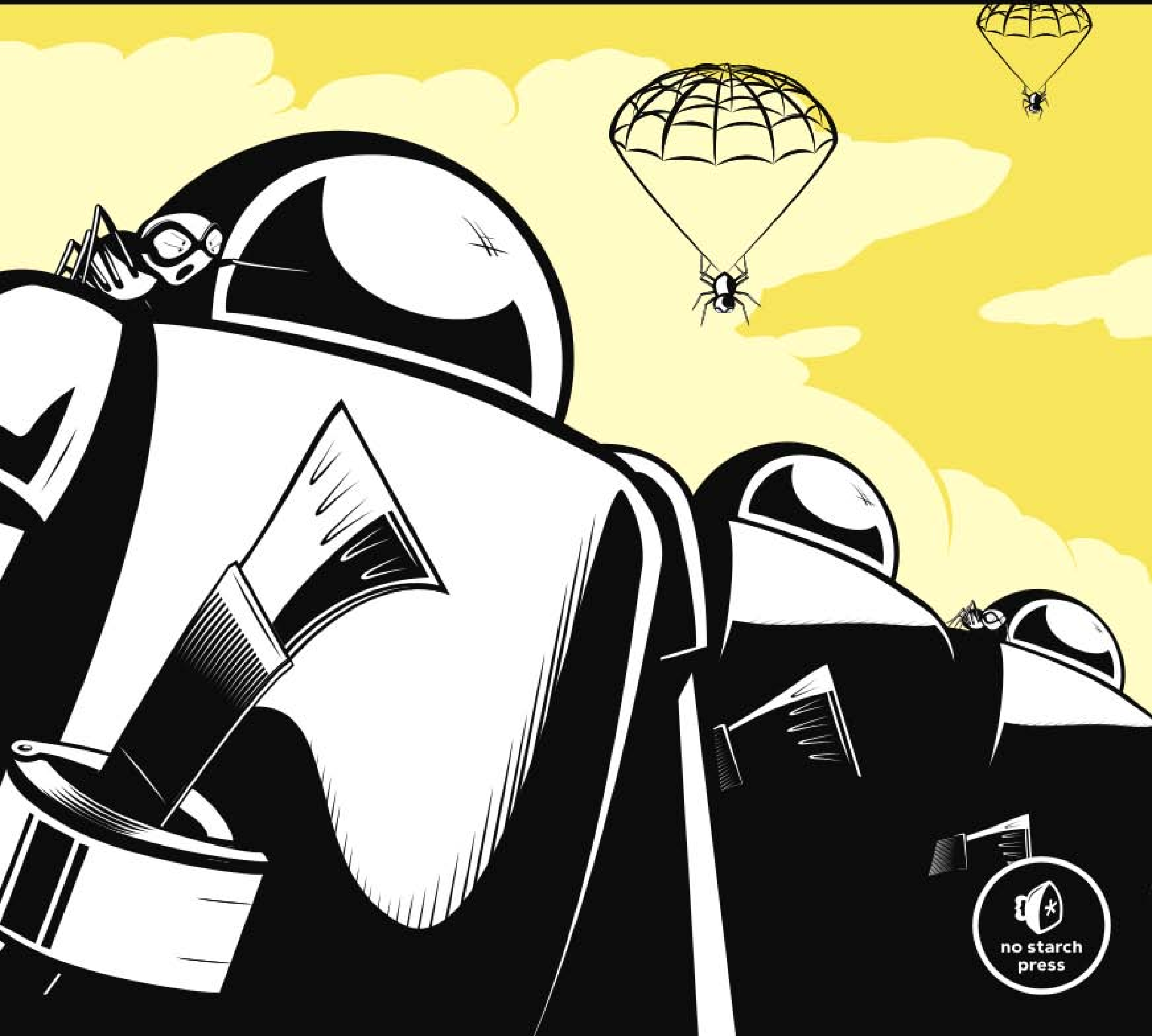
2ND EDITION

# WEBBOTS, SPIDERS, AND SCREEN SCRAPERS

A GUIDE TO DEVELOPING *INTERNET AGENTS* WITH *PHP/CURL*

MICHAEL SCHRENK



no starch press

# CONTENTS IN DETAIL

ABOUT THE AUTHOR xxiii

ABOUT THE TECHNICAL REVIEWER xxiii

ACKNOWLEDGMENTS xxv

INTRODUCTION 1

Old-School Client-Server Technology .......... 2
The Problem with Browsers .......... 2
What to Expect from This Book .......... 2
Learn from My Mistakes .......... 3
Master Webbot Techniques .......... 3
Leverage Existing Scripts .......... 3
About the Website .......... 3
About the Code .......... 4
Requirements .......... 5
Hardware .......... 5
Software .......... 6
Internet Access .......... 6
A Disclaimer (This Is Important) .......... 6

PART I: FUNDAMENTAL CONCEPTS AND TECHNIQUES 7

1
WHAT'S IN IT FOR YOU? 9

Uncovering the Internet's True Potential .......... 9
What's in It for Developers? .......... 10
Webbot Developers Are in Demand .......... 10
Webbots Are Fun to Write .......... 11
Webbots Facilitate "Constructive Hacking" .......... 11
What's in It for Business Leaders? .......... 11
Customize the Internet for Your Business .......... 12
Capitalize on the Public's Inexperience with Webbots .......... 12
Accomplish a Lot with a Small Investment .......... 12
Final Thoughts .......... 12

© 2012 Michael Schrenk

## 2
## IDEAS FOR WEBBOT PROJECTS 15

Inspiration from Browser Limitations .......... 15
Webbots That Aggregate and Filter Information for Relevance .......... 16
Webbots That Interpret What They Find Online .......... 17
Webbots That Act on Your Behalf .......... 17
A Few Crazy Ideas to Get You Started .......... 18
Help Out a Busy Executive .......... 18
Save Money by Automating Tasks .......... 19
Protect Intellectual Property .......... 19
Monitor Opportunities .......... 20
Verify Access Rights on a Website .......... 20
Create an Online Clipping Service .......... 20
