Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Preface

This book, *The Unicode Standard, Version 5.0*, together with the Unicode Character Database, is the authoritative source of information on Version 5.0 of the Unicode character encoding standard.

Version 5.0 of the standard is a significant departure from prior versions. It lays out much clearer requirements for supporting Unicode and provides more explicit guidance for implementers to quickly embrace the proliferation of new growth technologies and emerging markets while at the same time meeting users' needs for secure, robust software.

Why Buy This Book

In a major enhancement, Version 5.0 of the Unicode Standard is now available in a smaller, more convenient size while including much more textual content. Most notably, for the first time the book includes *all* of the Unicode Standard Annexes, which provide specifications for vital processes such as text normalization, bidirectional handling, and identifier parsing.

Version 5.0 contains the knowledge gained from many years of worldwide implementation experience and has been enhanced significantly: the text incorporates 15 years of user feedback, provides thorough answers to the many questions users of Unicode have raised, and is much more accessible—with greatly improved figures and tables, and with the text revised for clarity.

- Four-fifths of the figures are new.
- Two-thirds of the definitions are new.
- One-half of the Unicode Standard Annexes are new.
- One-third of the conformance clauses are new.
- One-fourth of the tables are new.

In addition, the text of Version 5.0 reflects advances in the computer implementation of writing systems. It substantially improves the descriptions of rendering Indic scripts to meet the demands of this area of growing market importance—Unicode-based implementations are supported by the government of India, and this book explains how to build them. Version 5.0 also highlights the newly established core CJK subset of characters, IICore, which is critical for rendering and interoperability in the East Asian market.

In short, *The Unicode Standard, Version 5.0*, enables developers to implement quickly the latest advances for worldwide software users while opening new opportunities in high-growth markets. The changes from Versions 3.0 and 4.0 to Version 5.0 are major and important—this is the one book all Unicode implementers must have.

Why Upgrade to Version 5.0

Version 5.0 of the Unicode Standard brings significant improvements beyond Versions 3.0 and 4.0. The industry has noticed and is quickly moving to Version 5.0—Windows Vista runs on 5.0; ICU, Google, and Yahoo! all have plans to upgrade to 5.0. Internet and W3C protocols are built on Unicode and are continually adapting to the latest versions. The International Standard ISO/IEC 10646 is also synchronized with Version 5.0.

This latest version of the Unicode Standard is the basis for Unicode security mechanisms, the Unicode collation algorithm, the locale data provided by the Common Locale Data Repository, and support for Unicode in regular expressions. Improved expression of the Unicode encoding model makes it much clearer how implementers need to support the representation of Unicode text in UTF-8 and other encoding forms. Character properties have been systematized and greatly extended to help implementers in support of Unicode text processing. The standard has also established principles of stability for casefolding and identifiers, crucial for interoperability and backward compatibility for formal language use and in other contexts that depend on exact usage and matching of identifiers.

Version 5.0 delivers a stable, practical character processing model in sync with today's information technology needs. Unicode now offers:

- Round-trip compatibility with the Chinese standards GB18030 and HKSCS
- The specification of the newly established core CJK subset of characters, IICore
- Refinements to casing and bidirectional behavior to meet industry requirements
- Improved Indic rendering guidelines
- Better guidance on the handling of combining characters, Unicode strings, variation selectors, line breaking, and segmentation

Implementers who want to keep pace with the industry and take advantage of a stable foundation for security, to align with the latest collation and locale data definitions, and, most importantly, to expand their market reach need to upgrade to Version 5.0 as soon as possible.

Detailed Change Information. See *Appendix D, Changes from Previous Versions*, for detailed information about the changes from previous versions of the standard, including character counts, stability guarantees, and updates to the Unicode Character Database and Unicode Standard Annexes.

Version 5.0 of the Unicode Standard corresponds to ISO/IEC 10646:2003 plus Amendments 1 and 2 to that standard and four characters to support Sindhi from Amendment 3.

Organization of This Book

This book and the Unicode Character Database define Version 5.0 of the Unicode Standard. The book gives the general principles, requirements for conformance, guidelines for implementers, character code charts and names, and the Unicode Standard Annexes.

Concepts, Architecture, Conformance, and Guidelines. The first five chapters of Version 5.0 introduce the Unicode Standard and provide the fundamental information needed to produce a conforming implementation. Basic text processing, working with combining marks, and encoding forms are all described. A special chapter on implementation guidelines answers many common questions that arise when implementing Unicode.

Chapter 1 introduces the standard's basic concepts, design basis, and coverage and discusses basic text handling requirements.

Chapter 2 sets forth the fundamental principles underlying the Unicode Standard and covers specific topics such as text processes, overall character properties, and the use of combining marks.

Chapter 3 constitutes the formal statement of conformance. This chapter also presents the normative algorithms for three processes: the canonical ordering of combining marks, the encoding of Korean Hangul syllables by conjoining *jamo*, and default casing.

