

Contents

Acknowledgments	iii
Unicode Consortium Members and Directors	vi
Current Full Members	vi
Current Associate Members	vi
Current Liaison Members	vi
Current Members of the Board of Directors	vi
Former Members of the Board of Directors	vi
Contents	vii
Figures	xiii
Tables	xv
Preface	xvii
1 Introduction	1-1
1.1 Design Goals	1-2
1.2 Coverage	1-2
1.3 About This Book	1-4
Notational Conventions	1-5
1.4 The Unicode Consortium	1-5
The Unicode Technical Committee	1-6
1.5 The Unicode Standard and ISO/IEC 10646	1-6
1.6 Resources	1-7
On-line Information Sources	1-7
How to Contact the Unicode Consortium	1-7
2 General Structure	2-1
2.1 Architectural Context	2-1
Basic Text Processes	2-1
Text Elements, Code Elements, and Text Processes	2-2
Text Processes and Encoding	2-2
2.2 Unicode Design Principles	2-3
Sixteen-Bit Characters	2-3
Full Encoding	2-4
Characters, Not Glyphs	2-4
Semantics	2-5
Plain Text	2-5
Logical Order	2-7
Unification	2-8
Dynamic Composition	2-9
Equivalent Sequence	2-9

Convertibility	2-9
2.3 Unicode Allocation	2-10
Allocation Areas	2-10
Code Space Assignment for Graphic Characters	2-12
Non-Graphic Characters, Reserved and Unassigned Codes	2-12
2.4 Special Character and Non-Character Values	2-13
Byte Order Mark	2-13
Special Non-Character Values	2-13
Separators	2-14
Layout and Format Control Characters	2-14
The Replacement Character	2-14
2.5 Combining Characters	2-14
Sequence of Base Characters and Diacritics	2-15
Multiple Combining Characters	2-15
Multiple Base Characters	2-17
Spacing Clones of European Diacritical Marks	2-17
2.6 Controls and Control Sequences	2-18
Control Characters	2-18
Representing Control Sequences	2-18
2.7 Conforming to the Unicode Standard	2-19
Characters Not Used in a Subset	2-20
3 Conformance	3-1
3.1 Conformance Requirements	3-1
Byte Ordering	3-1
Invalid Code Values	3-2
Interpretation	3-2
Modification	3-3
3.2 Semantics	3-3
3.3 Characters and Coded Representations	3-4
3.4 Simple Properties	3-4
3.5 Combination	3-5
3.6 Decomposition	3-6
Compatibility Decomposition	3-6
Canonical Decomposition	3-7
3.7 Surrogates	3-7
3.8 Special Character Properties	3-8
3.9 Canonical Ordering Behavior	3-9
Combining Classes	3-10
Canonical Ordering	3-10
3.10 Combining Jamo Behavior	3-11
Syllable Boundaries	3-12
Canonical Syllables	3-12
Hangul Syllable Composition	3-12
Hangul Syllable Decomposition	3-13
Hangul Syllable Name	3-14
3.11 Bidirectional Behavior	3-14
Directional Formatting Codes	3-15
Basic Display Algorithm	3-16
Bidirectional Character Types	3-17

Resolving Embedding Levels	3-17
Reordering Resolved Levels	3-21
Bidirectional Conformance	3-22
Usage	3-23
4 Character Properties	4-1
Disclaimer	4-1
4.1 Case	4-2
4.2 Combining Classes	4-2
4.3 Directionality	4-10
4.4 Jamo Short Names	4-12
4.5 Letters	4-14
4.6 Numeric Value	4-15
4.7 Mirrored	4-22
4.8 Unicode 1.0 Names	4-25
4.9 Mathematical Property	4-25
5 Implementation Guidelines	5-1
5.1 ANSI/ISO C wchar_t	5-1
5.2 Compression and Transmission	5-2
7-bit or 8-bit Transmission	5-2
5.3 Language Information	5-3
5.4 Unknown and Missing Characters	5-3
Unassigned and Private Use Character Codes	5-3
Interpretable but Unrenderable Characters	5-4
5.5 Handling Surrogate Characters	5-4
5.6 Handling Numbers	5-6
5.7 Transcoding to Other Standards	5-6
Disclaimer	5-7
Issues	5-7
Tables and Virtual Memory	5-7
5.8 Handling Properties	5-8
5.9 Normalization	5-9
5.10 Editing and Selection	5-10
Consistent Text Elements	5-10
5.11 Strategies for Handling Non-Spacing Marks	5-11
Keyboard Input	5-12
Truncation	5-13
5.12 Rendering Non-Spacing Marks	5-14
Positioning Methods	5-16
5.13 Locating Text Element Boundaries	5-18
Boundary Specification	5-18
Example Specifications	5-20
Character Boundaries	5-21
5.14 Identifiers	5-25
Terminology	5-26
Terminal Classes	5-26
Syntactic Rules	5-26
Character Properties	5-27