Plot Unauthorized Wi-Fi Networks .......... 21
Track Web Technologies .......... 21
Allow Incompatible Systems to Communicate .......... 21
Final Thoughts .......... 22

## 3
## DOWNLOADING WEB PAGES 23

Think About Files, Not Web Pages .......... 24
Downloading Files with PHP's Built-in Functions .......... 25
Downloading Files with fopen() and fgets() .......... 25
Downloading Files with file() .......... 27
Introducing PHP/CURL .......... 28
Multiple Transfer Protocols .......... 28
Form Submission .......... 28
Basic Authentication .......... 28
Cookies .......... 29
Redirection .......... 29
Agent Name Spoofing .......... 29
Referer Management .......... 30
Socket Management .......... 30
Installing PHP/CURL .......... 30
LIB_http .......... 30
Familiarizing Yourself with the Default Values .......... 31
Using LIB_http .......... 31
Learning More About HTTP Headers .......... 34
Examining LIB_http's Source Code .......... 35
Final Thoughts .......... 35

## 4
## BASIC PARSING TECHNIQUES 37

Content Is Mixed with Markup .......... 37
Parsing Poorly Written HTML .......... 38
Standard Parse Routines .......... 38
Using LIB_parse .......... 39
Splitting a String at a Delimiter: split_string() .......... 39
Parsing Text Between Delimiters: return_between() .......... 40

© 2012 Michael Schrenk

Parsing a Data Set into an Array: parse_array() .......... 41
Parsing Attribute Values: get_attribute() .......... 42
Removing Unwanted Text: remove() .......... 43
Useful PHP Functions .......... 44
Detecting Whether a String Is Within Another String .......... 44
Replacing a Portion of a String with Another String .......... 45
Parsing Unformatted Text .......... 45
Measuring the Similarity of Strings .......... 46
Final Thoughts .......... 46
Don't Trust a Poorly Coded Web Page .......... 46
Parse in Small Steps .......... 46
Don't Render Parsed Text While Debugging .......... 47
Use Regular Expressions Sparingly .......... 47

