Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

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

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Appendix D

Changes from Previous Versions

This appendix provides version history of the standard and summarizes updates that have been made to conformance specifications, character content, and data files in the Unicode Character Database since the publication of *The Unicode Standard, Version 3.0,* and *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.6, Other Unicode Online Resources*.

D.1 Improvements to the Standard

Version 5.0 of the Unicode Standard incorporates into the text the knowledge gained from many years of worldwide industry implementation experience. It supersedes all previous versions and offers round-trip compatibility with the Chinese standards GB18030 and HKSCS, improved alignment of the Bidirectional Algorithm with norms of the industry, improved guidance on the segmentation of text and processing Unicode strings, and enhanced descriptions of rendering Indic scripts.

This latest version of the Unicode Standard is the basis for Unicode security mechanisms, the Unicode collation algorithm, the locale data provided by the Common Locale Data Repository, and support for Unicode in regular expressions. The significant improvements to the standard since Versions 3.0 and 4.0 include the further development of the Unicode encoding model, the introduction of the character property model, and the establishment of casing and identifier stability.

The text of the standard has been enhanced significantly:

- Two thirds of the definitions are new.
- One third of the conformance clauses are new.
- One half of the character repertoire is new.
- One fourth of the tables are new.
- Four fifths of the figures are new or updated.

- One half of the Unicode Standard Annexes are new.
- All Unicode Standard Annexes are included in the book for the first time.
- The form factor has been improved dramatically to make the book smaller and lighter.

D.2 Versions of the Unicode Standard

The Unicode Technical Committee updates the Unicode Standard to respond to the needs of implementers and users while maintaining consistency with ISO/IEC 10646. 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
2005	Unicode 4.1	Web publication	Matches ISO 10646:2003, third version, plus Amd. 1
2006	Unicode 5.0	Addison-Wesley (2007)	Matches ISO 10646:2003, third version, plus Amd. 1, Amd. 2, and four characters from Amd. 3

The Unicode Standard has grown from having 28,294 assigned graphic and format characters in Version 1.0, to having 99,024 characters in Version 5.0. *Table D-2* documents the number of code points allocated in the different versions of the Unicode Standard. The row in *Table D-2* labeled "Graphic + Format" represents the traditional count of Unicode characters and is the typical answer to the question, "How many characters are in the Unicode Standard?"

Some of the values in *Table D-2* differ slightly from summary statistics published in earlier versions of the standard, primarily due to a refined accounting of the allocations in Unicode 1.0. Also note that the numbers for Han Compatibility include the 12 unified ideographs encoded in the CJK Compatibility Ideographs block.

Table D-2. Allocation of Code Points by Type

	V1.0.0	V1.0.1	V1.1	V2.0	V2.1	V3.0	V3.1	V3.2	V4.0	V4.1	V5.0
Alphabetics, Symbols	4,734	4,728	6,290	6,491	6,493	10,210	11,798	12,753	13,973	15,117	16,486
Han (URO)		20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902
Han (URO Extension)										22	22
Han Extension A						6,582	6,582	6,582	6,582	6,582	6,582
Han Extension B							42,711	42,711	42,711	42,711	42,711
Han Compatibility		302	302	302	302	302	844	903	903	1,009	1,009
Subtotal Han		21,204	21,204	21,204	21,204	27,786	71,039	71,098	71,098	71,226	71,226
Hangul Syllables	2,350	2,350	6,656	11,172	11,172	11,172	11,172	11,172	11,172	11,172	11,172
Graphic Characters	7,084	28,282	34,150	38,867	38,869	49,168	94,009	95,023	96,243	97,515	98,884
Format Characters	12	12	18	18	18	26	131	133	139	140	140
Graphic + Format	7,096	28,294	34,168	38,885	38,887	49,194	94,140	95,156	96,382	97,655	99,024
Controls	65	65	65	65	65	65	65	65	65	65	65
Private Use	5,632	6,144	6,400	137,468	137,468	137,468	137,468	137,468	137,468	137,468	137,468
Total Assigned	12,793	34,503	40,633	176,418	176,420	186,727	231,673	232,689	233,915	235,188	236,557
Surrogate Code Points				2,048	2,048	2,048	2,048	2,048	2,048	2,048	2,048
Noncharacters	2	2	2	34	34	34	66	66	66	66	66
Total Designated	12,795	34,505	40,635	178,500	178,502	188,809	233,787	234,803	236,029	237,302	238,671
Reserved Code Points	52,741	31,031	24,901	935,612	935,610	925,303	880,325	879,309	878,083	876,810	875,441

