

Contents

Acknowledgments	v
Foreword by Dare Obasanjo	xxvii
Foreword by Greg Reinacker	xxix
Introduction	xxxix

Part I: Understanding the Issues and Taking Control **1**

Chapter 1: Managing the Flow of Information: A Crucial Skill **3**

New Vistas of Information Flow	4
The Information Well and Information Flow	4
The Information Well	4
The Information Flow	5
The Information Flood	5
Managing Information	5
What Do You Want to Do with Information?	6
Browse and Discard	6
Read	6
Study and Keep	6
Taking Control of Information	7
Determining What Is Important to You	7
Avoiding Irrelevant Information	7
Determining the Quality of Information	8
Information Flows Other Than the Web	8
Books	8
Magazines	9
Newspapers	9
Broadcast Media	9
The Web and Information Feeds	10
New Information Opportunities	10
New Information Problems	10
The Need to Keep Up-to-Date	11
Distractions	11
Summary	11
Exercise	11

Chapter 2: Where Did Information Feeds Start?	13
The Nature of the Web	13
HTTP	14
HTML	14
XML	14
Polling the Web	14
Precursors to RSS	17
MCF and HotSauce	17
Netscape Channels	17
The Microsoft Channel Definition Format	18
RSS: An Acronym with Multiple Meanings	18
RSS 0.9	19
RSS 0.91	19
RSS 1.0	19
RSS 0.92, 0.93, and 0.94	19
RSS 2.0	19
Use of RSS and Atom Versions	20
Summary	21
Exercises	21
Chapter 3: The Content Provider Viewpoint	23
Why Give Your Content Away?	23
Selling Your Content	24
Creating Community	25
Content to Include in a Feed	25
The Importance of Item Titles	25
Internal Links	25
One Feed or Several?	26
Structuring Topics	26
Blogging Tools	26
Wikis	28
Publicizing Your Information Feed	28
Deciding on a Target Audience	28
Registering with Online Sites	28
How Information Feeds Can Affect Your Site's Visibility	29
Advertisements and Information Feeds	29
Power to the User?	30
Filtering Out Advertisements	30
Summary	31
Exercise	31

Chapter 4: The Content Recipient Viewpoint	33
Access to Information	34
Convenience	34
Timeliness of Access	35
Timeliness and Data Type	35
Timeliness and Data Source	35
Newsreaders and Aggregators	36
Aggregating for Intranet Use	36
Security and Aggregators	36
Directories	37
Finding Information about Interesting Feeds	38
The Known Sites Approach	38
The Blogroll Approach	39
The Directory Approach	41
Filtering Information Feeds	42
Filtering Blogs	42
Summary	42
Chapter 5: Storing, Retrieving, and Exporting Information	43
Storing Information	44
Storing URLs	44
Storing Content	44
Storing Static Files	44
Relational Databases	45
RDF Triple Stores	46
Two Examples of Longer-Term Storage	46
Onfolio 2.0	46
OneNote 2003	50
Retrieving Information	52
Search in Onfolio 2.0	53
Search in OneNote 2003	54
Exporting Information	55
Exporting in Onfolio 2.0	55
Exporting in OneNote 2003	55
Summary	56

Part II: The Technologies 57

Chapter 6: Essentials of XML 59

What Is XML?	60
XML Declaration	62
XML Names	62
XML Elements	62
XML Attributes	63
XML Comments	64
Predefined Entities	64
Character References	65
XML Namespaces	66
HTML, XHTML, and Feed Autodiscovery	68
Summary	70
Exercises	70

Chapter 7: Atom 0.3 71

Introducing Atom 0.3	72
The Atom 0.3 Specification	72
The Atom 0.3 Namespace	72
Atom 0.3 Document Structure	72
The feed Element	73
The title Element	74
The link Element	74
The author Element	74
The id Element	74
The generator Element	75
The copyright Element	75
The info Element	75
The modified Element	75
The tagline Element	75
The entry Element	76
Using Modules with Atom 0.3	78
Summary	79
Exercises	79

Chapter 8: RSS 0.91 and RSS 0.92 81

What Is RSS 0.91?	82
The RSS 0.91 Document Structure	82
The rss Element	83
The channel Element	83

