails *63*
webcrawlers *146*
webmaster *105, 201, 273*
web links 86, 134
branching *204*
Universal Resource Locator URL *87*
updates *87*
with PDF *87*
with URL *87*
web publishing
with CD-ROM *87*
Web search engines 62, 260
always see user's guide *243*
intuitive techniques *108*
more like that *63*
multiple search engines *73*
Web site
Adobe Home page *211*
home page *105*
home page set up *210-211*
like magazine cover *207*
presentation *105*
sub-pages *105*
weightless information 6
Wharton School of the University of Pennsylvania 50
What's Legal 70
Whitehouse, Kendall 50
wide area network WAN 73
wildcard 111

  - *examples* *140*
  - *fuzzy search* *155*
  - *in fields* *146*
  - *left-hand and right-hand truncations* *140*
- **Windows** **95, 184**
  - **(see also Microsoft)**
  - *registry function* *93*
- **wireless** **11**
- **Word** **25**
- **word processing** **44, 107**
- **Word Search** **10**
- **word stemming** **111, 114, 119, 140, 157**
- **WordPerfect** **45**
- **World Wide Web** **7, 69, 74**
  - **(see also Web)**
  - *publishing on* *52*
- **World Wide Web Consortium** **72**
- **WWW Virtual Library** **68**
- **www.search.com** **260**
- **Xerox** **67**
- **Xplor International** **3**
- **Yahoo** **73, 259**
- **ZIP file** **44**
- **zoom in & out** **50**
- **ZyLab**
  - *ZyImage for scanning, recognition and text retrieval* *240*
  - *ZyIndex* *240*

Acknowledgments

I must thank Patrick Ames, Publisher of Adobe Press for giving me this opportunity to work on this labor of love. My editors, Mike Britton and especially Suzanne Van Cleve provided insight and direction that shaped this book and made this book better than I ever imagined it could be. Bradford Foltz designed the pages and gave the lively feel and openness that a book about the Web should have. Diane Boccadoro fine-tuned the text and any errors remaining are purely my own.

I'm also compelled to acknowledge my colleagues, customers and friends who lived through the history of scanning and recognition with me, always providing comfort and encouragement: John Solomon, Wayne Crandall, Steve Goodfellow, Jon Karlin, Sam Couchara, Dave Abbott, Joanne Tremonte, Peter Atwood, Larry Shiohama, Mike Schofield, Bob Boerner and so many others I'd like to list. It's been a pleasure and a privilege working with Adobe people, Chris Hunt, Alan Williams, Judy Kirkpatrick, Anna Kjos; and the guys from Emerge, John Cook, Andy Young and Glenn Gernert. For making the Internet and the Web feel like my own backyard, I'll always be grateful to Jeff Miller (Sysop) at ONIX (my ISP) and online friends David Orr, Todd Gerbert, Andrew Sutherland, Philip Yip (Technoid, Camel, Egghead & Qwerty) and the rest of the gang. And for giving me the chance to contribute, I thank Bruce Hoard and Andy Moore of Imaging World.

I'd like to thank my Mom for believing all along that my degree in Creative Writing was not for naught. Thanks, Mom.

Finally, for reading everything and providing those critical first line editing services and patient analysis, support, and Common Sense, cutting out the dense jargon and shinola, translating everything to human-readable sentences: my wife, Pat McKinley, read every word of this and aimed to make it understandable by the largest number of people.

Introduction

In this book I've tried to share my excitement and enthusiasm for the great new technology that is providing easy and unprecedented access to information. It might seem like a strange obsession, but I've been dedicated to turning paper into digital form for over 15 years and finally — everything works!

Dating back to the earliest Optical Character Recognition machines, I've worked at finding better ways than typing to get information into computers. Of course, in the early days we mostly wanted to just get information off paper to simply edit it or recompose it and then print it right back out on paper. But, even that was exciting. Look, a machine that reads and types!

By the early '80s, more interesting things were being done with electronic information, and I had the opportunity to work with such online database pioneers as Mead Data Central, Chemical Abstracts and BIOSIS. Now that digital information itself was being consumed over online terminals, OCR seemed to have found a higher calling and I was lucky enough to work for Kurzweil Computer Products.

The information that we wanted to convert to digital form was now in typeset form, unlike the far simpler typewritten documents that early OCR could handle. We could ”train” our Kurzweil scanners to recognize magazines for Mead's NEXIS database, and for the giraffe-high stack of Chemical Abstracts published in hard cover. Compared to OCR's earlier limitations of being able to read only about a dozen typestyles, Kurzweil Intelligent Scanning Systems were a real breakthrough.

In those days, there were precious few prospects for our scanners. One limitation may have been the $50,000 + price tag, but the more severe limitation was that there just wasn't much you could do with digital documents. By the late '80s local area networks and personal computers had become more powerful and popular. But still, to equip a PC to run a moderately fast scanner, we had to add special hardware boards that often cost more than the computers they were installed in. By 1990, I took great pride in the fact that we could put together a decent one-seat scanning system for about $25,000.

But every step of the way was still very expensive, and everything had to be beefed up for imaging applications, including storage, networks, displays and printers. The high speed OCR systems still cost $20,000-$30,000 or more. Then, as Moore's law kicked in over a few generations of PCs the computers finally grew into the demanding requirements of imaging and recognition. But still, transmitting images beyond the LAN was demanding and expensive, requiring dedicated communications lines. And CD-ROM readers cost in the thousands, and CD recorders were beyond the reach of all but the hardiest companies boldly investing in electronic publishing.