D.3 Clause and Definition Numbering Changes

In the time since the publication of *The Unicode Standard, Version 2.0*, there have been very substantial additions to the numbered conformance clauses and formal definitions in the standard. In prior versions, some effort was made to keep the numbering of clauses and definitions consistent between versions. However, with Version 5.0, it has proven necessary to renumber completely.

To assist in comparison of Version 5.0 to earlier versions of the standard, *Table D-3* provides a cross-mapping of clause and definitions numbers between Version 5.0 and other major versions.

Table D-3. Clause and Definition Numbering

V5.0	V4.0	V3.0	V2.0
n/a	C1–C3		
C1	C4	C4	C4
C2	C5	C5	C5
C3	C6	C6	C6
C4	C7	C7	C7
C5	C8	C8	C8
C6	C9	C9	C9
C7	C10	C10	C10
C8	C11	C1, C2, C11	C1, C2
C9	C12	C12	
C10	C12a	C12	
C11	C12b	C3	C3
C12	C13	C13	
C13	C14		
C14	C15		
C15	C16		
C16	C17		
C17	C18		
C18	C19		
C19	C19a		
C20	C20		
D1	D1	D1	D1
n/a	D2		
D2	D2a		
D3	D2b	D2	D2
D4–D6			
D7	D3	D3	D3
D8	D4	D4	D4
D9	D4a		
D10	D4b		
D11	D5		
D12	D6	D6	D6
n/a	D7		
D13	D7a	D7a	
D14	D7b		
D15	D7c		
D16	D8	D8	D8
D17	D8a		
D18–D32			
n/a	n/a	D9	D9
n/a	n/a	D10	D10
n/a	n/a	D10a	
n/a	n/a	D10b	
n/a	n/a	D11	D11
D33	D9		
D34			
D35	D9a		
D36	D9b		
D37–D44			
D45	D9c		
D46	D9d		
D47	D10		
D48	D10a		
n/a	D11		
D49	D12	D12	D12
n/a	D13	D13	D13
D50–D51			
D52	D14	D14	D14

Table D-3. Clause and Definition Numbering (Continued)

V5.0	V4.0	V3.0	V2.0
D53	D15	D15	D15
D54			
D55	D16	D16	D16
D56	D17	D17	D17
D57	D17a	D17a	
D58–D62			
D63	D18	D18	D18
D64	D19	D19	D19
D65	D20	D20	D20
D66	D21	D21	D21
D67	D22	D22	D22
D68	D23	D23	D23
D69	D23a		
D70	D24	D24	D24
D71	D25	D25	D25
D72	D25a		
D73	D26	D26	D26
D74	D26a		
D75	D27	D27	D27
D76	D28	D28	D28
D77	D28a	D5	D5
D78	D28b	D7	D7
D79	D29	D29	
D80	D29a		
D81	D29b		
D82	D29c		
D83	D29d		
D84	D30	D30	
D85	D30a		
D86	D30b		
D87	D30c		
D88	D30d		
D89	D30e		
n/a	n/a	D31	
D90	D31		
n/a	D32	D32	
n/a	D33–D34		
D91	D35		
D92	D36		
D93	D37		
D94	D38		
D95	D39	D36	
D96	D40	D33	
D97	D41	D34	
D98	D42	D35	
D99	D43		
D100	D44		
D101	D45		
D102–D103			
D104	D46	D37	D29
D105–D119			
D120	D47		
D121–D122			
D123	D48		
D124–D131			

An entry "n/a" in the Version 5.0 (V5.0) or Version 4.0 (V4.0) columns indicates a clause or definition that has become obsolete or has been superseded. In some cases, because of rewording of clauses or definitions over time, two or more entries from an earlier version might correspond to a single Unicode 5.0 entry, or vice versa.

