L2/08-067

Proposed Draft

Unicode Standard Annex #44

UNICODE CHARACTER DATABASE

Version	Unicode 5.1 draft 1
Authors	Mark Davis (mark.davis@google.com) and Ken Whistler (ken@unicode.org)
Date	2008-1-11
This Version	http://www.unicode.org/reports/tr44/tr44-1.html
Previous Version	n/a
Latest Version	http://www.unicode.org/reports/tr44/
Revision	1

Summary

This annex consolidates information documenting the Unicode Character Database.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***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 corresponds to the version of the Unicode Standard of which it forms a part.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "Common References for Unicode Standard Annexes." For the latest version of the Unicode Standard, see [Unicode]. For a list of current Unicode Technical Reports, see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

> ***Warning:*** *the information in this file does not completely describe the use and interpretation of Unicode character properties and behavior. It must be used in conjunction with the data in the other files in the Unicode Character Database, and relies on the notation and definitions supplied in The Unicode Standard. All chapter references are to Version 5.0.0 of the standard unless otherwise indicated.*

1 Introduction

The Unicode Character Database (UCD) is a collection of data files which contain the Unicode character code points and character names and which define the Unicode character properties and mappings between Unicode characters (such as case mappings).

This annex describes the UCD and provides a guide to the various documentation files associated with it.

The current version of the UCD is always located on the Unicode Web site at:

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

The specific files for the UCD associated with this version of the Unicode Standard (5.1.0) are located at:

http://www.unicode.org/Public/5.1.0/

Stable, archived versions of the UCD associated with all earlier versions of the Unicode Standard can be accessed from:

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

See Section 4.1, "Unicode Character Database", in [Unicode] for a general discussion of the UCD and its use in defining properties.

2 Conformance

The Unicode Character Database is an integral part of the Unicode Standard.

The UCD contains normative property and mapping information required for implementation of various Unicode algorithms such as the Unicode Bidirectional Algorithm, Unicode Normalization, and Unicode Casefolding. The data files also contain additional informative and provisional character property information.

Each specification of a Unicode algorithm, whether specified in the text of [Unicode] or in one of the Unicode Standard Annexes, designates which data file(s) in the UCD are required for providing normative property information required by that algorithm.

For information on the meaning and application of the terms *normative, informative*, and *provisional*, see Section 3.5, "Properties" in [Unicode].

3 Documentation Files

The UCD also contains a number of documentation files, which provide information about the UCD as a whole, and about file formats, status, derivation of derived properties, and various other information.

3.1 UCD.html

UCD.html is the most important of the documentation files. It provides a complete listing of the UCD data files and character properties. It indicates which properties are normative and where they are defined. It provides further information required for the proper interpretation of some of the Unicode character properties.

UCD.html also records the modification history for the data files in the UCD, noting changes from version to version of the standard.

3.2 NamesList.html

NamesList.html formally describes (in BNF) the format of the NamesList.txt data file, the file which is used to drive the printing of the Unicode code charts and names list. See also Section 17.1, "Character Names List", in [Unicode] for a detailed discussion of the conventions used in the names list.

3.3 Unihan.html and UAX #38

Unihan.html describes the format and content of Unihan.txt, the data file which collects together all property information for CJK Unified Ideographs. As of Version 5.1.0 of the Unicode Standard, the content of Unihan.html has been incorporated into the new [UAX38], which is intended to supersede Unihan.html.

3.4 StandardizedVariants.html

StandardizedVariants.html documents standardized variants, showing a representative glyph for each. It is closely tied to the data file, StandardizedVariants.txt , which defines those sequences normatively.

3.5. Data File Comments

In addition to the specific documentation files for the UCD, individual data files often contain extensive header comments describing their content and any special conventions used in the data. In some instances, individual property definition sections are also commented with information about how the property may be derived.

4 Test Files

The UCD also contains a number of test data files, which specify, in standard formats, data which can be used to test implementation of Unicode algorithms.

4.1. NormalizationTest.txt

This file contains data which can be used to test an implementation of the Unicode Normalization Algorithm. (See [UAX15].)

4.2. LineBreakTest.txt

This file, located in the auxiliary directory of the UCD, contains data which can be used to test an implementation of the Unicode Linebreaking Algorithm. (See [UAX14].)

There is an associated documentation file, LineBreakTest.html, which displays the results of the Linebreaking Algorithm in an interactive chart form, with a documented listing of the rules.

4.3. Segmentation Test Files

The following three data files are also located in the auxiliary directory of the UCD:

- GraphemeBreakTest.txt
- SentenceBreakTest.txt
- WordBreakTest.txt

They contain data which can be used to test an implementation of the segmentation algorithms specified in [UAX29].

There are also associated documentation files, which display the results of the segmentation algorithms in an interactive chart form, with a documented listing of the rules:

- GraphemeBreakTest.html
- SentenceBreakTest.html
- WordBreakTest.html

5 UCD in XML

[UAX42] defines an XML schema which is used to incorporate all of the Unicode character property information into an XML version of the UCD.

Starting with Version 5.1.0, a set of XML data files using that schema are also released with each version of the UCD. Those data files make it possible to import and process the UCD property data using standard XML parsing tools, instead of the specialized parsing required for the various individual data files of the UCD.

Acknowledgments

Mark Davis and Ken Whistler are the authors of the initial version and have added to and maintained the text of this annex.

References

For references for this annex, see Unicode Standard Annex #41, “Common References for Unicode Standard Annexes.”

Modifications

The following summarizes modifications from previous revisions of this annex.

Revision 1

- Initial version