Understand Your Data and Be More Productive

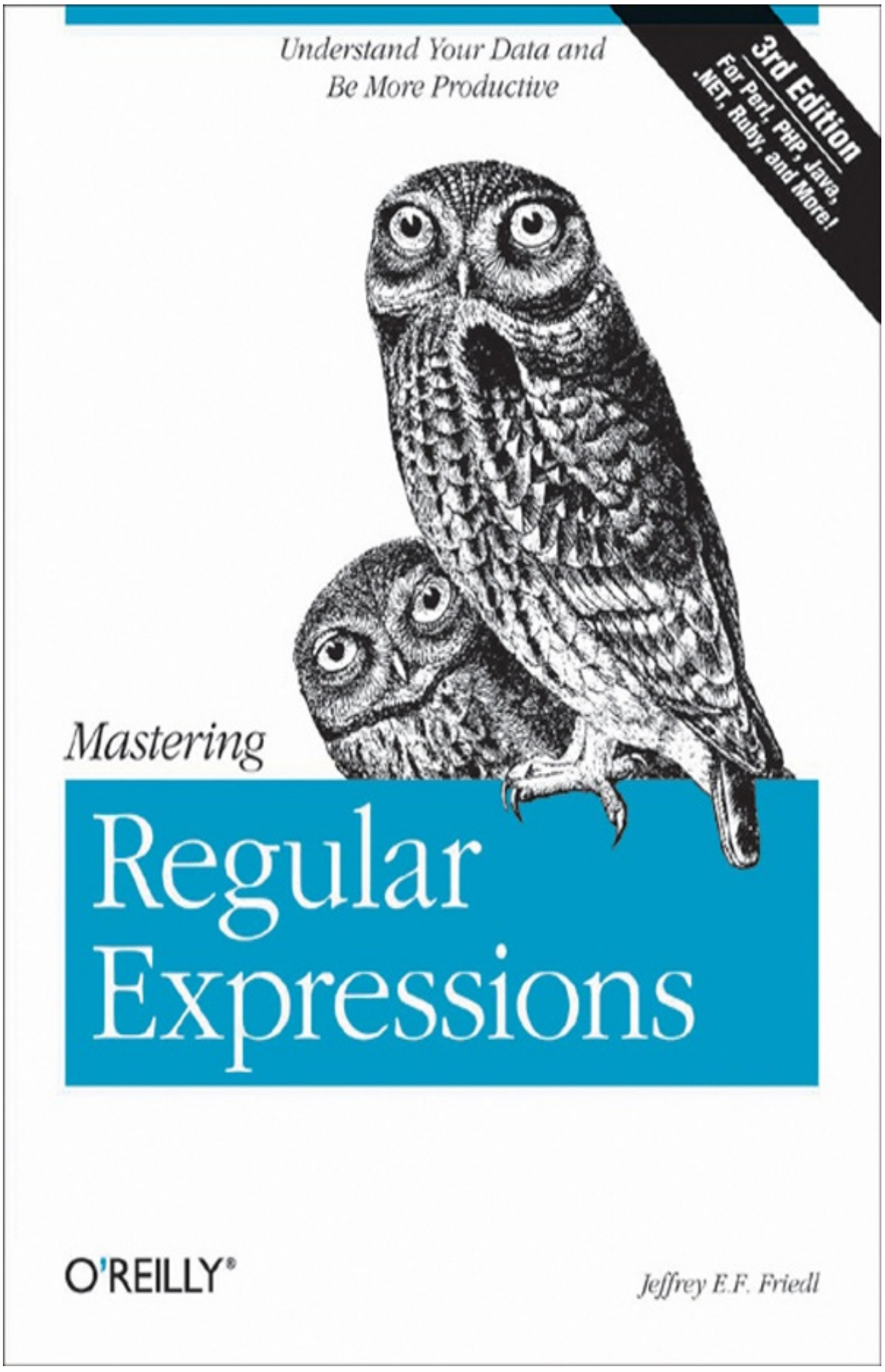
3rd Edition
For Perl, PHP, Java, .NET, Ruby, and More!

Mastering

Regular Expressions

O'REILLY®

Jeffrey E.F. Friedl


Understand Your Data and
Be More Productive
3rd Edition
For Perl, PHP, Java,
.NET, Ruby, and More!
Mastering
Regular
Expressions
O'REILLY®
Jeffrey E.F. Friedl

Programming

O'REILLY®

Mastering Regular Expressions


This book will get you up to speed on the productivity secrets that will make your life easier: regular expressions. Well-crafted regular expressions can reduce hours of tedious labor to a 15-second solution. Now a standard feature in a wide range of languages and popular tools—including Perl, PHP, Java, Python, Ruby, MySQL, VB.NET, and C# (and any language using the .NET Framework)—regular expressions will allow you to code complex and subtle text processing that you never imagined could be automated.

Mastering Regular Expressions, Third Edition, now includes material on PHP and its powerful regular expressions. This edition has been updated throughout to reflect advances in other languages, including expanded in-depth coverage of Sun's java.util.regex, with special attention to the many differences between Java 1.4.2 and Java 1.5/1.6.

Topics include:

- A comparison of features among many languages and tools
- How a regular expression engine works
- Optimization (major savings available here!)
- Matching only what you want, not what you don't
- Sections and chapters on individual languages

Written in the lucid, entertaining lone that makes a complex topic become crystal clear to programmers, and filled with solutions to difficult real-world problems, *Mastering Regular Expressions*, Third Edition, offers a wealth of information that you can put to immediate use to craft elegant, time-saving solutions to a wide range of issues.

