This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/standard/versions/enumeratedversions.html.

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, http://www.mehallo.com

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: http://www.awprofessional.com

Library of Congress Cataloging-in-Publication Data
The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-321-18578-1 (alk. paper)
1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004
005.7'2—dc21

2003052158

ISBN 0-321-18578-1
Text printed on recycled paper
1 2 3 4 5 6 7 8 9 10—CRW—0706050403
First printing, August 2003

Appendix D

Changes from Unicode Version 3.0

D.1 Versions of the Unicode Standard

The Unicode Technical Committee periodically updates the Unicode Standard to respond to the needs of implementers and users while maintaining consistency with ISO/IEC 10646.

The previous versions of the Unicode Standard are:

- *The Unicode Standard, Version 1.0*, Volume 1 (1991)
- *The Unicode Standard, Version 1.0*, Volume 2 (1992)
- *The Unicode Standard, Version 1.1*, Unicode Technical Report #4 (1993)
- *The Unicode Standard, Version 2.0* (1996)
- *The Unicode Standard, Version 2.1*, Unicode Technical Report #8 (1998)
- *The Unicode Standard, Version 3.0* (2000)
- *The Unicode Standard, Version 3.1,* Unicode Standard Annex #27 (2001)
- *The Unicode Standard, Version 3.2,* Unicode Standard Annex #28 (2002)

The relationship between these versions of Unicode and ISO/IEC 10646 is shown in *Table D-1*. For more detail on the relationship of Unicode and ISO/IEC 10646, see *Appendix C, Relationship to ISO/IEC 10646*.

Table D-1. Versions of Unicode and ISO/IEC 10646-1

Year	Version	Published	ISO/IEC 10646-1
1991	Unicode 1.0	Vol. 1, Addison-Wesley	Basis for Committee Draft 2 of 10646-1
1992	Unicode 1.0.1	Vol. 1, 2, Addison-Wesley	Interim merger version
1993	Unicode 1.1	Technical Report #4	Matches ISO 10646-1
1996	Unicode 2.0	Addison-Wesley	Matches ISO 10646-1 plus amendments
1998	Unicode 2.1	Technical Report #8	Matches ISO 10646-1 plus amendments
2000	Unicode 3.0	Addison-Wesley	Matches ISO 10646-1 second edition
2001	Unicode 3.1	Standard Annex #27	Matches ISO 10646-1 second edition plus two characters, 10646-2 first edition
2002	Unicode 3.2	Standard Annex #28	Matches ISO 10646-1 second edition plus amendment, 10646-2 first edition
2003	Unicode 4.0	Addison-Wesley	Matches ISO 10646:2003, third version

The Unicode Standard has grown from having 28,302 assigned graphic and format characters in Version 1.0, to having 96,382 characters in Version 4.0. *Table D-2*, *Table D-3*, and *Table D-4* document the number of code points allocated in the different versions of the Unicode Standard.

Table D-2. Allocation of BMP Code Points

	V 1.0	V 1.1	V 2.0	V 2.1	V 3.0	V 3.1	V 3.2	V4.0
Alphabetics, Symbols	4,748	6,309	6,509	6,511	10,236	10,238	11,195	11,649
Han (URO)	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902
Han Extension A					6,582	6,582	6,582	6,582
Han Compatibility [a]	302	302	302	302	302	302	361	361
Hangul Syllables	2,350	6,656	11,172	11,172	11,172	11,172	11,172	11,172
Graphic characters	**28,292**	**34,153**	**38,869**	**38,871**	**49,170**	**49,172**	**50,186**	**50,635**
Format	10	16	16	16	24	24	26	29
Control	65	65	65	65	65	65	65	65
Private Use	5,632	6,400	6,400	6,400	6,400	6,400	6,400	6,400
Code points assigned to abstract characters	**33,999**	**40,634**	**45,350**	**45,352**	**55,659**	**55,661**	**56,677**	**57,131**
Surrogate			2,048	2,048	2,048	2,048	2,048	2,048
Noncharacter	2	2	2	2	2	34	34	34
Designated code points	**34,001**	**40,636**	**47,400**	**47,402**	**57,709**	**57,743**	**58,759**	**59,213**
Undesignated code points (reserved)	31,535	24,900	18,136	18,134	7,827	7,793	6,777	6,323

[a] Includes 12 unified ideographs.