D.4 Changes from Version 4.1 to Version 5.0

New Characters Added

In total, 1,369 new character assignments were made to the Unicode Standard, Version 5.0. These additions include new characters for Cyrillic, Greek, Hebrew, Kannada, Latin, math, phonetic extensions, symbols, and five new scripts: Balinese, N'Ko, Phags-pa, Phoenician, and Sumero-Akkadian Cuneiform.

The new character additions were to both the BMP and the SMP (Plane 1). For more information on these character allocations, see the file DerivedAge.txt in the Unicode Character Database (UCD).

Unicode Character Database Changes

The Unicode Character Database was extended to cover the character repertoire additions, and new block definitions and script values were added. A number of other updates were made, as listed here.

New Properties. Normative_Name_Alias was added to provide alternate identifiers for characters with problems in their names. The metaproperty termed "Deprecated" was added. The Jamo_Short_Name property was documented as a contributory property in PropertyAliases.txt and PropertyValueAliases.txt.

General Category. U+103D0 OLD PERSIAN WORD DIVIDER was changed to Po. The bracket characters U+23B4..U+23B6 were changed to So. U+0294 LATIN LETTER GLOTTAL STOP was changed to Lo. U+2132 TURNED CAPITAL F and U+2183 ROMAN NUMERAL REVERSED ONE HUNDRED were changed to Lu. U+10341 GOTHIC LETTER NINETY was changed to Nl.

Numeric Properties. The Numeric_Type of U+10341 GOTHIC LETTER NINETY was changed from None to Numeric, and it was given the Numeric_Value 90.

Bidirectional Behavior. The Bidi_Class property for Old Persian numerals U+103D1..U+103D5 and U+2132 TURNED CAPITAL F was changed to L. The list of characters with the Bidi_Mirrored property was made consistent for brackets and quotation marks.

Scripts. Unassigned code points were given a new Script property value of "Zzzz". Three Mongolian punctuation marks, U+1802, U+1803, and U+1805, were changed to Script=Common. U+1DBF MODIFIER LETTER SMALL THETA was changed to Script=Greek. U+2132 TURNED CAPITAL F was changed to Script=Latin.

Unihan. The kIICore field was made a normative property, and three new provisional properties were added: kCheungBauer, kCheungBauerIndex, and kFourCornerCoverage. There were numerous additions to the kCangjie property.

Text Breaking. Grapheme_Link was deprecated as a property and moved from PropList.txt to DerivedCoreProperties.txt.

Line Break. The Line_Break property of several punctuation characters (U+1735, U+1736, U+17D9, U+203D, U+2047..U+2049) and bracket characters (U+23B4..U+23B6) was changed to AL to better match their expected behavior. Numerous characters for Southeast Asian scripts, which require complex contextual line breaking, were changed to SA.

Case-Related Properties. The addition of the second member of a few case pairs (such as U+0243 LATIN CAPITAL LETTER B WITH STROKE) led to the revision of the uppercase, lowercase, and titlecase mappings for the already-encoded first member of the case pair.

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

Changes Affecting Conformance and Stability

Chapter 3, Conformance, was substantially improved by incorporating much of the Unicode Property Model, enhancing the treatment of combining characters, and further clarifying canonical ordering behavior through the addition of clearly defined principles. Additionally, conformance clauses and definitions were renumbered for overall readability and clarity of the text.

Significant clarifications or modifications to character behavior include those listed below.

Bidirectional Behavior. The Bidirectional Algorithm was modified to tighten up the conformance requirements for characters with the Bidi_Mirrored property.

Stability of Cased Letters. If uppercase characters are added in cased scripts, the corresponding lowercase characters will be added as well, so that case folding is stable.

Stability of Named Character Sequences. An initial provisional phase was incorporated into the process for defining Named Character Sequences, so that approved Named Character Sequences will be immutable.

Disunification of Diacritics. Criteria for disunifying diacritics were established.