> ***"If you use regular expressions as part of your professional work (even if you already have a good book on whatever language you're programming in) I would strongly recommend this book to you."***
>
> —Dr Chris Brown, Linux Format

> **"Mastering Regular Expressions** ***is the definitive guide to the subject, and an outstanding resource that belongs on every programmer's bookshelf. Ten out of Ten Horseshoes."***

—Jason Menard, Java Ranch

"There isn't a better (or more useful) book available on regular expressions."

—Zak Greant, Planet PHP

www.oreilly.com

ISBN 0-596-52812-4 US $44.99 CAN $58.99

9 780596 528126 6 36920 52812 8

Mastering Regular Expressions

Third Edition

Jeffrey E. F. Friedl

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

Mastering Regular Expressions, Third Edition
by Jeffrey E. F. Friedl

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly Media, Inc. books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*safari.oreilly.com*). For more information contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Andy Oram

Production Editor: Jeffrey E. F. Friedl

Cover Designer: Edie Freedman

Printing History:

January 1997:	First Edition.
July 2002:	Second Edition.
August 2006:	Third Edition.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Mastering Regular Expressions*, the image of owls, and related trade dress are trademarks of O'Reilly Media, Inc. Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

This book uses RepKover™, a durable and flexible lay-at binding.

ISBN: 0-596-52812-4
[M]

FOR Fumie 文枝

For putting up with me.

And for the years I worked on this book,

for putting up without me.

Table of Contents

Preface

This book is about a powerful tool called “regular expressions”. It teaches you how to use regular expressions to solve problems and get the most out of tools and languages that provide them. Most documentation that mentions regular expressions doesn’t even begin to hint at their power, but this book is about *mastering* regular expressions.

Regular expressions are available in many types of tools (editors, word processors, system tools, database engines, and such), but their power is most fully exposed when available as part of a programming language. Examples include Java and JScript, Visual Basic and VBScript, JavaScript and ECMAScript, C, C++, C#, elisp, Perl, Python, Tcl, Ruby, PHP, *sed*, and *awk*. In fact, regular expressions are the very heart of many programs written in some of these languages.

There’s a good reason that regular expressions are found in so many diverse languages and applications: they are extremely powerful. At a low level, a regular expression describes a chunk of text. You might use it to verify a user’s input, or perhaps to sift through large amounts of data. On a higher level, regular expressions allow you to master your data. Control it. Put it to work for you. To master regular expressions is to master your data.

The Need for This Book

I finished the first edition of this book in late 1996, and wrote it simply because there was a need. Good documentation on regular expressions just wasn’t available, so most of their power went untapped. Regular-expression documentation was available, but it centered on the “low-level view.” It seemed to me that they were analogous to showing someone the alphabet and expecting them to learn to speak.

The five and a half years between the first and second editions of this book saw the popular rise of the Internet, and, perhaps more than just coincidentally, a considerable expansion in the world of regular expressions. The regular expressions of almost every tool and language became more powerful and expressive. Perl, Python, Tcl, Java, and Visual Basic all got new regular-expression backends. New languages with regular expression support, like PHP, Ruby, and C#, were developed and became popular. During all this time, the basic core of the book — how to truly understand regular expressions and how to get the most from them — remained as important and relevant as ever.

Yet, the first edition gradually started to show its age. It needed updating to reflect the new languages and features, as well as the expanding role that regular expressions played in the Internet world. It was published in 2002, a year that saw the landmark

releases of `java.util.regex`, Microsoft's .NET Framework, and Perl 5.8. They were all covered fully in the second edition. My one regret with the second edition was that it didn't give more attention to PHP. In the four years since the second edition was published, PHP has only grown in importance, so it became imperative to correct that deficiency.

This third edition features enhanced PHP coverage in the early chapters, plus an all new, expansive chapter devoted entirely to PHP regular expressions and how to wield them effectively. Also new in this edition, the Java chapter has been rewritten and expanded considerably to reflect new features of Java 1.5 and Java 1.6.

Intended Audience

This book will interest anyone who has an opportunity to use regular expressions. If you don't yet understand the power that regular expressions can provide, you should benefit greatly as a whole new world is opened up to you. This book should expand your understanding, even if you consider yourself an accomplished regular-expression expert. After the first edition, it wasn't uncommon for me to receive an email that started "I *thought* I knew regular expressions until I read *Mastering Regular Expressions. Now* I do."

Programmers working on text-related tasks, such as web programming, will find an absolute gold mine of detail, hints, tips, and *understanding* that can be put to immediate use. The detail and thoroughness is simply not found anywhere else.

Regular expressions are an idea—one that is implemented in various ways by various utilities (many, many more than are specifically presented in this book). If you master the general concept of regular expressions, it's a short step to mastering a particular implementation. This book concentrates on that idea, so most of the knowledge presented here transcends the utilities and languages used to present the examples.

How to Read This Book

This book is part tutorial, part reference manual, and part story, depending on when you use it. Readers familiar with regular expressions might feel that they can immediately begin using this book as a detailed reference, flipping directly to the section on their favorite utility. I would like to discourage that.

You'll get the most out of this book by reading the first six chapters as a story. I have found that certain habits and ways of thinking help in achieving a full understanding, but are best absorbed over pages, not merely memorized from a list.

The story that is the first six chapters form the basis for the last four, covering specifics of Perl, Java, .NET, and PHP. To help you get the most from each part, I've used cross references liberally, and I've worked hard to make the index as useful as possible. (Over 1,200 cross references are sprinkled throughout the book; they are often presented as "☞" followed by a page number.)