5.15	Sorting and Searching	5-27
	Culturally Expected Sorting	5-27
	Unicode Character Equivalence	5-28
	Similar Characters	5-28
	Levels of Comparison	5-29
	Ignorable Characters	5-30
	Multiple Mappings	5-31
	Collating Out-of-Scope Characters	5-31
	Unmapped Characters	5-32
	Parameterization	5-32
	Optimizations	5-32
	Searching	5-32
	Sublinear Searching	5-33
6	Character Block Descriptions	6-1
6.1	General Scripts Area	6-2
	Basic Latin: U+0000—U+007F	6-3
	Latin-1 Supplement: U+0080—U+00FF	6-5
	Latin Extended-A: U+0100—U+017F	6-7
	Latin Extended-B: U+0180—U+024F	6-8
	IPA Extensions: U+0250—U+02AF	6-10
	Spacing Modifier Letters: U+02B0—U+02FF	6-12
	Combining Diacritical Marks: U+0300—U+036F	6-14
	Greek: U+0370—U+03FF	6-16
	Cyrillic: U+0400—U+04FF	6-18
	Armenian: U+0530—U+058F	6-19
	Hebrew: U+0590—U+05FF	6-20
	Arabic: U+0600—U+06FF	6-22
	Devanagari: U+0900—U+097F	6-33
	Bengali: U+0980—U+09FF	6-45
	Gurmukhi: U+0A00—U+0A7F	6-46
	Gujarati: U+0A80—U+0AFF	6-47
	Oriya: U+0B00—U+0B7F	6-48
	Tamil: U+0B80—U+0BFF	6-49
	Telugu: U+0C00—U+0C7F	6-54
	Kannada: U+0C80—U+0CFF	6-55
	Malayalam: U+0D00—U+0D7F	6-56
	Thai: U+0E00—U+0E7F	6-57
	Lao: U+0E80—U+0EFF	6-58
	Tibetan: U+0F00—U+0FBF	6-59
	Georgian: U+10A0—U+10FF	6-61
	Hangul Jamo: U+1100—U+11FF	6-62
	Latin Extended Additional: U+1E00—U+1EFF	6-63
	Greek Extended: U+1F00—U+1FFF	6-64
6.2	Symbols Area	6-67
	General Punctuation: U+2000—U+206F	6-68
	Superscripts and Subscripts: U+2070—U+209F	6-75
	Currency Symbols: U+20A0—U+20CF	6-76
	Combining Marks for Symbols: U+20D0—U+20FF	6-77
	Letterlike Symbols: U+2100—U+214F	6-78
	Number Forms: U+2150—U+218F	6-79
	Arrows: U+2190—U+21FF	6-80
	Mathematical Operators: U+2200—U+22FF	6-81

Miscellaneous Technical: U+2300—U+23FF	6-83
Control Pictures: U+2400—U+243F	6-84
Optical Character Recognition: U+2440—U+245F	6-85
Enclosed Alphanumerics: U+2460—U+24FF	6-86
Box Drawing: U+2500—U+257F	6-87
Block Elements: U+2580—U+259F	6-88
Geometric Shapes: U+25A0—U+25FF	6-89
Miscellaneous Symbols: U+2600—U+26FF	6-90
Dingbats: U+2700—U+27BF	6-91
6.3 CJK Phonetics and Symbols Area	6-93
CJK Symbols and Punctuation: U+3000—U+303F	6-94
Hiragana: U+3040—U+309F	6-95
Katakana: U+30A0—U+30FF	6-96
Bopomofo: U+3100—U+312F	6-97
Hangul Compatibility Jamo: U+3130—U+318F	6-98
Kanbun: U+3190—U+319F	6-99
Enclosed CJK Letters and Months: U+3200—U+32FF	6-100
CJK Compatibility: U+3300—U+33FF	6-101
6.4 CJK Ideographs Area	6-103
CJK Unified Ideographs: U+4E00—U+9FFF	6-104
6.5 Hangul Syllables Area	6-113
Hangul Syllables: U+AC00—U+D7A3	6-114
6.6 Surrogates Area	6-117
Surrogates Area: U+D800—U+DFFF	6-118
6.7 Private Use Area	6-119
Private Use Area: U+E000—U+F8FF	6-120
6.8 Compatibility Area and Specials	6-121
CJK Compatibility Ideographs: U+F900—U+FAFF	6-123
Alphabetic Presentation Forms: U+FB00—U+FB4F	6-124
Arabic Presentation Forms-A: U+FB50—U+FDFF	6-125
Combining Half Marks: U+FE20—U+FE2F	6-126
CJK Compatibility Forms: U+FE30—U+FE4F	6-127
Small Form Variants: U+FE50—U+FE6F	6-128
Arabic Presentation Forms-B: U+FE70—U+FEFF	6-129
Halfwidth and Fullwidth Forms: U+FF00—U+FFEF	6-130
Specials: U+FEFF, U+FFFO—U+FFFF	6-131
7 Code Charts	7-1
7.1 Character Names List	7-1
Images in the Code Charts and Character Lists	7-1
Cross References	7-2
Case Form Mappings	7-2
Decompositions	7-2
Information About Languages	7-3
Reserved Characters	7-3
7.2 CJK Unified Ideographs	7-3
7.3 Hangul Syllables	7-4
8 Han Radical-Stroke Index	8-1
A Transformation Formats	A-1
A.1 UTF-7	A-1

