

# Preface

This is *The Unicode Standard, Version 5.2*. It supersedes all earlier versions of the Unicode Standard.

## **What's New**

For the first time in a minor version of the Unicode Standard, all new content and corrections have been incorporated into this consolidated text for the core specification. All changes for both Version 5.1 and for Version 5.2 are fully included in this text.

Key new features that have been defined and documented since the book publication of *The Unicode Standard, Version 5.0* include:

- increased security in data exchange
- significant character additions for Indic and Southeast Asian scripts
- expanded identifier specifications for Indic and Arabic scripts
- improvements in support for Tamil, Malayalam, and other Indic scripts
- modification to line breaking conformance to better support HTML and other protocols
- strengthened normalization stability and case pair stability
- additional stability guarantees for properties, property aliases and property value aliases

The following discussion highlights further, specific details, indicating whether important changes were already present for Version 5.1 or were more recent additions new in Version 5.2.

**Conformance, General Structure and Property Updates.** Version 5.2 improves data security by giving a tighter specification of the modification of character sequences in conformance clause C7. The conformance requirements now disallow the removal of noncharacter code points from uninterpreted text strings, except where strings are explicitly being modified.

Version 5.2 also improves the documentation of conformance requirements for the specification of normalization forms, canonical ordering, and the status of types of properties. This latest version incorporates into *Chapter 3, Conformance* the formal definitions of normalization previously presented in Unicode Standard Annex #15, “Unicode Normalization Forms.” It revises *Section 3.5, Properties* to better explain the status of Normative, Informative, Provisional, and Contributory properties, and it improves the description of compatibility characters in *Chapter 2, General Structure*. The types of code points were clarified in Chapters 2, 3, and 4, with coordinated updates in Unicode Standard Annex #44, “Unicode Character Database.”

The PropertyAliases.txt file in the Unicode Character Database is now designated as the normative listing of Unicode character properties and their names.

**Block Descriptions.** Updates to the text in Version 5.2 further improve the descriptions of Indic scripts, including extensive new documentation on rendering Malayalam chillus, and a complete specification of the representation of Tamil atomic elements: consonants,

stand-alone vowels, and syllables. Version 5.1 also included an enhanced block description for Myanmar.

**Data File Descriptions and Updates.** The documentation of the Unicode Character Database has been significantly revised in Unicode Standard Annex #44, “Unicode Character Database.” New since Version 5.0 is documentation of the Unihan Database in a Unicode Standard Annex: Unicode Standard Annex #38, “Unicode Han Database (Unihan).”

Implementers also will find:

- Updated conformance test data for the Unicode Line Breaking Algorithm and, new for the first time, test data to assess conformance to the Unicode Bidirectional Algorithm
- XML data files that encapsulate all of the Unicode character properties
- Unihan data files reorganized for better access

**Important Property and Behavioral Updates.** Since Version 5.0, the standard has been updated with significant changes to properties and behavioral specifications. Several important property definitions were extended, improving line breaking for Polish and Portuguese hyphenation. The Unicode Text Segmentation Algorithms, covering sentences, words, and characters, were greatly enhanced to improve the processing of Tamil and other Indic scripts. The Unicode Normalization Algorithm has defined stabilized strings and has provided guidelines for buffering, and Unicode Standard Annex #31, “Identifier and Pattern Syntax,” has clarified the use of scripts in identifiers.

Version 5.2 clarifies the definition of “Deprecated” and its relationship to the use of the wording, “strongly discouraged,” updates the set of deprecated characters in view of this clearer definition, and updates best practices for the use of replacement characters. Overall, character properties have been systematized and greatly extended to help implementers of Unicode text processing.

Property stability guarantees have been added that make it possible to use property aliases and property value aliases as stable identifiers. Version 5.2 adds stability guarantees for normative and informative properties—building on Version 5.1, which added stability guarantees for property aliases and property value aliases. This latest version also adds a policy on alias uniqueness.

**Support for Languages and Symbol Sets.** Version 5.1 extended support for additional languages with the addition of eight contemporary use scripts: Cham, Kayah Li, Lepcha, Ol Chiki, Rejang, Saurashtra, Sundanese, and Vai. Version 5.2 continues support for new languages with seven more contemporary use scripts: Bamum, Javanese, Lisu, Meetei Mayek, Samaritan, Tai Tham, and Tai Viet.

New character additions to existing scripts provide greater support for the language communities of Pakistan, North Africa, and Abkhazia, and users of Canadian Aboriginal Syllabics and of the Coptic, Devanagari, Malayalam, and Myanmar scripts. Of particular note are Devanagari additions to support Vedic Sanskrit. Encoding Vedic is significant because Sanskrit is one of the principal languages for the religious heritage of India, and because Vedic represents the earliest attested phase of the language.

Standardized named sequences were added for Lithuanian and Tamil.

The fifteen additional contemporary scripts, standardized named sequences, and newly encoded individual characters expand support of language and orthographic communities in Africa, India, China, Central Asia, South Asia, Southeast Asia, and the Middle East.

Other additions include important modern use symbols, currency signs, and historic characters. With Unicode Version 5.2, scholars will now have access to the Gardiner set of Egypt-

tian Hieroglyphs as well as other historic scripts: Imperial Aramaic, Avestan, Kaithi, Old South Arabian, and Old Turkic. Several key symbol sets were added or expanded: the ARIB set of Japanese broadcasting symbols, additional number forms used in India, and currency symbols.

**CJK.** A major feature since the publication of Version 5.0 is the enabling of ideographic variation sequences. These sequences allow standardized representation of glyphic variants needed for Japanese, Chinese, and Korean text. The first registered collection is now available at:

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

**Industry Standards.** Industry standards, including 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 this latest version of the Unicode Standard.

Version 5.2 of the Unicode Standard provides the basis for the most up-to-date Unicode security mechanisms, the Unicode Collation Algorithm, the locale data provided by the Unicode Locales Project, the Common Locale Data Repository, 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 define Version 5.2 of the Unicode Standard. The core specification contains the general principles, requirements for conformance, and the guidelines for implementers. The character code charts and names are also available online.

**Concepts, Architecture, Conformance, and Guidelines.** The first five chapters of Version 5.2 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 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* 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.

**Han Radical-Stroke Index.** *Chapter 18* describes the Han radical-stroke index for 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 5.0.

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

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

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

<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 on the Unicode Web site:

<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*. 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/>

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 Code Charts***

The Unicode code charts contain the character encoding assignments and the names list. The archival, reference set of versioned 5.2 code charts may be found on the Unicode Web site:

<http://www.unicode.org/charts/PDF/Unicode-5.2/>

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 on the Unicode Web site:

<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 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.