Chapter 4 describes character properties in detail, both normative (required) and informative. Tables giving additional character property information appear in the Unicode Character Database.

Chapter 5 discusses implementation issues, including compression, strategies for dealing with unknown and unsupported characters, and transcoding to other standards.

Character Block Descriptions. *Chapters 6 through 16* contain the character block descriptions that give basic information about each script or group of symbols and may discuss specific characters or pertinent layout information. Some of this information is required to produce conformant implementations of these scripts and other collections of characters.

Code Charts. *Chapter 17* gives the code charts and the Character Names List. The code charts contain the normative character encoding assignments, and the names list contains normative information as well as useful cross references and informational notes.

Han Radical-Stroke Index. *Chapter 18* provides a Han radical-stroke index for the IICore subset of CJK ideographs. This index aids in locating specific, common ideographs encoded in the Unicode Standard.

Appendices. The appendices contain detailed background information on important topics regarding the history of the Unicode Standard and its relationship to ISO/IEC 10646.

Appendix A documents the notational conventions used by the standard.

Appendix B provides abstracts of Unicode Technical Reports and lists other important Unicode resources.

Appendix C details the relationship between the Unicode Standard and ISO/IEC 10646.

Appendix D lists the changes to the Unicode Standard since Version 4.0.

Appendix E describes the history of Han unification in the Unicode Standard.

Appendix F reproduces the text of the policies of the Unicode Consortium regarding character encoding stability.

Glossary, References, and Indices. The appendices are followed by a glossary of terms, a bibliography, and two indices: an index to Unicode characters and an index to the text of the book.

Unicode Standard Annexes

The Unicode Standard Annexes are printed in the back of this book, following the indices. These annexes form an integral part of the Unicode Standard. Conformance to a version of the Unicode Standard includes conformance to its Unicode Standard Annexes.

Unicode Standard Annex #9, "The Bidirectional Algorithm," describes specifications for the positioning of characters in mixed-directional text, such as Arabic or Hebrew.

Unicode Standard Annex #11, "East Asian Width," presents the specification of an informative property for Unicode characters that is useful when interoperating with East Asian legacy character sets.

Unicode Standard Annex #14, "Line Breaking Properties," presents the specification of line breaking properties for Unicode characters.

Unicode Standard Annex #15, "Unicode Normalization Forms," describes specifications for four normalized forms of Unicode text.

Unicode Standard Annex #24, "Script Names," specifies an assignment of script names to all Unicode code points.

Unicode Standard Annex #29, "Text Boundaries," describes guidelines for determining default boundaries between certain significant text elements: grapheme clusters ("user-perceived characters"), words, and sentences.

Unicode Standard Annex #31, "Identifier and Pattern Syntax," describes specifications for recommended defaults for the use of Unicode in the definitions of identifiers and in pattern-based syntax.

Unicode Standard Annex #34, "Unicode Named Character Sequences," defines the concept of a Unicode named character sequence and a set of rules constraining possible names applied to character sequences.

Unicode Standard Annex #41, "Common References for Unicode Standard Annexes," contains the listing of references shared by other Unicode Standard Annexes.

The 5.0.0 version of each UAX is included on the CD-ROM. All versions, including the most up-to-date versons of all Unicode Standard Annexes, are available on the Unicode Web site:

http://www.unicode.org/reports/

The Unicode Character Database

The Unicode Character Database (UCD) is a collection of data files containing character code points, character names, and character property data. It is described more fully in *Section 4.1, Unicode Character Database*. All versions, including the most up-to-date version of the Unicode Character Database, are found on the Unicode Web site:

http://www.unicode.org/ucd/

The files for Version 5.0.0 of the Unicode Character Database are also supplied on the CD-ROM that accompanies this book.

Information on versioning and on all versions of the Unicode Standard can be found on the Unicode Web site:

http://www.unicode.org/versions/

Unicode Technical Standards and Unicode Technical Reports

Unicode Technical Reports and Unicode Technical Standards are separate publications and do not form part of the Unicode Standard.

All versions of all Unicode Technical Reports and Unicode Technical Standards are available on the Unicode Web site:

http://www.unicode.org/reports/

The latest available version of each document at the time of publication is included on the CD-ROM. See *Appendix B, Unicode Publications and Resources*, for a summary overview of important Unicode Technical Standards and Unicode Technical Reports.

On the CD-ROM

The CD-ROM contains additional information, such as sample code, which is maintained on the Unicode FTP site:

ftp.unicode.org

It is also available via HTTP:

http://www.unicode.org/Public/

For the contents of the CD-ROM, see its ReadMe.txt file.

Updates and Errata

Reports of errors in the Unicode Standard, including the Unicode Character Database and the Unicode Standard Annexes, may be reported using the online reporting form:

http://www.unicode.org/reporting.html

A list of known errata is maintained on the Unicode Web site:

http://www.unicode.org/errata/

Any currently listed errata will be fixed in subsequent versions of the standard.