Table D-3. Allocation of Supplementary Code Points

	V 1.0	V 1.1	V 2.0	V 2.1	V 3.0	V 3.1	V 3.2	V4.0
Alphabetics, Symbols						1,691	1,691	2,465
Han Extension B						42,711	42,711	42,711
Han Compatibility						542	542	542
Graphic characters						**44,839**	**44,839**	**45,613**
Format						105	105	105
Private Use			131,068	131,068	131,068	131,068	131,068	131,068
Code points assigned to abstract characters			**131,068**	**131,068**	**131,068**	**176,012**	**176,012**	**176,786**
Noncharacter			32	32	32	32	32	32
Designated code points			**131,100**	**131,100**	**131,100**	**176,044**	**176,044**	**176,818**
Undesignated code points			917,476	917,476	917,476	872,532	872,532	871,758

Table D-4. Overall Allocation of Unicode Code Points

	V 1.0	V 1.1	V 2.0	V 2.1	V 3.0	V 3.1	V 3.2	V4.0
Graphic	28,292	34,153	38,869	38,871	49,170	94,011	95,025	96,248
Format	10	16	16	16	24	129	131	134
Control	65	65	65	65	65	65	65	65
Private Use	5,632	6,400	137,468	137,468	137,468	137,468	137,468	137,468
Code points assigned to abstract characters	**33,999**	**40,634**	**176,418**	**176,420**	**186,727**	**231,673**	**232,689**	**233,915**
Surrogate			2,048	2,048	2,048	2,048	2,048	2,048
Noncharacter	2	2	34	34	34	66	66	66
Designated code points	**34,001**	**40,636**	**178,500**	**178,502**	**188,809**	**233,787**	**234,803**	**236,029**
Undesignated code points (reserved)	31,535	24,900	935,612	935,610	925,303	880,325	879,309	878,083

This appendix summarizes updates to conformance specifications, character content, and data files made to the Unicode Standard, Version 4.0. For specific details on conformance requirements, always refer to *Chapter 3, Conformance*. Further information on all major, minor, and update versions of the Unicode Standard can be found on the Unicode Web site. Also see the subsection "Versions" in *Section B.4, Other Unicode References*.

D.2 Changes from Unicode Version 3.0 to Version 3.1

New Characters Added

44,946 new character assignments were made to the Unicode Standard, Version 3.1, including a very large collection of additional CJK ideographs, historic scripts, and several sets of symbols. The CJK ideograph additions provide significant coverage for dictionary and historical usage. For the first time, graphic and format characters were added to the supplementary planes:

- Supplementary Multilingual Plane (SMP), U+10000..U+1FFFF
- Supplementary Ideographic Plane (SIP), U+20000..U+2FFFF
- Supplementary Special-purpose Plane (SSP), U+E0000..U+EFFFF

Several historic scripts, including Old Italic, Gothic, and Deseret, and sets of symbols covering mathematical alphanumeric and musical symbols, were added to the Supplementary Multilingual Plane, or Plane 1. The Supplementary Ideographic Plane, or Plane 2, saw the addition of a very large collection of unified Han ideographs as well as additional Han compatibility ideographs. A set of 97 tag characters was added to the Supplementary Special-purpose Plane, or Plane 14.

Additionally, two mathematical symbols were added to the BMP, and 32 more code points were allocated as noncharacters. For more information on these character allocations, see the file DerivedAge.txt in the Unicode Character Database.

Unicode Character Database Changes

The Unicode Character Database (UCD) was extended to cover the character repertoire addition. The supplementary property list file, PropList.txt, was significantly reorganized, and a number of derived data files were added. New properties were added for case folding and scripts. All of the General Category values were made normative. Case mappings were also made normative.

Five-digit hex notation was used for the encoded supplementary characters.

Changes Affecting Conformance

There were four major changes affecting conformance. The first was the addition of new noncharacters and a clarification regarding noncharacter status. The second was a major corrigendum to the definition of UTF-8 to address security issues, by excluding non-shortest forms. The third was the inclusion of UTF-32 as part of this standard. See *Chapter 3, Conformance*, for the updated definitions and conformance clauses. The fourth was a corrigendum affecting normalization: U+FB1D was added to CompositionExclusions.txt in the Unicode Character Database.

To allow for finer control over ligature formation, the semantics of U+200D ZERO WIDTH JOINER and U+200C ZERO WIDTH NON-JOINER were broadened to cover ligatures, as well as

cursive connection. Additionally, the Unicode Character Encoding Stability policy was documented.

Unicode Standard Annexes

The following Technical Reports were upgraded in status to Unicode Standard Annexes:

- UAX #9: The Bidirectional Algorithm
- UAX #19: UTF-32

D.3 Changes from Unicode Version 3.1 to Version 3.2

New Characters Added

1,016 new character assignments were made to the Unicode Standard, Version 3.2. These additions included a large collection of mathematical symbols in support of MathML, other symbols such as recyling symbols, minority Philippine scripts, and a number of special characters, including U+034F COMBINING GRAPHEME JOINER and U+2060 WORD JOINER. The additional symbol sets benefit technical publishing needs.

All new character additions were to the BMP. For more information on these character allocations, see the file DerivedAge.txt in the Unicode Character Database.

Unicode Character Database Changes

The Unicode Character Database (UCD) was extended to cover the character repertoire additions, and a number of other updates were made:

- PropertyAliases.txt and PropertyValueAliases.txt were made normative when used to refer to Unicode properties or property values.
- Normative blocks defined in Blocks.txt were adjusted slightly.
- A specification of when variation selectors can be used was added.
- New properties were added, including properties for ideographic description categories, code points that are ignorable by default, deprecated characters, IDS operators, Han properties, grapheme properties, and the soft dotted property.