5
ADVANCED PARSING WITH REGULAR EXPRESSIONS 49

Pattern Matching, the Key to Regular Expressions .......... 50
PHP Regular Expression Types .......... 50
PHP Regular Expressions Functions .......... 50
Resemblance to PHP Built-In Functions .......... 52
Learning Patterns Through Examples .......... 52
Parsing Numbers .......... 53
Detecting a Series of Characters .......... 53
Matching Alpha Characters .......... 53
Matching on Wildcards .......... 54
Specifying Alternate Matches .......... 54
Regular Expressions Groupings and Ranges .......... 55
Regular Expressions of Particular Interest to Webbot Developers .......... 55
Parsing Phone Numbers .......... 55
Where to Go from Here .......... 59
When Regular Expressions Are (or Aren't) the Right Parsing Tool .......... 60
Strengths of Regular Expressions .......... 60
Disadvantages of Pattern Matching While Parsing Web Pages .......... 60
Which Are Faster: Regular Expressions or PHP's Built-In Functions? .......... 62
Final Thoughts.......... 62

6
AUTOMATING FORM SUBMISSION 63

Reverse Engineering Form Interfaces .......... 64
Form Handlers, Data Fields, Methods, and Event Triggers .......... 65
Form Handlers .......... 65
Data Fields .......... 66
Methods .......... 67
Multipart Encoding .......... 69
Event Triggers .......... 70
Unpredictable Forms .......... 70
JavaScript Can Change a Form Just Before Submission .......... 70
Form HTML Is Often Unreadable by Humans .......... 70
Cookies Aren't Included in the Form, but Can Affect Operation .......... 70
Analyzing a Form .......... 71

Final Thoughts ........................................ 74
Don't Blow Your Cover ........................................ 74
Correctly Emulate Browsers ........................................ 75
Avoid Form Errors ........................................ 75

## 7
## MANAGING LARGE AMOUNTS OF DATA 77

Organizing Data ........................................ 77
Naming Conventions ........................................ 78
Storing Data in Structured Files ........................................ 79
Storing Text in a Database ........................................ 80
Storing Images in a Database ........................................ 83
Database or File? ........................................ 85
Making Data Smaller ........................................ 85
Storing References to Image Files ........................................ 85
Compressing Data ........................................ 86
Removing Formatting ........................................ 88
Thumbnailing Images ........................................ 89
Final Thoughts ........................................ 90

# PART II: PROJECTS 91

## 8
## PRICE-MONITORING WEBBOTS 93

The Target ........................................ 94
Designing the Parsing Script ........................................ 95
Initialization and Downloading the Target ........................................ 95
Further Exploration ........................................ 100

## 9
## IMAGE-CAPTURING WEBBOTS 101

Example Image-Capturing Webbot ........................................ 102
Creating the Image-Capturing Webbot ........................................ 102
Binary-Safe Download Routine ........................................ 103
Directory Structure ........................................ 104
The Main Script ........................................ 105
Further Exploration ........................................ 108
Final Thoughts ........................................ 108

## 10
## LINK-VERIFICATION WEBBOTS 109

Creating the Link-Verification Webbot ........................................ 109
Initializing the Webbot and Downloading the Target ........................................ 109
Setting the Page Base ........................................ 110
Parsing the Links ........................................ 111
Running a Verification Loop ........................................ 111
Generating Fully Resolved URLs ........................................ 112

Webbots, Spiders, and Screen Scrapers, 2nd Edition
© 2012 Michael Schrenk

Downloading the Linked Page ... 113
Displaying the Page Status ... 113
Running the Webbot ... 114
LIB_http_codes ... 114
LIB_resolve_addresses ... 115
Further Exploration ... 115