Required Child Elements of channel	83
Optional Child Elements of channel	84
The image Element	85
The textInput Element	85
The item Element	86
Introducing RSS 0.92	87
The RSS 0.92 Document Structure	87
New Child Elements of the item Element	87
The cloud Element	88
Summary	88
Exercises	88
 Chapter 9: RSS 1.0	 89
 What Is RSS 1.0?	 89
RSS 1.0 Is RDF	90
RSS 1.0 Uses XML Namespaces	90
RSS 1.0 Uses Modules	91
The RSS 1.0 Document Structure	91
The channel Element	93
The items Element	93
The image Element	94
The item Element	94
The textinput Element	95
Some Real-World RSS 1.0	95
Summary	98
Exercise	99
 Chapter 10: RSS 1.0 Modules	 101
 RSS Modules	 101
The RSS 1.0 Official Modules	102
RDF Parser Compatibility	103
Module Compatibility	103
The Content Module	103
The content:items Element	104
The content:item Element	104
The Dublin Core Module	105
The Syndication Module	107
Including Other Modules in RSS 1.0 Feed Documents	108
Adding the Namespace Declaration	108
The Admin Module	108
The FOAF Module	108
Summary	109

Chapter 11: RDF: The Resource Description Framework **111**

What Is RDF? **112**

Simple Metadata 112

Simple Facts Expressed in RDF 112

The RDF Triple 113

Using URIs in RDF 113

Directed Graphs 114

How RDF and XML Are Related 115

What RDF Is Used For 115

RDF and RSS 1.0 116

RDF Vocabularies **117**

Dublin Core 117

FOAF 117

RDF Toolkits **118**

Jena 118

Redland 118

RDFLib 118

rdftdata.org 118

Summary **118**

Chapter 12: RSS 2.0: Really Simple Syndication **121**

What Is RSS 2.0? **121**

XML Namespaces in RSS 2.0 122

New Elements in RSS 2.0 122

The RSS 2.0 Document Structure **122**

The rss Element 122

The channel Element 123

The image Element 124

The cloud Element 125

The textinput Element 125

The item Element 126

An Example RSS 2.0 Document 126

RSS 2.0 Extensions **127**

The blogChannel RSS Module 127

Summary **128**

Chapter 13: Looking Forward to Atom 1.0 **129**

Why Another Specification? **130**

Aiming for Clarity 130

Archiving Feeds 130

The RDF Issue	130
The Standards Body Aspect	131
What Is Atom?	131
Further Information about Atom 1.0	131
Atom 1.0 Document Structure	132
The XML Declaration	132
The feed Element	132
The head Element	133
The entry Element	133
Extending Atom 1.0 Documents	134
Other Aspects of Atom 1.0	134
The Atom Publishing Protocol	134
An Atom Entry Document	135
The Atom Notification Protocol	135
Atom Feed Autodiscovery	136
Summary	136
 Part III: The Tools	 137
 Chapter 14: Feed Production Using Blogging Tools	 139
Movable Type	140
WordPress	146
Blojsom	147
Summary	154
 Chapter 15: Aggregators and Similar Tools	 155
Overview of Desktop Aggregators	156
Managing Subscriptions	156
Updating Feeds	156
Viewing Web Pages	156
Individual Desktop Aggregators	157
Abilon	157
ActiveRefresh	159
Aggreg8	160
BlogBridge	161
Firefox Live Bookmarks	162
Google News Alerts	162
Habari Xenu	163
NewzCrawler	164
Newsgator	165
Omea	166

Contents

Onfolio	168
RSS Bandit	169
Thunderbird 1.0	170
Summary	171
Exercise	172
Chapter 16: Long-Term Storage of Information	173
Choosing an Approach to Long-Term Storage	174
The Use Case for Long-Term Storage	174
Storage Options	174
Choosing Information to Store	174
Determining Who Will Access the Data	175
Ease of Storage	175
Choosing a Backup Strategy	176
Characteristics of Long-Term Storage	176
Scalability	176
Availability	176
Software to Support Long-Term Storage	177
Onfolio	177
Microsoft OneNote	178
Microsoft SQL Server 2005	179
Summary	180
Chapter 17: Online Tools	181
Advantages and Disadvantages of Online Tools	181
Registration	182
Interface Issues	182
Online Tools When Mobile	182
Cost of Online Tools	182
Stability of Access	183
Performance	183
Notifications	183
Cookies	183
Using Multiple Machines	183
Making a Choice	184
Choosing Between Individual Online Tools	184
Bloglines	184
Feedster	185
Furl	186
Google News Alerts	187
Kinja	188

