

Unraveling Regular Expressions, Step-by-Step

Introducing

Regular Expressions



O'REILLY®

Michael Fitzgerald

Copyrighted material

Introducing Regular Expressions

by Michael Fitzgerald

Copyright © 2012 Michael Fitzgerald. All rights reserved.

Printed in the United States of America.

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

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

Editor: Simon St. Laurent

Production Editor: Holly Bauer

Proofreader: Julie Van Keuren

Indexer: Lucie Haskins

Cover Designer: Karen Montgomery

Interior Designer: David Futato

Illustrator: Rebecca Demarest

July 2012: First Edition

Revision History for the First Edition:

2012-07-10 First release

2012-12-12 Second release

See <http://oreilly.com/catalog/errata.csp?isbn=9781449392680> for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Introducing Regular Expressions*, the image of a fruit bat, 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 authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-39268-0

[LSI]

What Is a Regular Expression?

Regular expressions are specially encoded text strings used as patterns for matching sets of strings. They began to emerge in the 1940s as a way to describe regular languages, but they really began to show up in the programming world during the 1970s. The first place I could find them showing up was in the QED text editor written by Ken Thompson.

“A regular expression is a pattern which specifies a set of strings of characters; it is said to match certain strings.” —Ken Thompson

Regular expressions later became an important part of the tool suite that emerged from the Unix operating system—the *ed*, *sed* and *vi* (*vim*) editors, *grep*, *AWK*, among others. But the ways in which regular expressions were implemented were not always so regular.



This book takes an inductive approach; in other words, it moves from the specific to the general. So rather than an example after a treatise, you will often get the example first and then a short treatise following that. It's a learn-by-doing book.

Regular expressions have a reputation for being gnarly, but that all depends on how you approach them. There is a natural progression from something as simple as this:

`\d`

a *character shorthand* that matches any digit from 0 to 9, to something a bit more complicated, like:

`^(\{\d{3}\})|^{\d{3}}[.-]?^{\d{3}}[.-]?^{\d{4}}$`