11
SEARCH-RANKING WEBBOTS 117

Description of a Search Result Page ... 118
What the Search-Ranking Webbot Does ... 120
Running the Search-Ranking Webbot ... 120
How the Search-Ranking Webbot Works ... 120
The Search-Ranking Webbot Script ... 121
Initializing Variables ... 121
Starting the Loop ... 122
Fetching the Search Results ... 123
Parsing the Search Results ... 123
Final Thoughts ... 126
Be Kind to Your Sources ... 126
Search Sites May Treat Webbots Differently Than Browsers ... 126
Spidering Search Engines Is a Bad Idea ... 126
Familiarize Yourself with the Google API ... 127
Further Exploration ... 127

12
AGGREGATION WEBBOTS 129

Choosing Data Sources for Webbots ... 130
Example Aggregation Webbot ... 131
Familiarizing Yourself with RSS Feeds ... 131
Writing the Aggregation Webbot ... 133
Adding Filtering to Your Aggregation Webbot ... 135
Further Exploration ... 137

13
FTP WEBBOTS 139

Example FTP Webbot ... 140
PHP and FTP ... 142
Further Exploration ... 143

14
WEBBOTS THAT READ EMAIL 145

The POP3 Protocol ... 146
Logging into a POP3 Mail Server ... 146
Reading Mail from a POP3 Mail Server ... 146
Executing POP3 Commands with a Webbot ... 149
Further Exploration ... 151
Email-Controlled Webbots ... 151
Email Interfaces ... 152

## 15
## WEBBOTS THAT SEND EMAIL 153

Email, Webbots, and Spam ........ 153
Sending Mail with SMTP and PHP ........ 154
Configuring PHP to Send Mail ........ 154
Sending an Email with mail() ........ 155
Writing a Webbot That Sends Email Notifications ........ 157
Keeping Legitimate Mail out of Spam Filters ........ 158
Sending HTML-Formatted Email ........ 159
Further Exploration ........ 160
Using Returned Emails to Prune Access Lists ........ 160
Using Email as Notification That Your Webbot Ran ........ 161
Leveraging Wireless Technologies ........ 161
Writing Webbots That Send Text Messages ........ 161

## 16
## CONVERTING A WEBSITE INTO A FUNCTION 163

Writing a Function Interface ........ 164
Defining the Interface ........ 165
Analyzing the Target Web Page ........ 165
Using describe_zipcode() ........ 167
Final Thoughts ........ 169
Distributing Resources ........ 169
Using Standard Interfaces ........ 170
Designing a Custom Lightweight "Web Service" ........ 170

# PART III: ADVANCED TECHNICAL CONSIDERATIONS 171

## 17
## SPIDERS 173

How Spiders Work ........ 174
Example Spider ........ 175
LIB_simple_spider ........ 176
harvest_links() ........ 177
archive_links() ........ 178
get_domain() ........ 178
exclude_link() ........ 179
Experimenting with the Spider ........ 180
Adding the Payload ........ 181
Further Exploration ........ 181
Save Links in a Database ........ 181
Separate the Harvest and Payload ........ 182
Distribute Tasks Across Multiple Computers ........ 182
Regulate Page Requests ........ 183

© 2012 Michael Schrenk

## 18
## PROCUREMENT WEBBOTS AND SNIPERS 185

Procurement Webbot Theory ........ 186
Get Purchase Criteria ........ 186
Authenticate Buyer ........ 187
Verify Item ........ 187
Evaluate Purchase Triggers ........ 187
Make Purchase ........ 187
Evaluate Results ........ 188
Sniper Theory ........ 188
Get Purchase Criteria ........ 188
Authenticate Buyer ........ 189
Verify Item ........ 189
Synchronize Clocks ........ 189
Time to Bid? ........ 191
Submit Bid ........ 191
Evaluate Results ........ 191
Testing Your Own Webbots and Snipers ........ 191
Further Exploration ........ 191
Final Thoughts ........ 192

## 19
## WEBBOTS AND CRYPTOGRAPHY 193

Designing Webbots That Use Encryption ........ 194
SSL and PHP Built-in Functions ........ 194
Encryption and PHP/CURL ........ 194
A Quick Overview of Web Encryption ........ 195
Final Thoughts ........ 196