Indic Scripts. Descriptions of Indic scripts were improved substantially. ZERO WIDTH JOINER and ZERO WIDTH NON-JOINER can now be used to encourage or discourage ligation in Bengali; the sequence for Gurmukhi double vowels was determined; and the shaping of *ra* in Tamil was updated.

Combining Marks. The use of the combining grapheme joiner with Latin script diacritics was clarified.

Unicode Standard Annexes

Changes to the Unicode Standard Annexes were made, as listed here.

> In UAX #9, "Bidirectional Algorithm," the definition of directional run was changed to be the same as level run, rule L4 and HL6 were updated in conjunction with the change in handling Bidi_Mirrored, and a caution was added on the use of higher-level protocols.

In UAX #14, "Line Breaking Properties," a number of rules were modified, the use of soft hyphen in cursive scripts was documented, the conformance clauses were restated and the algorithm was reorganized into tailorable and non-tailorable sections, and the normative status was made consistent with *Chapter 3, Conformance*. As a result of the restatement of conformance, the Line_Break property became normative.

In UAX #15, "Unicode Normalization Forms," the stability section was updated.

In UAX #29, "Text Boundaries," the definition of numeric was changed for both word and sentence breaks, the definition of ALetter was tied to the Line_Break property SA, text was added to explain why modifier letters are not parts of words, breaks within CRLF were forbidden, and rule 0 of section 4 was removed.

UAX #31, "Identifier and Pattern Syntax," introduced profiles and added notes on profiles of identifiers for natural languages and the use of spaces in identifiers.

Errata

An itemized list of errata incorporated since the publication of the Unicode Standard, Version 4.1, can be found online. See "Updates and Errata" in *Section B.6, Other Unicode Online Resources*.

D.5 Changes from Version 4.0 to Version 4.1

New Characters Added

In total, 1,273 new character assignments were made to the Unicode Standard, Version 4.1. These additions include characters to complete round-trip mapping of the HKSCS and GB 18030 standards; five new currency signs; new characters for Arabic, Ethiopic, Hebrew, Indic, and Korean; and eight new scripts: Buginese, Coptic, Glagolitic, Kharoshthi, New Tai Lue, Old Persian, Syloti Nagri, and Tifinagh. The addition of the Coptic script constituted a disunification of Coptic from Greek.

The Nuskhuri forms of Khutsuri Georgian were added. The new Nuskhuri forms are now to be taken as the lowercase pairs of the Asomtavruli Georgian and are a change from the previous documentation about Georgian.

The new character additions were to both the BMP and the SMP (Plane 1). For more information on the character allocations, see the file DerivedAge.txt in the Unicode Character Database.

Unicode Character Database Changes

The Unicode Character Database was extended to cover the character repertoire additions, and new block definitions and script values were added. A number of other updates were made, including those listed below.

New Properties. Grapheme_Cluster_Break, Sentence_Break, Word_Break, and STerm were added in support of text boundary determination. Other_ID_Continue was added to support identifier stability. Pattern_Syntax and Pattern_White_Space were added in support of pattern syntax matching. The property Variation_Selector was added as a convenience for referring to all variation selector characters.

Case Mapping. The case mapping contexts defined in SpecialCasing.txt were updated and supersede *Table 3-13*, "Context Specification for Casing," in *The Unicode Standard, Version 4.0*.

Alphabetic. The Alphabetic property was modified to be a superset of Lowercase and Uppercase for compatibility with POSIX-style character classes. The uppercase and lowercase circled letters A through Z, U+24B6..U+24E9, were added to Other_Alphabetic.

Bidirectional Behavior. The values of the Bidi_Class property were harmonized for a few compatibility equivalents of characters whose Bidi_Class values changed for Unicode 4.0.1. The Bidi_Class property for Braille symbols was changed to L. The properties for characters related to number and date formatting were changed.

General Category. U+30FB KATAKANA MIDDLE DOT and U+FF65 HALFWIDTH KATAKANA MIDDLE DOT were changed from Pc to Po. The Ethiopic digits (U+1369 and following) were changed from Nd to No. U+200B ZERO WIDTH SPACE was changed to Cf. U+A015 YI SYLLABLE WU was changed from Lo to Lm.

