The Unicode Standard
Version 8.0 – Core Specification

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

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Appendix B

Unicode Publications and Resources

This appendix provides information about the Unicode Consortium and its activities, particularly regarding publications other than the Unicode Standard. The Unicode Consortium publishes a number of technical standards and technical reports, and the current list of those, with abstracts of their content, is included here for convenient reference.

The Unicode website also has many useful online resources. *Section B.6, Other Unicode Online Resources*, provides a guide to the kinds of information available online.

B.1 The Unicode Consortium

The Unicode Consortium was incorporated in January 1991, under the name Unicode, Inc., to promote the Unicode Standard as an international encoding system for information interchange, to aid in its implementation, and to maintain quality control over future revisions.

To further these goals, the Unicode Consortium cooperates with the Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC1). It holds a Class C liaison membership with ISO/IEC JTC1/SC2; it participates in the work of both JTC1/SC2/WG2 (the technical working group for the subcommittee within JTC1 responsible for character set encoding) and the Ideographic Rapporteur Group (IRG) of WG2. The Consortium is a member company of the InterNational Committee for Information Technology Standards, Technical Committee L2 (INCITS/L2), an accredited U.S. standards organization. Many members of the Unicode Consortium have representatives in many countries who also work with other national standards bodies. In addition, a number of organizations are Liaison Members of the Consortium. For a list, see the Unicode website.

Membership in the Unicode Consortium is open to organizations and individuals anywhere in the world who support the Unicode Standard and who would like to assist in its extension and widespread implementation. Full, Institutional, Supporting, and Associate Members represent a broad spectrum of corporations and organizations in the computer and information processing industry. For a list, see the Unicode website. The Consortium is supported financially solely through membership dues.

The Unicode Technical Committee

The Unicode Technical Committee (UTC) is the working group within the Consortium responsible for the creation, maintenance, and quality of the Unicode Standard. The UTC follows an open process in developing the Unicode Standard and its other technical publications. It coordinates and reviews all technical input to these documents and decides their contents. For more information on the UTC and the process by which the Unicode Standard and the other technical publications are developed, see:

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

Other Activities

Going beyond developing technical standards, the Unicode Consortium acts as registration authority for the registration of script identifiers under ISO 15924, and it has a technical committee dedicated to the maintenance of the Unicode Common Locale Data Repository (CLDR). The repository contains a large and rapidly growing body of data used in the locale definition for software internationalization. For further information about these and other activities of the Unicode Consortium, visit:

http://www.unicode.org

B.2 Unicode Publications

In addition to the Unicode Standard, the Unicode Consortium publishes Unicode Technical Standards and Unicode Technical Reports. These materials are published as electronic documents only and, unlike Unicode Standard Annexes, do not form part of the Unicode Standard.

A *Unicode Standard Annex* (UAX) forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document is always the same as the version of the Unicode Standard of which it forms a part.

A *Unicode Technical Standard* (UTS) is an independent specification. Conformance to the Unicode Standard does not imply conformance to any UTS.

A *Unicode Technical Report* (UTR) contains informative material. Conformance to the Unicode Standard does not imply conformance to any UTR. Other specifications, however, are free to make normative references to a UTR.

In the past, some normative material was published as Unicode Technical Reports. Currently, however, such material is published either as a Unicode Technical Standard or a Unicode Standard Annex.

The Unicode website is the source for the most current version of all three categories of technical reports:

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

The following sections provide lists of abstracts for current Unicode Technical Standards and Unicode Technical Reports. They are listed numerically within each category. There are gaps in the numerical sequence become some of the reports have been superseded or have been incorporated into the text of the standard.

B.3 Unicode Technical Standards

UTS #6: A Standard Compression Scheme for Unicode

This report presents the specifications of a compression scheme for Unicode and sample implementation.

UTS #10: Unicode Collation Algorithm

This report provides the specification of the Unicode Collation Algorithm, which provides a specification for how to compare two Unicode strings while remaining conformant to the requirements of The Unicode Standard. The UCA also supplies the Default Unicode Collation Element Table (DUCET) as the data specifying the default collation order for all Unicode characters.