## 20
## AUTHENTICATION 197

What Is Authentication? ........ 197
Types of Online Authentication ........ 198
Strengthening Authentication by Combining Techniques ........ 198
Authentication and Webbots ........ 199
Example Scripts and Practice Pages ........ 199
Basic Authentication ........ 199
Session Authentication ........ 202
Authentication with Cookie Sessions ........ 202
Authentication with Query Sessions ........ 205
Final Thoughts ........ 207

## 21
## ADVANCED COOKIE MANAGEMENT 209

How Cookies Work ........ 209
PHP/CURL and Cookies ........ 211

How Cookies Challenge Webbot Design ........ 212
Purging Temporary Cookies ........ 212
Managing Multiple Users' Cookies ........ 213
Further Exploration ........ 214

## 22 SCHEDULING WEBBOTS AND SPIDERS 215

Preparing Your Webbots to Run as Scheduled Tasks ........ 216
The Windows XP Task Scheduler ........ 216
Scheduling a Webbot to Run Daily ........ 217
Complex Schedules ........ 218
The Windows 7 Task Scheduler ........ 220
Non-calendar-based Triggers ........ 223
Final Thoughts ........ 225
Determine the Webbot's Best Periodicity ........ 225
Avoid Single Points of Failure ........ 225
Add Variety to Your Schedule ........ 225

## 23 SCRAPING DIFFICULT WEBSITES WITH BROWSER MACROS 227

Barriers to Effective Web Scraping ........ 229
AJAX ........ 229
Bizarre JavaScript and Cookie Behavior ........ 229
Flash ........ 229
Overcoming Webscraping Barriers with Browser Macros ........ 230
What Is a Browser Macro? ........ 230
The Ultimate Browser-Like Webbot ........ 230
Installing and Using iMacros ........ 230
Creating Your First Macro ........ 231
Final Thoughts ........ 237
Are Macros Really Necessary? ........ 237
Other Uses ........ 237

## 24 HACKING IMACROS 239

Hacking iMacros for Added Functionality ........ 240
Reasons for Not Using the iMacros Scripting Engine ........ 240
Creating a Dynamic Macro ........ 241
Launching iMacros Automatically ........ 245
Further Exploration ........ 247

## 25 DEPLOYMENT AND SCALING 249

One-to-Many Environment ........ 250
One-to-One Environment ........ 251

Many-to-Many Environment .......... 251
Many-to-One Environment .......... 252
Scaling and Denial-of-Service Attacks .......... 252
    Even Simple Webbots Can Generate a Lot of Traffic .......... 252
    Inefficiencies at the Target .......... 252
    The Problems with Scaling Too Well .......... 253
Creating Multiple Instances of a Webbot .......... 253
    Forking Processes .......... 253
    Leveraging the Operating System .......... 254
    Distributing the Task over Multiple Computers .......... 254
Managing a Botnet .......... 255
    Botnet Communication Methods .......... 255
Further Exploration .......... 262

## PART IV: LARGER CONSIDERATIONS 263

### 26 DESIGNING STEALTHY WEBBOTS AND SPIDERS 265

Why Design a Stealthy Webbot? .......... 265
    Log Files .......... 266
    Log-Monitoring Software .......... 269
Stealth Means Simulating Human Patterns .......... 269
    Be Kind to Your Resources .......... 269
    Run Your Webbot During Busy Hours .......... 270
    Don’t Run Your Webbot at the Same Time Each Day .......... 270
    Don’t Run Your Webbot on Holidays and Weekends .......... 270
    Use Random, Intra-fetch Delays .......... 270
Final Thoughts .......... 270

### 27 PROXIES 273

What Is a Proxy? .......... 273
Proxies in the Virtual World .......... 274
Why Webbot Developers Use Proxies .......... 274
    Using Proxies to Become Anonymous .......... 274
    Using a Proxy to Be Somewhere Else .......... 277
Using a Proxy Server .......... 277
    Using a Proxy in a Browser .......... 278
    Using a Proxy with PHP/CURL .......... 278
Types of Proxy Servers .......... 278
    Open Proxies .......... 279
    Tor .......... 281
    Commercial Proxies .......... 282
Final Thoughts .......... 283
    Anonymity Is a Process, Not a Feature .......... 283
    Creating Your Own Proxy Service .......... 283