Numeric Properties. The Numeric_Type of U+1034A GOTHIC LETTER NINE HUNDRED was changed from None to Numeric, and it was given the Numeric_Value 900.

New Data Files. NamedSequences.txt was added. This data file defines specific names for some significant Unicode character sequences, giving their Unicode Sequence Identifier (USI) values.

Line Break. The Line_Break properties of Runic, certain Indic characters, Mongolian, Tibetan punctuation, Hangul, and spacing clones of European diacritical marks were revised to better match their expected behavior.

Unihan. Five provisional properties were added: kFennIndex, kGSR, kHDZRadBreak, kHanyuPinlu, and kRSAdobe_Japan1_6. kIICore was added as an informative property. kAlternateKangXi and kAlternateMorohashi were dropped.

Block Ranges. The end of the CJK Unified Ideographs range was changed from U+9FA5 to U+9FBB.

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

Changes Affecting Conformance and Stability

Several updates were made to *Chapter 3, Conformance*. *Section 3.13, Default Case Operations*, was updated to define case-ignorable characters and case-ignorable sequences. *Table 3-13*, "Context Specification for Casing," was replaced by a description of each context followed by the equivalent regular expression(s) describing the context before a character and the context after a character, or both. In *Section 3.7, Decomposition*, the phrase "according to the decomposition mappings found in the names list of *Section 16.1, Character Names List*," was changed to "according to the decomposition mappings found in the Unicode Character Database" in the relevant definitions.

Significant clarifications or modifications to character behavior include those listed below.

Space and No-Break Space. U+0020 SPACE is no longer recommended as a suitable base character for display of isolated nonspacing marks. Instead, U+00A0 NO-BREAK SPACE is the preferred base character for this function. Additionally, NO-BREAK SPACE is no longer considered equivalent to <U+FEFF ZERO WIDTH NO-BREAK SPACE, U+0020 SPACE, U+FEFF ZERO WIDTH NO-BREAK SPACE>.

Combining Grapheme Joiner. The function of U+034F COMBINING GRAPHEME JOINER was clarified.

Other Combining Characters. The control of the positioning of U+05BD HEBREW POINT METEG and the rendering of Thai combining marks were clarified.

Indic. U+09CE BENGALI LETTER KHANDA TA was added. This necessitates adjustment of Bengali script implementations.

Arabic. The joining type of U+06C2 ARABIC LETTER HEH GOAL WITH HAMZA ABOVE was changed to dual joining to align with U+06C1 ARABIC LETTER HEH GOAL.

Bidirectional Behavior. The Bidi_Class properties for +, -, and / were changed. This change affected number and date formatting. It was made to better align with industry practice.

Stability of Numeric Assignments. All characters with the property value Numeric_Type= Numeric are guaranteed to have a non-null Numeric_Value.

Unicode Standard Annexes

The following Unicode Standard Annexes were added:

- UAX #31: Identifier and Pattern Syntax
- UAX #34: Unicode Named Character Sequences

Changes to the Unicode Standard Annexes were made, as listed here.

> UAX #9, "Bidirectional Algorithm," added a note after N1 and clarified the example after N2. Overriding expected behavior by a higher-level protocol was constrained.

UAX #14, "Line Breaking Properties," was updated.

UAX #15, "Unicode Normalization Forms," corrected definition D2, which defines *blocked* characters.

In UAX #29, "Text Boundaries," the definition of numeric for both word and sentence breaks was changed to include all characters of the General Category Nd plus U+066B ARABIC DECIMAL SEPARATOR and U+066C ARABIC THOUSANDS SEPARATOR.

UAX #31, "Identifier and Pattern Syntax," was updated to contain information on identifiers formerly published in *Section 5.15* of *The Unicode Standard, Version 4.0*, as well as information formerly published in Annex 7 of UAX #15, "Unicode Normalization Forms."

Errata

An itemized list of errata incorporated since the publication of *The Unicode Standard, Version 4.0* can be found online. See "Updates and Errata" in *Section B.6, Other Unicode Online Resources*.