Newsgator Online	190
Rollup.org	191
Syndic8.com	192
Technorati	193
PubSub.com	194
Summary	194
Chapter 18: Language-Specific Developer Tools	195
Python Tools	195
PyXML	196
RSS.py	196
Universal Feed Parser	196
xpath2rss	196
Chumpologica	196
PHP Tools	196
lastRSS	197
MagpieRSS	197
Java Tools	197
ROME	197
Jena	197
The Redland RDF Application Framework	198
XSLT Tools	198
Perl Tools	198
XML::RSS	199
Support for XSLT	199
The rss2html Module	199
The Regular Expressions Approach	200
Miscellaneous Tools	200
Useful Sources	200
Summary	201
Part IV: The Tasks	203
Chapter 19: Systematic Overview	205
Before You Start	206
Python	206
Recommended Downloads	206
States and Messages: Exchanges Between Client and Server	207
Resources and Representations	207
States and Statelessness	208
RPC vs. Document-Oriented Messaging	208

Contents

Communication Layers	210
Web Languages	210
The HTTP Client	210
<i>Try It Out: Client Requests</i>	210
Server Response	212
Acronym City	214
Application Layers	214
Server Producer: Producing Feeds	214
The HTTP Server	214
Sockets and Ports	215
<i>Try It Out: Serving an RSS Feed</i>	216
Getting Flexible	219
<i>Try It Out: Improved Version with Python</i>	220
Serving and Producing	221
Client Consumer: Viewing Feeds	222
Single-Page View	223
Three-Pane View	223
Syndication, the Protocol	224
Polling	225
Reducing the Load	226
<i>Try It Out: HTTP Client with gzip Compression</i>	227
The Trade Off	230
Client Producer: Blogging APIs	230
XML-RPC	231
RESTful Communications	232
Server Consumer	233
Weblog Comments	233
Trackback	233
Architectural Approaches	234
Summary	234
Exercise	234
 Chapter 20: Modeling Feed Data	 237
Model or Syntax?	237
The Conceptual Feed	238
The Feed-Oriented Model	238
Source or Styled?	239
<i>Try It Out: RSS with CSS</i>	243
Drawbacks of a Feed-Oriented Model	245
Item-Oriented Models	247
Twenty-First Century Press Clippings	247
Microcontent and Memes	247

Including Item Provenance	250
Element Order	252
Common Features Among Formats	252
Entities and Relationships	252
Entities in Feeds	253
Relationships in Feeds	253
Identity	254
Uniform and Unique	254
GUID	255
Crisis, What Crisis?	255
Parallel Elements	256
Groups of Feeds	256
Extending Structures	256
An XML Model	257
<i>Try It Out: Reverse Engineering a Schema</i>	259
An Object-Oriented Model: XML to Java	260
<i>Try It Out: Creating a Feed</i>	262
Reading a Feed	264
<i>Try It Out: A Demo Channel</i>	264
Problems of Autogeneration	266
The RDF Models	266
There Is No Syntax!	266
<i>Try It Out: Yet Another Syntax</i>	268
RSS 1.0 RDF Schema	268
Groups of Feeds in RDF	272
A Relational Model	274
Syndication Applications of Databases	274
Content Storage	274
Metadata Management	274
User Tracking	274
Query and Search	275
Requirements for Storing Feed Data	275
Storing Items	275
Feed Lists and State Management	275
The Relational Model, Condensed	275
Relations and Tables	275
Tables, Take Two	276
Tables, Take Three	277
Keys	279
Extensibility in Databases	279
Tables, Take Four	280
Summary	281
Exercises	282

Contents

Chapter 21: Storing Feed Data	285
The Document Object Model	285
A (Forgetful) DOM-Based Weblog	289
XML in PHP	290
Source Code	290
Representing Feed Lists in OPML	301
Creating a SQL Database for Feeds	302
Download Requirements	302
<i>Try It Out: Working with SQL</i>	302
RDF in Relational Databases	308
A Minimal Triple Store	308
Filling the Store	310
RSS 1.0 Parser	310
Connecting to a Database	313
<i>Try It Out: Interfacing with a Database</i>	313
Aggregating Triples	316
Code Overview	316
Store Wrapper	316
Connecting Parser to Store	321
The Joy of SQL VIEWS	322
<i>Try It Out: Using the Triplestore</i>	323
Customizing for RSS	324
Utility Methods	324
A View of Items	325
Starting with a Blogroll	326
<i>Try It Out: Aggregating with a Triple Store</i>	327
Pros and Cons of Using an RDF Store	329
Summary	330
Exercises	330
Chapter 22: Consuming Feeds	333
Consumer Strategies	333
Parsing Data	335
Scanning for Tokens	335
Parsing for Structure	336
Interpretation for Semantics	336
XML Parsing	336
Semantics in Feeds	337
A More Practical Consumer	337
All Kinds of XML	338
Markup in Content	339
Other Approaches	339