**28**
**WRITING FAULT-TOLERANT WEBBOTS 285**

- Types of Webbot Fault Tolerance ... 286
  - Adapting to Changes in URLs ... 286
  - Adapting to Changes in Page Content ... 291
  - Adapting to Changes in Forms ... 292
  - Adapting to Changes in Cookie Management ... 294
  - Adapting to Network Outages and Network Congestion ... 294
- Error Handlers ... 295
- Further Exploration ... 296

**29**
**DESIGNING WEBBOT-FRIENDLY WEBSITES 297**

- Optimizing Web Pages for Search Engine Spiders ... 297
  - Well-Defined Links ... 298
  - Google Bombs and Spam Indexing ... 298
  - Title Tags ... 298
  - Meta Tags ... 299
  - Header Tags ... 299
  - Image alt Attributes ... 300
- Web Design Techniques That Hinder Search Engine Spiders ... 300
  - JavaScript ... 300
  - Non-ASCII Content ... 301
- Designing Data-Only Interfaces ... 301
  - XML ... 301
  - Lightweight Data Exchange ... 302
  - SOAP ... 305
  - REST ... 306
- Final Thoughts... 307

**30**
**KILLING SPIDERS 309**

- Asking Nicely ... 310
  - Create a Terms of Service Agreement ... 310
  - Use the robots.txt File ... 311
  - Use the Robots Meta Tag ... 312
- Building Speed Bumps ... 312
  - Selectively Allow Access to Specific Web Agents ... 312
  - Use Obfuscation ... 313
  - Use Cookies, Encryption, JavaScript, and Redirection ... 313
  - Authenticate Users ... 314
  - Update Your Site Often ... 314
  - Embed Text in Other Media ... 314
- Setting Traps ... 315
  - Create a Spider Trap ... 315
  - Fun Things to Do with Unwanted Spiders ... 316
- Final Thoughts ... 316

## 31
## KEEPING WEBBOTS OUT OF TROUBLE 317

It's All About Respect .......... 318
Copyright .......... 319
Do Consult Resources .......... 319
Don't Be an Armchair Lawyer .......... 319
Trespass to Chattels .......... 322
Internet Law .......... 324
Final Thoughts .......... 325

## A
## PHP/CURL REFERENCE 327

Creating a Minimal PHP/CURL Session .......... 327
Initiating PHP/CURL Sessions .......... 328
Setting PHP/CURL Options .......... 328
CURLOPT_URL .......... 329
CURLOPT_RETURNTRANSFER .......... 329
CURLOPT_REFERER .......... 329
CURLOPT_FOLLOWLOCATION and CURLOPT_MAXREDIRS .......... 329
CURLOPT_USERAGENT .......... 330
CURLOPT_NOBODY and CURLOPT_HEADER .......... 330
CURLOPT_TIMEOUT .......... 331
CURLOPT_COOKIEFILE and CURLOPT_COOKIEJAR .......... 331
CURLOPT_HTTPHEADER .......... 331
CURLOPT_SSL_VERIFYPEER .......... 332
CURLOPT_USERPWD and CURLOPT_UNRESTRICTED_AUTH .......... 332
CURLOPT_POST and CURLOPT_POSTFIELDS .......... 332
CURLOPT_VERBOSE .......... 333
CURLOPT_PORT .......... 333
Executing the PHP/CURL Command .......... 333
Retrieving PHP/CURL Session Information .......... 334
Viewing PHP/CURL Errors .......... 334
Closing PHP/CURL Sessions .......... 335

## B
## STATUS CODES 337

HTTP Codes .......... 337
NNTP Codes .......... 339

## C
## SMS GATEWAYS 341

Sending Text Messages .......... 342
Reading Text Messages .......... 342
A Sampling of Text Message Email Addresses .......... 342

## INDEX 345