UTS #18: Unicode Regular Expressions

This document describes guidelines for how to adapt regular expression engines for use with the Unicode Standard.

UTS #22: Character Mapping Markup Language (CharMapML)

This document specifies an XML format for the interchange of mapping data for character encodings. It provides a complete description for such mappings in terms of a defined mapping to and from Unicode code points, and a description of alias tables for the interchange of mapping table names.

UTS #35: Unicode Locale Data Markup Language (LDML)

This document describes an XML format (*vocabulary*) for the exchange of structured locale data. This format is used in the Unicode Common Locale Data Repository.

UTS #37: Unicode Ideographic Variation Database

This document describes the organization of the Ideographic Variation Database and the procedure to add sequences to that database.

UTS #39: Unicode Security Mechanisms

Because Unicode contains such a large number of characters and incorporates the varied writing systems of the world, incorrect usage can expose programs or systems to possible security attacks. This report specifies mechanisms that can be used in detecting possible security problems.

UTS #46: Unicode IDNA Compatibility Processing

Client software, such as browsers and emailers, faces a difficult transition from the version of international domain names approved in 2003 (IDNA2003), to the revision approved in 2010 (IDNA2008). The specification in this document provides a mechanism that minimizes the impact of this transition for client software, allowing client software to access domains that are valid under either system.

B.4 Unicode Technical Reports

UTR #16: UTF-EBCDIC

This document presents the specifications of UTF-EBCDIC: EBCDIC Friendly Unicode (or UCS) Transformation Format.

UTR #17: Unicode Character Encoding Model

This document clarifies a number of the terms used to describe character encodings, and where the different forms of Unicode fit in. It elaborates the Internet Architecture Board (IAB) three-layer "text stream" definitions into a four-layer structure.

UTR #20: Unicode in XML and Other Markup Languages

This document contains guidelines on the use of the Unicode Standard in conjunction with markup languages such as XML.

UTR #23: The Unicode Character Property Model

This document presents a conceptual model of character properties defined in the Unicode Standard.

UTR #25: Unicode Support for Mathematics

The Unicode Standard includes virtually all standard characters used in mathematics. This set supports a wide variety of math usage on computers, including in document presentation languages like TEX, in math markup languages like MathML and OpenMath, in internal representations of mathematics for applications like Mathematica, Maple, and MathCAD, in computer programs, and in plain text. This technical report describes the Unicode support for mathematics and gives some of the imputed default math properties for Unicode characters.

UTR #26: Compatibility Encoding Scheme for UTF-16: 8-Bit (CESU-8)

This document specifies an 8-bit Compatibility Encoding Scheme for UTF-16 (CESU) that is intended for internal use within systems processing Unicode to provide an ASCII-compatible 8-bit encoding that is similar to UTF-8 but preserves UTF-16 binary collation. *It is not intended or recommended as an encoding used for open information exchange.* The Unicode Consortium does not encourage the use of CESU-8, but does recognize the existence of data in this encoding and supplies this technical report to clearly define the format and to distinguish it from UTF-8. This encoding does not replace or amend the definition of UTF-8.

UTR #33: Unicode Conformance Model

This report defines conformance terminology, specifies different areas and levels of conformance, and describes what it means to make a claim of conformance or "support" of the standard. This conformance model presented here is not a framework for conformance verification testing.

UTR #36: Unicode Security Considerations

Because Unicode contains such a large number of characters and incorporates the varied writing systems of the world, incorrect usage can expose programs or systems to possible security attacks. This is especially important as more and more products are internationalized. This document describes some of the security considerations that programmers, system analysts, standards developers, and users should take into account, and provides specific recommendations to reduce the risk of problems.