Rule 1: Direct Encoding	A-3
Rule 2: Unicode Shifted Encoding	A-3
Rule 3: ASCII Equivalents	A-3
Sample Implementation of the UTF-7 Conversions	A-3
A.2 UTF-8	A-7
Sample Implementation of the UTF-8 Conversions	A-8
B Submitting New Characters	B-1
B.1 Proposals	B-1
C Relationship to ISO/IEC 10646	C-1
C.1 Timeline	C-1
C.2 Structure of ISO/IEC 10646	C-2
C.3 UTF-16	C-3
C.4 The Unicode Standard and ISO/IEC 10646	C-3
C.5 The Unicode Standard as a Profile of 10646	C-4
C.6 Character Names	C-4
C.7 Character Functional Specifications	C-5
D Cumulative Changes	D-1
D.1 Versions of the Unicode Standard	D-1
D.2 Changes from Unicode 1.0 to 1.1	D-1
Areas Redefined	D-1
Characters Removed	D-1
Characters Unified	D-2
Characters Moved	D-3
Code Values Whose Assignment Has Changed	D-3
New Characters Added	D-4
Character Name Changes	D-6
Character Semantics Changes	D-7
D.3 Changes from Unicode 1.1 to Unicode 2.0	D-7
Areas Redefined	D-8
Characters Moved	D-8
New Characters Added	D-8
Character Name Changes	D-8
Character Semantics Changes	D-9
E Han Unification History	E-1
G Glossary	G-1
R References	R-1
R.1 Source Standards	R-1
R.2 Source Dictionaries for Han Unification	R-4
R.3 Selected Resources	R-4
I Indices	I-1
I.1 Unicode Names Index	I-1
I.2 General Index	I-25

Figures

Figure 1-1.	Wide ASCII	1-1
Figure 1-2.	Universal, Efficient, and Unambiguous.	1-3
Figure 2-1.	Characters Versus Glyphs	2-5
Figure 2-2.	Unicode Character Code to Rendered Glyph	2-6
Figure 2-3.	Bidirectional Ordering	2-7
Figure 2-4.	Equivalent Sequences	2-9
Figure 2-5.	Unicode Allocation	2-11
Figure 2-6.	Indic Vowel Signs	2-15
Figure 2-7.	Stacking Sequences	2-16
Figure 2-8.	Interacting Combining Characters.	2-16
Figure 2-9.	Overriding Behavior	2-17
Figure 2-10.	Multiple Base Characters	2-17
Figure 3-1.	Enclosing Marks.	3-9
Figure 3-2.	Positioning of Double Diacritics	3-9
Figure 5-1.	Ideographic Numbers	5-6
Figure 5-2.	Two-Stage Tables.	5-8
Figure 5-3.	Normalization	5-9
Figure 5-4.	Consistent Character Boundaries.	5-10
Figure 5-5.	Dead Keys Versus Handwriting Sequence.	5-13
Figure 5-6.	Truncating Composed Character Sequences	5-13
Figure 5-7.	Inside-Out Rule	5-14
Figure 5-9.	Bidirectional Placement	5-15
Figure 5-8.	Fallback Rendering	5-15
Figure 5-10.	Justification	5-16
Figure 5-11.	Positioning with Ligatures	5-16
Figure 5-12.	Positioning with Contextual Forms.	5-17
Figure 5-13.	Positioning with Enhanced Kerning	5-17
Figure 5-14.	Random Access	5-24
Figure 5-16.	Naïve Comparison.	5-29
Figure 5-15.	Character Equivalence.	5-29
Figure 5-17.	Levels of Comparison	5-30
Figure 5-18.	Orientation.	5-30
Figure 5-19.	Ignorable Characters	5-30
Figure 5-20.	Multiple Mappings	5-31
Figure 5-21.	Sublinear Searching	5-33
Figure 6-1.	General Scripts	6-2
Figure 6-2.	Diacritics on i	6-7
Figure 6-3.	Tone Letters	6-13
Figure 6-4.	Double Diacritics	6-14
Figure 6-5.	Ordering of Double Diacritics	6-14
Figure 6-6.	Reversal and Cursive Connection	6-22