It was a lot of fun, but document imaging never had the great breakout that we all expected for so long. We were still limited to niche applications, not far removed from the earliest industrial and professional imaging applications. Then, the Web came along and woke up the slumbering giant of the Internet. Suddenly we had something to do with all the digital documents we could create!

The fact that the PCs, CD drives, storage, monitors, modems and every other piece continually got faster and cheaper certainly helped move things along too.

But my personal expectations remained unfulfilled. Optical character recognition was still expensive because it was difficult to reproduce the appearance of paper documents and it was costly to edit and clean up the results of OCR.

Then one day a couple of years ago, I was walking by the Adobe booth at the AIIM show, the biggest annual event in the imaging industry. It was curious that Adobe was even there, but what I was seeing on the big screen monitor was even more curious. They had just scanned a page, done recognition, and were displaying the results. Unlike the results of OCR, which contains tildes for unrecognizable characters in the text, there were small images of the suspect words blended into the document! And the output looked just like the original! To someone who had spent 15 years toiling in the field of OCR, spanning hundreds of installations and millions of dollars worth of systems, this was absolutely incredible.

As soon as it became available, I got my hands on a copy of Adobe Acrobat Capture 1.0 and put it to the test in our Online OCR Lab. I was stunned. I literally sat in front of my computer in wide-eyed astonishment. Now, I don't expect everyone to feel that way, but then not everyone has personally done more than 2,000 OCR demos like I have. If I'd had Capture over the years I could have met the needs of hundreds and hundreds of customers that conventional OCR simply could not satisfy.

I have one more confession to make, and that is that my interest in converting paper to digital documents goes back beyond my years in OCR. My inspiration for this field comes from Buckminster Fuller, in his 1962 book Education Automation. In that typically freewheeling talk Bucky proposed a universally accessible digital library that would enable anyone, anywhere to study, learn and grow. Bucky figured that this intellectual freedom of the masses would bring humanity's best ideas to reality.

Now, that may sound like a lofty goal, but if we're going to get there, the path is clear before us. The theme of this book is that we can now provide superior access to information. It's not just that we can digitally miniaturize books and get them out of the library without leaving home. The key is that we can provide instant access to the information within the documents.

In my study of Acrobat Capture, I explored the efficiencies of Adobe's Portable Document Format, which offers a range of features from cross-platform viewing and printing to a built-in set of management and search capabilities. Not only does Capture do what my 2,000 OCR demo customers ask for, but PDF seems to meet the vaguely defined but clearly required needs of digital documents and digital libraries.

I've tried to address the entire experience of moving from paper to digital documents on the Web. After finding the most efficient way to create the new documents, we move on to the even more important question of how to organize the information within the digital library. Finally, the issues and techniques of Information Retrieval are explored to give both publishers and users some useful tips and tools for finding what they seek on the Web.

While the lessons learned in this book came from many years in business and academic applications of this technology, I still aspire to the ideals espoused by Bucky Fuller. Now that the technology to make the world a better place that he predicted is here, we're on our way.

— T.M. - 1/20/97

Dedication

I lovingly dedicate this book to my wife, Patty
and my children, Laura, Tony and Hugh.

About the Author

Tony McKinley is a industry analyst, writer and consultant, and is a principal in Intelligent Imaging. Mr. McKinley's dedication to document imaging and recognition began in 1978, and includes five years at Kurzweil Computer Products, and six years as President and Founder of an early imaging systems integration company. Mr. McKinley's Online OCR Lab has performed testing for some of the largest vendors in the field, and recent analysis and writing clients have included Adobe, Xerox, Caere, Fujitsu, Canon, Open Text, Excalibur Technologies and ZyLab. He is a Contributing Editor for *ImagingWorld Magazine*. He lives near Valley Forge, Pennsylvania.

Copyright © 1997 Adobe Systems Incorporated. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of Adobe Systems Incorporated.

Library of Congress Catalog No.: 96-78987

ISBN: 1-56830-345-9

10 9 8 7 6 5 4 3 2 First Printing: February 1997

Published by Adobe Press, Adobe Systems Inc., http://www.adobe.com

The information in this book is furnished for informational use only, is subject to change without notice, and should not be construed as a commitment by Adobe Systems Incorporated. Adobe Systems Incorporated assumes no responsibility for any errors or inaccuracies that may appear in this book. The software mentioned in this book are furnished under license and may only be used or copied in accordance with the terms of such license. Contact the software manufacturer directly for terms of software licenses for any software mentioned in this book not originating from Adobe Systems Incorporated.

Adobe, the Adobe Press logo, Acrobat, Acrobat Capture, Acrobat Exchange, Acrobat Reader, Adobe FrameMaker, Adobe Illustrator, Adobe PageMill, Adobe PageMaker, Adobe Photoshop, Adobe SiteMill, Distiller, PDFWriter, and Postscript are trademarks of Adobe Systems Incorporated. All other brand or product names are the trademarks or registered trademarks of their respective holders.

Printed in the United States of America by GAC Shepard Poorman, Indianapolis, Indiana.

Published simultaneously in Canada.

Editorial and Production: Van Cleve Britton Publishing Ltd. Design: Bradford Foltz

Chapter opener images: ©1996, PhotoDisc, Inc.

Adobe Press books are published and distributed by Macmillan Computer Publishing USA. For individual, educational, corporate, or retail sales accounts, call 1-800-428-5331, or 317-581-3500. For information address Macmillan Computer Publishing USA, 201 West 103rd Street, Indianapolis, IN 46290, or at http://www.mcp.com.