

# The Unicode® Standard

## Version 10.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://www.unicode.org/versions/latest/>.

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 the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

The authors and publisher have taken care in the preparation of this specification, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

© 2017 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 10.0.

Includes bibliographical references and index.

ISBN 978-1-936213-16-0 (<http://www.unicode.org/versions/Unicode10.0.0/>)

1. Unicode (Computer character set) I. Unicode Consortium.

QA268.U545 2017

ISBN 978-1-936213-16-0

Published in Mountain View, CA

June 2017

# Preface

This is *The Unicode Standard, Version 10.0*. It supersedes all earlier versions of the Unicode Standard.

## **Why Unicode?**

The Unicode Standard and its associated specifications provide programmers with a single universal character encoding, extensive descriptions, and a vast amount of data about how characters function. The specifications and data describe how to form words and break lines; how to sort text in different languages; how to format numbers, dates, times, and other elements appropriate to different languages; how to display languages whose written form flows from right to left, such as Arabic and Hebrew, or whose written form splits, combines, and reorders, such as languages of South Asia. These specifications include descriptions of how to deal with security concerns regarding the many “look-alike” characters from alphabets around the world. Without the properties and algorithms in the Unicode Standard and its associated specifications, interoperability between different implementations would be impossible, and much of the vast breadth of the world’s languages would lie outside the reach of modern software.

## **What’s New?**

Unicode 10.0 adds 8,518 characters, for a total of 136,690 characters. Other important updates in Unicode Version 10.0 include:

- Improvements in the charts for the Mongolian script
- Significant updates to segmentation algorithms

**Support for Languages and Symbol Sets.** The following new scripts were added in Version 10.0:

- Masaram Gondi, used to write Gondi in Central and Southeast India
- Nüshu, used by women in China to write poetry and other discourses until the late twentieth century
- Soyombo and Zanabazar Square, used in historic Buddhist texts to write Sanskrit, Tibetan, and Mongolian

Support for lesser-used languages was extended worldwide, including:

- Syriac letters used for writing Suriyani Malayalam, also known as Garshuni and Syriac Malayalam
- Gujarati signs used for the transliteration of the Arabic script into Gujarati by Ismaili Khoja communities

Further additions were also made for Chinese, Japanese, and Korean:

- A set of 285 Hentaigana characters used in Japan (historic variant of Hiragana characters)
- CJK Extension F (7,473 Han characters)

Important symbol additions include:

- Bitcoin sign
- 56 emoji characters
- A set of Typicon marks and symbols

**Property and Behavioral Updates.** The core data files of the Unicode Character Database were updated for the new additions in Version 10.0.

**Detailed Change Information.** See *Appendix D, Version History of the Standard* and also <http://www.unicode.org/versions/Unicode10.0.0/> for detailed information about the changes from the previous versions of the standard, including character counts, and significant changes to the Unicode Character Database and Unicode Standard Annexes.

## ***Organization of This Standard***

This core specification, together with the Unicode code charts, the Unicode Character Database, and the Unicode Standard Annexes, defines Version 10.0 of the Unicode Standard. The core specification contains the general principles, requirements for conformance, and guidelines for implementers. The character code charts and names are available online.

**Concepts, Architecture, Conformance, and Guidelines.** The first five chapters of Version 10.0 introduce the Unicode Standard and provide the fundamental information needed to produce a conforming implementation. Basic text processing, working with combining marks, encoding forms, and normalization 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 several processes, including normalization, Korean syllable boundary determination, and default casing.

*Chapter 4* describes character properties in detail, both normative (required) and informative. Additional character property information appears in Unicode Standard Annex #44, “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 23* contain the character block descriptions that provide 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 24* describes the conventions used in the code charts and the list of character names. 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.

**Appendices.** The appendices contain additional information.

*Appendix A* documents the notational conventions used by the standard.

*Appendix B* provides information about Unicode publications and links to other important Unicode resources.

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

*Appendix D* lists version history and details code point allocation history.

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

*Appendix F* provides additional documentation for characters encoded in the CJK Strokes block (U+31C0..U+31EF).

**References and Index.** The appendices are followed by a bibliography and an index to the text of this core specification.

**Glossary and Character Index.** A glossary of Unicode terms and the Unicode Character Name Index may be found at:

<http://www.unicode.org/glossary/>

<http://www.unicode.org/charts/charindex.html>

## **Unicode Standard Annexes**

The Unicode Standard Annexes form an integral part of the Unicode Standard. Conformance to a version of the Unicode Standard includes conformance to its Unicode Standard

Annexes. All versions, including the most up-to-date versions of all Unicode Standard Annexes, are available at:

<http://www.unicode.org/reports/index.html#annexes>

The following is the list of Unicode Standard Annexes:

Unicode Standard Annex #9, “Unicode Bidirectional Algorithm,” describes specifications for the positioning of characters in text containing characters flowing from right to left, 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, “Unicode Line Breaking Algorithm,” presents the specification of line breaking properties for Unicode characters.

Unicode Standard Annex #15, “Unicode Normalization Forms,” describes Unicode normalization and provides examples and implementation strategies for it.

Unicode Standard Annex #24, “Unicode Script Property,” describes two related Unicode code point properties. Both properties share the use of Script property values. The Script property itself assigns single script values to all Unicode code points, identifying a primary script association, where possible. The Script\_Extensions property assigns sets of Script property values, providing more detail for cases where characters are commonly used with multiple scripts.

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

Unicode Standard Annex #31, “Unicode 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.

Unicode Standard Annex #38, “Unicode Han Database (UniHan),” describes the organization and content of the UniHan database.

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

Unicode Standard Annex #42, “Unicode Character Database in XML,” describes an XML representation of the Unicode Character Database.

Unicode Standard Annex #44, “Unicode Character Database,” provides the core documentation for the Unicode Character Database (UCD). It describes the layout and organization of the Unicode Character Database and how the UCD specifies the formal definition of Unicode character properties.

Unicode Standard Annex #45, “U-Source Ideographs,” describes U-source ideographs as used by the Ideographic Rapporteur Group (IRG) in its CJK ideograph unification work.

Unicode Standard Annex #50, “Unicode Vertical Text Layout,” describes the Unicode character property, `Vertical_Orientation`, which can serve as a stable default orientation for characters for reliable document interchange.

## ***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* and in Unicode Standard Annex #44, “Unicode Character Database.” All versions, including the most up-to-date version of the Unicode Character Database, are found at:

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

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

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

## ***Unicode Code Charts***

The Unicode code charts contain the character encoding assignments and the names list. The archival, reference set of versioned 10.0 code charts may be found at:

<http://www.unicode.org/charts/PDF/Unicode-10.0/>

For easy lookup of characters, see the current code charts:

<http://www.unicode.org/charts/>

An interactive radical-stroke index to CJK ideographs is located at:

<http://www.unicode.org/charts/unihanrsindex.html>

## ***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 at:

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

### ***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 reporting form:

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

A list of known errata is maintained at:

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

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

### ***Acknowledgements***

The Unicode Standard, Version 10.0 is the result of the dedication and contributions of many people over several years. We would like to acknowledge the individuals whose contributions were central to the design, authorship, and review of this standard. A complete listing of acknowledgements can be found at:

<http://www.unicode.org/acknowledgements/>

Current editorial contributors can be found at:

<http://www.unicode.org/consortium/edcom.html>