D.6 Changes from Unicode Version 3.2 to Version 4.0

New Characters Added

In total, 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

The Unicode Character Database was extended to cover the character repertoire additions, and new block definitions and script values were added. A number of other updates were made, including those listed below.

Provisional and Fallback Properties. 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.

New Properties and Values. The Hangul_Syllable_Type and Other_ID_Start properties were added. For Unihan.txt, the Unicode_Radical_Stroke property was added and classified as informative; all other non-normative Unihan properties were classified as provisional. A number of numeric values were added for CJK ideographs.

General Category. U+00AD SOFT HYPHEN was changed to Cf, which also resulted in it gaining the Default_Ignorable_Code_Point property. Modifier letters U+02B9..U+02BA and U+02C6..U+02CF were changed to Lm. U+180E MONGOLIAN VOWEL SEPARATOR was changed to Zs.

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

Stabilized Properties. The Hyphen property was stabilized.

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

Changes Affecting Conformance and Stability

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. In addition, the conformance section of UAXes, UTSes and UTRs was clarified. A section on default case operations was added, based on material incorporated from UAX #21, "Case Mappings."

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

Bidirectional Behavior. The Bidirectional Algorithm was made to be invariant under canonical equivalence.

Significant clarifications or modifications to character behavior include those listed below.

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

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.

Unicode Standard Annexes

The following Unicode Standard Annex was added:

- UAX #29: Text Boundaries

UAX #29, "Text Boundaries," was updated to contain information on text boundary conditions formerly 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 the 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," was updated to contain information on the Bidirectional Algorithm formerly published in Chapter 3 of *The Unicode Standard, Version 3.0*.

Errata

An itemized list of errata incorporated since the publication of the Unicode Standard, Version 3.2 can be found online. See "Updates and Errata" in *Section B.6, Other Unicode Online Resources*.

D.7 Changes from Unicode Version 3.1 to Version 3.2

New Characters Added

In total, 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 was extended to cover the character repertoire additions, and new block definitions and script values were added. A number of other updates were made, including those listed below.

Property Aliases. The new data files PropertyAliases.txt and PropertyValueAliases.txt were introduced, normatively specifying aliases for character properties and for the values of character properties.

Blocks. Normative blocks defined in Blocks.txt were adjusted slightly.

Variation Selectors. A specification of when variation selectors can be used was added.

New Properties. 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.

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

Changes Affecting Conformance

There was a significant update to the definition of UTF-8 to eliminate irregular sequences and bring the Unicode specification more in line with other specifications of UTF-8. Corrigendum #3 corrected the canonical decomposition mapping for U+F951 to map to U+964B.

Significant clarifications or modifications to character behavior include those listed below.

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.

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.

Unicode Standard Annexes

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

- UAX #21: Case Mappings

Errata

An itemized list of errata incorporated since the publication of the Unicode Standard, Version 3.1 can be found online. See "Updates and Errata" in *Section B.6, Other Unicode Online Resources*.

D.8 Changes from Unicode Version 3.0 to Version 3.1

New Characters Added

In total, 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 was extended to cover the character repertoire addition, and new block definitions and script values were added. A number of other updates were made, including those listed below.

PropList.txt. The supplementary property list file, PropList.txt, was significantly reorganized.

New Properties. A number of derived data files were added, and new properties were added for case folding and scripts.

Hex Notation. The convention of using five-digit hex notation for the representation of supplementary characters was introduced.

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

Changes Affecting Conformance and Stability

There were four major changes affecting conformance. The first was the addition of new noncharacters and a clarification regarding noncharacter status. The second was Corrigen-

dum #1, which updated 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 the standard. The fourth was Corrigendum #2, affecting normalization: U+FB1D was added to CompositionExclusions.txt in the Unicode Character Database.

Significant clarifications or modifications to character behavior include those listed below.

Ligature Formation. 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.

Stability Policy. The Unicode Character Encoding Stability policy was documented.

New Normative Properties. All of the General Category values and case mappings were made normative.

Unicode Standard Annexes

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

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

Errata

An itemized list of errata incorporated since the publication of the Unicode Standard, Version 3.0 can be found online. See "Updates and Errata" in *Section B.6, Other Unicode Online Resources*.

