



**INSTANT**

Short | Fast | Focused

# Web Scraping with Java

Build simple scrapers or vast armies of Java-based bots to untangle and capture the Web

Ryan Mitchell

**[PACKT]**  
PUBLISHING

# Instant Web Scraping with Java

Build simple scrapers or vast armies of Java-based bots  
to untangle and capture the Web

**Ryan Mitchell**

**[PACKT]**  
PUBLISHING

BIRMINGHAM - MUMBAI

# **Instant Web Scraping with Java**

Copyright © 2013 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: August 2013

Production Reference: 1230813

Published by Packt Publishing Ltd.  
Livery Place  
35 Livery Street  
Birmingham B3 2PB, UK.

ISBN 978-1-84969-688-3

[www.packtpub.com](http://www.packtpub.com)

# Credits

**Author**

Ryan Mitchell

**Proofreader**

Jonathan Todd

**Reviewers**

Benjamin Hill

Sumant Kumar Raja

**Graphics**

Abhinash Sahu

**Acquisition Editor**

Mary Nadar

**Production Coordinator**

Conidon Miranda

**Commissioning Editors**

Sharvari Tawde

Ameya Sawant

**Cover Work**

Conidon Miranda

**Cover Image**

Sheetal Aute

**Technical Editor**

Akashdeep Kundu

**Project Coordinator**

Joel Goveya

# About the Author

**Ryan Mitchell** has ten years of programming experience, including Java, C, Perl, PHP, and Python. In addition to “traditional” programming, she specializes in web technologies, with three years of Drupal development experience, and is Sitecore developer certified.

She graduated from Olin College of Engineering and is currently studying at the Harvard University Extension School for a Masters in Software Engineering.

In addition to academic life, she currently works at Velir Studios as a Web Systems Analyst, and has also worked as a developer for Harvard University and Abine Inc.

# About the Reviewers

**Benjamin Hill** is a product manager working in online and mobile applications. His background includes everything from a bootstrapped crowdsourcing startup in 2006 to companies as big as Comcast and eBay. During that time he has learned a variety of techniques to automate websites that don't yet provide APIs.

**Sumant Kumar Raja** is an entrepreneur, integration architect, and independent consultant with seven years of experience working for global clients. He has worked on a wide range of complex technical environments including J2EE and SAP Process Integration. He is a Certified ScrumMaster.

He is director of Denarit Consulting Ltd. UK. In past, he has worked in various roles with Accenture for clients in the UK and the U.S.

---

I am thankful to my family for their unconditional support.

I would like to extend my thanks to Packt Publishing for giving me the opportunity to review this book.

---

# www.PacktPub.com

## **Support files, eBooks, discount offers and more**

You might want to visit [www.PacktPub.com](http://www.PacktPub.com) for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at [www.PacktPub.com](http://www.PacktPub.com) and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at [service@packtpub.com](mailto:service@packtpub.com) for more details.

At [www.PacktPub.com](http://www.PacktPub.com), you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<http://PacktLib.PacktPub.com>

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

## **Why Subscribe?**

- ▶ Fully searchable across every book published by Packt
- ▶ Copy and paste, print and bookmark content
- ▶ On demand and accessible via web browser

## **Free Access for Packt account holders**

If you have an account with Packt at [www.PacktPub.com](http://www.PacktPub.com), you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

# Table of Contents

<b>Preface</b>	<b>1</b>
<b>Instant Web Scraping with Java</b>	<b>7</b>
Setting up your Java Environment (Simple)	9
Writing and executing HelloWorld.java (Simple)	12
Writing a simple scraper (Simple)	14
Writing more complicated scraper (Intermediate)	18
Handling errors (Simple)	22
Writing robust, scalable code (Advanced)	25
Persisting data (Advanced)	33
Writing tests (Intermediate)	36
Going undercover (Intermediate)	39
Submitting a basic form (Advanced)	42
Scraping Ajax Pages (Advanced)	46
Faster scraping through threading (Intermediate)	50
Faster scraping with RMI (Advanced)	54