Reading Feed Data	341
The Polite Client	341
Push and Pull	342
HTTP Status Codes	342
Timing Issues	344
XML-Level Timing Hints	344
HTTP Compression	345
Conditional GET	345
A Reasonably Polite Client	346
<i>Try It Out: Downloading a Feed</i>	350
Parsing XML with SAX	356
<i>Try It Out: Reading an RDF/XML Channel List</i>	359
Feed/Connection Management Subsystem	362
A Fetcher That Fetches the Feeds	364
<i>Try It Out: Running the Consumer</i>	375
Implementation Notes	379
Summary	379
Exercises	380
 Chapter 23: Parsing Feeds	 381
A Target Model	382
Parsing Feeds as XML	384
<i>Try It Out: Parsing RSS with SAX</i>	387
The Trouble with Feeds	392
The Syndication Stack	393
The Postel Paradox	393
The Bozo Bit	394
<i>Try It Out: Draconian Parsing</i>	394
Tag Soup	395
Soup Parser	395
Different Parser, Same Handler	406
<i>Try It Out: Parsing Tag Soup</i>	408
HTTP, MIME, and Encoding	409
Encoding Declaration in XML	410
Encoding Declaration in HTTP	410
The Bad News	411
The Good News	411
Gluing Together the Consumer	411
<i>Try It Out: A Consumer Subsystem</i>	417
One More Thing...	419
A Universal Feed Consumer?	420
Summary	420
Exercises	421
	xix

Contents

Chapter 24: Producing Feeds **423**

Content Management Systems **423**

Content and Other Animals	424
Metadata	424
Applications and Services	424
Accessibility	424
Device Independence	425
Personalization	425
Security	425
Inter-System Communication	425
Legacy Integration	426
Managing Content: Many Tiers	426

SynWiki: A Minimal CMS **427**

What Is PHP?	427
Installing PHP	428
Open Web	428
Basic Operation	428
System Operation	431
System Requirements	432
Files	432
Data Store	432
<i>Try It Out: Interacting with the Database</i>	435
Clients Are Clients	439
Utilities	439
Authoring Page	442
Viewing Page	444
Making Modules	447
URL Rewriting	448
Blog Module	448
Reverse-Chrono Here We Go	450
Producing RSS 1.0	452
Alternative Techniques	455
Steps in Producing RSS from a Content Database	456

Summary **456**

Exercise **456**

Chapter 25: Queries and Transformations **459**

Paths and Queries **459**

XQuery Runner	460
Source Code	460
<i>Try It Out: XPath Queries</i>	465

XQuery and Syndication Formats	467
<i>Try It Out: Querying Live Feeds</i>	467
XQuery Is More Than XPath	469
Getting Selective	470
<i>Try It Out: A Selective XQuery</i>	470
Introducing XSLT	472
Templates	472
More Templates	474
XSLT Processors	476
<i>Try It Out: Transforming XML to Text</i>	479
Output Format	481
XSLT in Syndication	481
Transforming to RSS	481
Transforming from RSS	483
Going Live	485
Graphic Display: SVG in a Very Small Nutshell	486
Graphic RSS	489
Handling Text Blocks	489
Multiformat Handling	494
SVG Output	496
Summary	502
Exercise	502
 Chapter 26: The Blogging Client	 503
 Talking to the Server with a Variety of Clients	 503
Browser Scripts	504
The Fat Client	505
The Language of HTTP	506
HTML's Legacy	506
Authentication	507
HTTP Basic Authentication	508
HTTP Digest	508
HTTPS	509
Base64 Encoding	509
Hashes and Digests	509
Nonces	509
Protocols for Blogging	509
The Blogger API and XML-RPC	510
Command-Line Tool	513
Handling Account Details	515
Minimal XML-RPC Client	518
<i>Try It Out: Running the Blogger Client</i>	520