Figure 6-7. Using Joiner 6-22

Figure 6-8. Using Non-Joiner 6-23

Figure 6-9. Combinations of Joiner and Non-Joiner. 6-23

Figure 6-10. Dependent versus Independent Vowels 6-35

Figure 6-11. Dead Consonants. 6-35

Figure 6-12. Conjunct Formations 6-36

Figure 6-13. Preventing Conjunct Forms 6-37

Figure 6-14. Half-Consonants 6-37

Figure 6-15. Independent Half-Forms 6-37

Figure 6-16. Consonant Forms 6-38

Figure 6-17. Rendering Order 6-42

Figure 6-18. Spacing Forms of Vowels 6-52

Figure 6-19. Vietnamese Letters and Tone Marks 6-63

Figure 6-20. Symbols. 6-67

Figure 6-21. CJK Misc. 6-93

Figure 6-22. CJK Ideographs 6-103

Figure 6-23. Han Spelling. 6-107

Figure 6-24. Context for Characters 6-107

Figure 6-25. Three-Dimensional Conceptual Model. 6-108

Figure 6-26. Preserving Variants 6-109

Figure 6-27. Not Cognates, Not Unified. 6-109

Figure 6-28. Component Structure 6-109

Figure 6-29. The Most Superior Node of a Component 6-109

Figure 6-30. Hangul Syllables. 6-113

Figure 6-31. Surrogates. 6-117

Figure 6-32. Private Use 6-119

Figure 6-33. Compatibility, Specials 6-121

Figure 6-34. Combining Half-Marks. 6-126

Tables

Table 2-1.	The Ten Unicode Design Principles	2-3
Table 3-1.	Sample Combining Classes	3-11
Table 3-2.	Canonical Ordering Results	3-11
Table 3-3.	Hangul Syllable Break Rules	3-12
Table 3-4.	Syllable Break Examples	3-12
Table 3-5.	Bidirectional Character Types	3-17
Table 3-6.	BIDI Example Abbreviations	3-17
Table 3-7.	Resolving Implicit Levels	3-21
Table 4-1.	Normative Character Properties	4-1
Table 4-2.	Informative Character Properties	4-1
Table 4-3.	Combining Classes	4-3
Table 4-4.	Bidirectional Character Types	4-11
Table 4-5.	Jamo Short Names	4-13
Table 4-6.	Numeric Properties	4-15
Table 4-7.	Mirrored Characters	4-22
Table 5-1.	Surrogate Support Levels	5-5
Table 5-2.	Surrogate Level Examples	5-5
Table 6-1.	Non-Spacing Marks Used with Greek	6-16
Table 6-2.	Digit Names	6-24
Table 6-3.	Arabic Joining Classes	6-24
Table 6-4.	Arabic Glyph Types	6-25
Table 6-5.	Ligature Notation	6-26
Table 6-6.	Dual-Joining Arabic Characters	6-27
Table 6-7.	Right-Joining Arabic Characters	6-29
Table 6-8.	Other Arabic Character Joining Classes	6-29
Table 6-9.	Indexed Arabic Character Joining Classes	6-29
Table 6-10.	Sample Half-Forms	6-42
Table 6-11.	Sample Ligatures	6-43
Table 6-12.	Sample Half-Ligature Forms	6-44
Table 6-13.	Tamil Letter Summary	6-49
Table 6-14.	Vowel Reordering	6-50
Table 6-15.	Vowel Splitting and Reordering	6-51
Table 6-16.	Ligating Vowel Signs	6-52
Table 6-17.	Greek Spacing and Non-Spacing Pairs	6-64
Table 6-18.	Unicode Space Characters	6-69
Table 6-19.	Unicode Dash Characters	6-69
Table 6-20.	Bidirectional Ordering Codes	6-72
Table 6-21.	Primary Source Standards for Unified Han	6-105
Table 6-22.	Secondary Source Standards for Unified Han	6-105
Table 6-23.	Common Han Characters	6-106
Table 6-24.	Ideographs Not Unified	6-110

Table 6-25. Ideographs Unified6-110
 Table 6-26. Han Ideograph Arrangement6-111
 Table 6-27. Line-Based Placement of Jungseong6-114
 Table 7-1. Han Character Chart Entries7-4
 Table A-1. UTF-7 Set D Special Characters A-2
 Table A-2. UTF-7 Set O A-2
 Table A-3. UTF-8 Bit Distribution A-7
 Table C-1. Zero Extending C-3
 Table D-1. Versions of the Unicode Standard D-1
 Table D-2. Characters Unified D-2
 Table D-3. Characters Moved D-3
 Table D-4. Changed Assignments D-3
 Table D-5. Reordered Character Groups D-3
 Table D-6. 1.1 Name Changes D-7
 Table D-7. 2.0 Name Changes D-8