

The Unicode Standard

Version 6.0 – 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.

Copyright © 1991–2011 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 Julie D. Allen ... [et al.]. — Version 6.0.

Includes bibliographical references and index.

ISBN 978-1-936213-01-6 (<http://www.unicode.org/versions/Unicode6.0.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2011

ISBN 978-1-936213-01-6

Published in Mountain View, CA

February 2011

Preface

This is *The Unicode Standard, Version 6.0*. 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 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 Version 6.0 is the first major version of the Unicode Standard to be published solely in online format. All new content and corrections have been incorporated into this version.

Key new features that have been defined and documented since the publication of *The Unicode Standard, Version 5.2* include:

- addition of the Indian rupee sign
- significant character additions for *emoji* (pictographic) use in mobile telephones
- increased security for spoof detection
- consolidation of normalization conformance specifications
- addition of important Han characters
- expanded support to meet African language requirements
- additional stability guarantees for decimal numbers
- updated collation and IDNA compatibility processing
- charts for CJK Compatibility Ideographs in a multicolumn format showing sources, comparable to charts for CJK unified ideographs

Support for Languages and Symbol Sets. 2,088 new characters were added in the Unicode Standard, Version 6.0. Of these characters, 230 characters were added to the basic multilingual plane (BMP—characters with code points less than U+10000), while 1,858 characters were added in the supplementary planes. For the first time in the history of the Unicode Standard, the majority of the regular encoded characters (graphic and format) are outside of the BMP.

Over 1,000 new symbols were added—chief among them 729 *emoji* symbols which are in widespread use, particularly in Japan, and which have become an essential part of text messages. These *emoji* symbols include symbols for many domains: map and transport symbols, phases of the moon, UI symbols such as fast-forward, new emoticons, and many others.

Other significant symbol additions include the newly created official Indian currency symbol, the Indian rupee sign at U+20B9. The design was presented to the public by the government of India in July of 2010, and it was quickly adopted in financial reporting by newspapers and for product pricing.

More than 250 other symbols were added, including math symbols, playing cards, the set of alchemical symbols for scholarly use, and many others. Computer hardware and software implementers should note that the world grid symbol (U+1F310) was added in Version 6.0. This symbol is often used to indicate keyboard settings or international settings for software.

African language support was enhanced through the addition of 603 characters, including extensions to the Tifinagh, Ethiopic, and Bamum scripts. Version 6.0 also has three additional scripts: Batak, Mandaic, and Brahmi.

Conformance Updates. The most important changes to conformance requirements in Unicode Version 6.0 are:

- clarification of the explanatory text of D92 in *Section 3.9, Unicode Encoding Forms* regarding maximal subpart
- clarification regarding the star operator used in the regex for Final_Sigma in *Table 3-15*
- update of D136 in *Section 3.13, Default Case Algorithms* regarding Case_Ignorable with a rewrite of the definitions and descriptions for caseless matching
- addition of text in *Chapter 3, Conformance* after C4 to clarify the implications of conformance to Unicode semantics

In addition, small clarifications of the conformance clauses with no significant changes to conformance requirements were made in Unicode Standard Annex #9, “The Unicode Bidirectional Algorithm.” There were major editorial revisions of Unicode Standard Annex #44, “Unicode Character Database,” and Unicode Standard Annex #15, “Unicode Normalization Forms.” However, no significant changes to conformance requirements were made to those documents.

Important Property and Behavioral Updates. The significant property changes in Unicode 6.0 include:

- reversal of the default grapheme cluster boundary determination for Thai and Lao to the behavior specified in Unicode 5.0
- formal deprecation of one Arabic character (U+0673)
- change to the general category for one New Tai Lue numeric character (U+19DA) which disqualifies it from identifiers unless grandfathered
- change to the general category for two Kannada characters (U+0CF1, U+0CF2) which makes them newly eligible for inclusion in identifiers

A few properties were deprecated: Hyphen, ISO_Comment, and certain derived normalization properties. The ISO_Comment property was also stabilized, because it is no longer needed for chart generation, and its values are null for all code points.

Block Descriptions. Updates to the text in Version 6.0 further improve the descriptions of Indic scripts, including Devanagari and Tamil, and other scripts such as Arabic, Tibetan, Balinese, Javanese, Han, and Hangul. Block descriptions were also updated to document significant new character additions: Indian rupee sign, Malayalam *dot reph* character, Oriya fractions, a new Kana supplement, Ethiopic additions, Tifinagh bi-consonants, Bamum supplement, alchemical symbols, and a new section on the *emoji* characters for mobile phones.

Data File Descriptions and Updates. There are four new data files in Version 6.0:

- an *emoji* source file maps the *emoji* symbols to their original Japanese telecommunication source sets
- a scripts extensions file provides provisional information for data used across multiple scripts for segmentation, regular expressions, and spoof detection
- two Indic property files provide provisional information for the specification of syllabic structure for Indic scripts

Additionally, a number of improvements were made to the tests for line breaking and bidirectional data. The description of Unihan.txt in Unicode Standard Annex #38, “Unicode Han Database” was updated to clarify the status of certain Japanese readings and to provide updated regular expressions. Unicode Standard Annex #44, “Unicode Character Database” was enhanced to add tables listing the Deprecated and Stabilized properties, update the matching rules, and to add documentation for new properties and data files.

Stability Policy Updates. While the Unicode stability policies are not formally part of the standard, important changes were made to them which affect the Unicode Standard, Version 6.0.

The Property Alias Uniqueness stability policy was updated to make it clear that uniqueness is defined specifically by the UAX44-LM3 matching rule, rather than by a generic reference to all of the UAX #44 matching rules. The UAX44-LM3 matching rule also was clarified regarding the status of any property aliases beginning with the sequence of characters “is” (or “Is” or “IS”). This reflects the prevalence of implementations of Unicode character properties or property values with APIs prefixed with “Is”, as for example IsNumeric() for the Unicode Numeric property.

Another Property Value Stability constraint was added to make it clear that decimal digits (Numeric_Type=Decimal) only occur in contiguous ranges of 10 characters, with ascending numeric values from 0 to 9.

CJK. Unicode Version 6.0 adds five characters in the main block for Hong Kong to cover HKSCS (2008). It also adds 222 characters in a new CJK Extension D including urgently needed characters for government registries for China and Japan, as well as characters to support the Hakka dialect in Taiwan.

Industry Standards. Industry standards, including Internet and W3C protocols, are built on Unicode and are periodically synchronized to the latest versions. The International Standard ISO/IEC 10646 is also aligned with this latest version of the Unicode Standard.

Version 6.0 of the Unicode Standard provides the basis for the most up-to-date Unicode security mechanisms, the Unicode Collation Algorithm, IDNA compatibility processing, locale data, and support for Unicode in regular expressions.

Detailed Change Information. See *Appendix D, Changes from Previous Versions* for detailed information about the changes from the previous versions of the standard, including character counts, conformance clause and definition updates, 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 6.0 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 also available online.

Concepts, Architecture, Conformance, and Guidelines. The first five chapters of Version 6.0 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 16* 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 17* 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 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 5.0.

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+C130..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/>

The following is a list of Unicode Standard Annexes:

Unicode Standard Annex #9, “Unicode 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, “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,” discusses the Script property specified in the Unicode Character Database.

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.

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 6.0 code charts may be found at:

<http://www.unicode.org/charts/PDF/Unicode-6.0/>

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/>

See *Appendix B, Unicode Publications and Resources*, for a summary overview of important Unicode Technical Standards and Unicode Technical 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 6.0 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/>