UTR #50: Unicode Vertical Text Layout

The Unicode code charts generally show characters oriented for horizontal presentation. However, some of the glyphs are actually oriented for vertical presentation. A few characters change shape or orientation when the text is rotated from horizontal to vertical. This report describes a Unicode character property which can serve as a stable default orientation of characters for reliable document interchange.

UTR #51: Unicode Emoji

This document aims to improve the interoperability of emoji characters across implementations by providing guidelines and data.

B.5 Unicode Technical Notes

Unicode Technical Notes provide information on a variety of topics related to Unicode and internationalization technologies.

These technical notes are independent publications, not approved by any of the Unicode Technical Committees, nor are they part of the Unicode Standard or any other Unicode specification. Publication does not imply endorsement by the Unicode Consortium in any way. These documents are not subject to the Unicode Patent Policy. Unicode Technical Notes can be found on the Unicode website at:

http://www.unicode.org/notes/

The technical notes cover the following topics (among others):

- Algorithms
- Collation
- Compression and code set conversions
- Language identification
- Migration of software
- Modern and historical scripts
- Text layout and rendering
- Tutorials
- Social and cultural issues

B.6 Other Unicode Online Resources

The Unicode Consortium provides a number of online resources for obtaining information and data about the Unicode Standard as well as updates and corrigenda.

Unicode Online Resources

Unicode Web Site

http://www.unicode.org

Unicode Anonymous FTP Site

ftp://ftp.unicode.org

Charts. The charts section of the website provides code charts for all of the Unicode characters, plus specialized charts for normalization, collation, case mapping, script names, and Unified CJK Ideographs.

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

Character Index. Online index by character name, to look up Unicode code points. This index also makes it easy to look up the location of scripts in the standard, and indexes common alternative names for characters as well.

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

Conferences. The Internationalization and Unicode Conferences are of particular value to anyone implementing the Unicode Standard or working on internationalization. A variety of tutorials and conference sessions cover current topics related to the Unicode Standard, the World Wide Web, software, internationalization, and localization.

http://www.unicode.org/conference/

E-mail Discussion List. Subscription instructions for the public e-mail discussion list are posted on the Unicode website.

FAQ (Frequently Asked Questions). The FAQ pages provide an invaluable resource for understanding the Unicode Standard and its implications for users and implementers.

http://www.unicode.org/conference/

Glossary. Online listing of definitions for technical terms used in the Unicode Standard and other publications of the Unicode Consortium.

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

Online Unicode Character Database. This page supplies information about the online Unicode Character Database (UCD), including links to documentation files and the most up-to-date version of the data files, as well as instructions on how to access any particular version of the UCD.

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

Online Unihan Database. The online Unihan Database provides interactive access to all of the property information associated with CJK ideographs in the Unicode Standard.

http://www.unicode.org/chart/unihan.html

Policies. These pages describe Unicode Consortium policies on stability, patents, and Unicode website privacy. The stability policies are particularly important for implementers, documenting invariants for the Unicode Standard that allow implementations to be compatible with future and past versions.

http://www.unicode.org/policies/

Unicode Common Locale Data Repository (CLDR). Machine-readable repository, in XML format, of locale information for use in application and system development.

http://www.unicode.org/cldr/

Updates and Errata. This page lists periodic updates with corrections of typographic errors and new clarifications of the text.

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

Versions. This page describes the version numbering used in the Unicode Standard, the nature of the Unicode character repertoire, and ways to cite and reference the Unicode Standard, the Unicode Character Database, and Unicode Technical Reports. It also specifies the exact contents of each and every version of the Unicode Standard, back to Unicode 1.0.0.

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

Where Is My Character? This page provides basic guidance to finding Unicode characters, especially those whose glyphs do not appear in the charts, or that are represented by sequences of Unicode characters.

http://www.unicode.org/standard/where/

How to Contact the Unicode Consortium

The best way to contact the Unicode Consortium to obtain membership information is via the website:

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

The website also lists the current telephone, fax, and courier delivery address. The Consortium's postal address is:

P.O. Box 391476
Mountain View, CA 94039-1476
USA