

K84

Reilly Media, Inc. All rights reserved.
States of America.

Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

e purchased for educational, business, or sales promotional use. Online editions
most titles (*safari.oreilly.com*). For more information, contact our
l sales department: (800) 998-9938 or *corporate@oreilly.com*.

rent
dam Witwer
olby
Witwer

Indexer: Joe Wizda
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrators: Robert Romano and Jessamyn Read

First Edition.



the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of
Unicode Explained, the image of a long-tailed glossy starling, and related trade
of O'Reilly Media, Inc.

mark of the Unicode Consortium.

ions used by manufacturers and sellers to distinguish their products are claimed as
those designations appear in this book, and O'Reilly Media, Inc. was aware of a
designations have been printed in caps or initial caps.

on has been taken in the preparation of this book, the publisher and author assume
errors or omissions, or for damages resulting from the use of the information

uses RepKover™, a durable and flexible lay-flat binding.

K

Table of Contents

Preface	ix
---------------	----

Part I. Working with Characters

1. Characters as Data	3
Introduction to Characters and Unicode	3
What's in a Character?	6
Variation of Writing Systems	27
Glyphs and Fonts	29
Definitions of Character Repertoires	36
Numbering Characters	39
Encoding Characters as Octet Sequences	43
Working with Encodings	49
Working with Fonts	56
Summaries	65
2. Writing Characters	69
Method Varieties	69
Keyboard Variation and Settings	73
Virtual Keyboards	77
Program Commands	80
Character Maps	88
Replacements on the Fly	93
Special Techniques	99
Escape Sequences	102
Specialized Editors	111
Exercise	112
3. Character Sets and Encodings	117
Good Old ASCII	117
ISO 8859 Codes	122

Windows Latin 1 and Other Windows Codes	125
Other 8-bit Codes	127
Unicode and UTF-8	135
Encodings for East Asian Language	140
Converters and Transcoding	143
Using Character Codes	145

Part II. A Systematic Look at Unicode

4. The Structure of Unicode	153
Design Principles	153
Versions of Unicode	169
Coding Space	170
Unicode Terms	181
Guide to the Unicode Standard	183
Unicode and Fonts	193
Criticism of Unicode	198
Questions and Answers	206
5. Properties of Characters	209
Character Classification	210
An Overview of Properties	213
Compositions and Decompositions	224
Normalization	237
Case Properties	244
Collation and Sorting	248
Text Boundaries	256
Directionality	257
Line-Breaking Properties	264
Unicode Conformance Requirements	281
Effects on Choosing Characters	289
6. Unicode Encodings	293
Unicode Encodings in General	293
UTF-32 and UCS-4	295
UTF-16 and UCS-2	296
UTF-8	298
Byte Order	300
Conversions Between Unicode Encodings	303
Other Encodings	303
Auto-Detecting the Encoding	317
Choosing an Encoding	317

Part III. Advanced Unicode Topics

7. Characters and Languages	323
Writing Systems and IT	323
Character Requirements of Languages	339
Transliteration and Transcription	350
Language Metadata	358
Languages and Fonts	368
8. Character Usage	371
Basics of Character Usage	371
ASCII (Basic Latin)	374
Latin-1 Supplement (ISO 8859-1)	393
Other Latin Letters	400
Other European Alphabetic Scripts	401
Diacritic Marks	402
Letterlike Symbols	411
General Punctuation	411
Line Structure Control	424
Mathematical and Technical Symbols	427
Other Blocks	434
9. The Character Level and Above	445
Levels of Text Representation and Processing	445
Characters and Markup	465
Media Types for Text	481
10. Characters in Internet Protocols	487
Information About Encoding	488
Characters in MIME	493
Content Negotiation and Multilingual Sites	515
Characters in Protocol Headers	527
Characters in Domain Names and URLs	531
11. Characters in Programming	535
Characters in Computer Languages	535
Character and String Data	545
The Preparedness Principle	566
Character Input and Output	573
Processing Form Data	579
Identifiers, Patterns, and Regular Expressions	583
International Components for Unicode (ICU)	598

Appendix: Tables for Writing Characters 611

Index 635