Changes Affecting Conformance

There was a major corrigendum to the definition of UTF-8 to eliminate irregular sequences and bring the Unicode specification more in line with other specifications of UTF-8. In addition, the canonical decomposition mapping for U+F951 was corrected to map to U+964B.

Significant clarifications or modifications to character behavior include the following:

Word Joiner. U+2060 WORD JOINER was defined as the preferred character to express the word joining semantics previously implied by U+FEFF ZERO WIDTH NO-BREAK SPACE. This leaves ZERO WIDTH NO-BREAK SPACE to be used solely with the semantic of the byte order mark (BOM).

Special Properties. A number of characters with special properties, including boundary control, joining, and variation selection, were added to the Unicode Standard in Version

3.2. See *Section 4.11, Characters with Unusual Properties*, and *Chapter 15, Special Areas and Format Characters*, for more information.

Behavior of Hangul Syllables, Conjoining Jamo, and Combining Marks. Discussions of the application of combining marks to Hangul syllables and the behavior of syllable boundaries in a sequence of conjoining jamo were updated. See *Section 3.11, Canonical Ordering Behavior*, and *Section 3.12, Conjoining Jamo Behavior*.

Unicode Standard Annexes

The following Technical Report was upgraded in status to a Unicode Standard Annex:

- UAX #21: Case Mappings

D.4 Changes from Unicode Version 3.2 to Version 4.0

New Characters Added

1,226 new character assignments were made to the Unicode Standard, Version 4.0. These additions include currency symbols, additional Latin and Cyrillic characters, the Limbu and Tai Le scripts; Yijing Hexagram symbols, Khmer symbols, Linear B syllables and ideograms, Cypriot, Ugaritic, and a new block of variation selectors. Double diacritic characters were added for dictionary use. In total, 452 characters were added to the BMP; 774 were added to the supplementary planes.

These new characters extend the set of modern currency symbols, and represent a greater coverage of minority and historical scripts. For more information on the allocations, see the file DerivedAge.txt in the Unicode Character Database.

In addition, substantial improvements were made to the script descriptions, particularly for Indic scripts.

Unicode Character Database Changes

Unicode 4.0 introduced the concept of provisional properties, clarified the relationships between properties, and provided precisely defined fallback properties for characters not explicitly defined in the data files.

Other property changes include the following:

Prefix Format Control. U+06DD ARABIC END OF AYAH and U+070F SYRIAC ABBREVIATION MARK were reclassified and have significantly different behavior as prefix format control characters. The new characters U+0600..U+0603 were given this behavior as well.

New Properties. The Hangul Syllable Type and identifier Other_ID_Start properties were added. The Unicode Radical Stroke property was classified as informative; all other Unihan properties were classified as provisional.

Soft Hyphen. U+00AD SOFT HYPHEN was changed to General Category Cf and other ignorable.

Modifier Letters. The General Category of U+02B9..U+02BA, U+02C6..U+02CF changed to General Category Lm.

Mongolian Vowel Separator. U+180E MONGOLIAN VOWEL SEPARATOR was changed to General Category Zs.

Deprecated Characters. Two Khmer characters, U+17A3 KHMER INDEPENDENT VOWEL QAQ and U+17D3 KHMER SIGN BATHAMASAT, were deprecated.

For more information, see the file UCD.html in the Unicode Character Database.

Changes Affecting Conformance

Chapter 3, Conformance, was substantially improved by incorporating the Unicode Character Encoding Model, resulting in fully specified definitions and conformance requirements of UTF-8, UTF-16, and UTF-32. Clearer terminology was introduced for code point assignments. The conformance section of UAXs, UTSs and UTRs was clarified.

Identifiers. A structure for ensuring backward-compatible programming language identifiers was introduced using the new property Other_ID_Start.

Bidi. The bidirectional algorithm was made to be invariant under canonical equivalence.

Line Breaking and Boundaries. U+00AD SOFT HYPHEN was reclassified. Text boundaries were clarified.

Case Folding. UAX #21, "Case Mappings," was updated for case folding and other new properties.

Unicode Standard Annexes

The following Unicode Standard Annex was added:

- UAX #29: Text Boundaries

UAX #29, "Text Boundaries," now contains information on text boundary conditions fomerly published in Chapter 5 of *The Unicode Standard, Version 3.0*.

The following Unicode Technical Report was upgraded in status to a Unicode Standard Annex:

- UAX #24: Script Names

The following Standard Annexes were superseded as a result of their incorporation into the text of this book:

- UAX #13: Unicode Newline Guidelines
- UAX #19: UTF-32
- UAX #21: Case Mappings
- UAX #27: Unicode 3.1
- UAX #28: Unicode 3.2

UAX #9, "The Bidirectional Algorithm," now contains information on the bidirectional algorithm formerly published in Chapter 3 of *The Unicode Standard, Version 3.0*.

Errata

An itemized list of errata rolled up since the publication of the Unicode Standard, Version 3.2 can be found online. See "Updates and Errata" in *Appendix B.4, Other Unicode References*.