Contents

Adding Post Support	523
MetaWeblog and other APIs	525
Atom Protocol	525
Atom Protocol, SOAP and WSDL	526
Atom Protocol — RESTful HTTP	529
Building Your Own Atomic Device	531
Other Formats, Transports, and Protocols	535
Jabber/XMPP	535
WebDAV	535
Annotea	536
InterWiki	536
Usenet, E-mail, and Friends	536
Summary	536
Exercise	536
 Chapter 27: Building Your Own Planet	 537
Why Aggregate?	537
Planet Core	538
Implementing a Planet	539
Extending the Entry	540
Entry Collection	543
Top-Level Runner	546
Serializing Output	550
Templates	555
Running the Planet	556
<i>Try It Out: Aggregating the Planet</i>	557
Summary	560
 Chapter 28: Building a Desktop Aggregator	 561
Desktop Aggregator Overview	562
System Construction	563
Feed Data Model	563
Event Notification	568
User Interface	569
DOM-Based Implementation	575
Using an RDF Model	579
Automating Updates	581
Summary	582

Chapter 29: Social Syndication	583
What Is Society on the Web?	583
Social Software and Syndication	584
Avoiding Overload	584
Counting Atoms	584
Friends as Filters	585
A Personal Knowledgebase	585
Kowari RDF Store	585
RDF Queries	585
<i>Try It Out: An iTQL Session</i>	585
Going Programmatic	589
<i>Try It Out: Kowari Blogroll Channels</i>	591
In an Ideal World...	595
Fetching and Normalizing	595
RDF to RDF	596
XML to RDF	596
Tag Soup to RDF	596
Coupling the Store	596
Supervisor	597
<i>Try It Out: Fetching and Normalizing Feeds</i>	604
Loading the Data Store	605
<i>Try It Out: Adding Normalized Data to Store</i>	607
Summary	608
Exercise	608
 Chapter 30: Additional Content	 609
Syndicated Broadcasting	609
Publishing Audio	610
Digital Audio in Brief	610
Compression Issues	611
An Imp for the Web	612
User Interface of Imp	612
Application Classes	613
AudioTool Interface	614
Audio in Java	614
Lines and Audio Formats	615
Audio Recorder Class	615
Audio Player Class	619
Imp User Interface	621

Contents

Publishing to a Feed	628
Syndication and Enclosures	628
Supporting the Enclosure Element	630
Serializing an Entry	631
Publisher Class	631
FTP Upload	635
Building the Imp	636
<i>Try It Out: Running the Imp</i>	637
Troubleshooting	637
Developments in Syndicated Media	638
Summary	639
 Chapter 31: Loose Ends, Loosely Coupled	 641
 Finding Feeds	 642
Autodiscovery	642
Practical Approaches	643
Consommé Code	643
<i>Try It Out: Parsing HTML</i>	646
From Browser to Aggregator	647
A Scheme Handler	648
Registering a URL Scheme in Windows	648
Writing a Bookmarklet	648
<i>Try It Out: Using a Bookmarklet</i>	649
Registering a URI Scheme for Real	650
Using MIME Types	651
GRDDL	651
Robots, Crawlers, Harvesters, and Scutters	652
Inter-Site Communication	654
Clouds	654
Weblog Pings	654
Publish/Subscribe Innovations	656
Feed Deltas	656
Trackback	657
Right Back to the Track	657
More Autodiscovery	660
Pulling Out the Comments	661
Defining Extension Modules	664
RSS, Atom, and XML Namespaces	665
Expressing Things Clearly	666
Other Resources	667
Summary	667

Chapter 32: What Lies Ahead in Information Management	669
Filtering the Flood of Information	670
Flexibility in Subscribing	670
Finding the Best Feeds	670
Filtering by Personality	670
Finding the Wanted Non-Core Information	671
Point to Point or Hub and Spoke	671
New Technologies	671
A User-Centric View of Information	672
Information for Projects	672
Automated Publishing Using Information Feeds	672
Ownership of Data	672
How Users Can Best Access Data	673
How to Store Relevant Information	673
How to Retrieve Information from Storage	673
Revisions of Information Feed Formats	674
Non-Text Information Delivered Through Information Feeds	674
Product Support Uses	674
Educational Uses	675
Summary	675
Appendix A: Answers to Exercises	677
Appendix B: Useful Online Resources	697
Appendix C: Glossary	701
Index	705