Glossary

Abstract Character. A unit of information used for the organization, control, or representation of textual data. (See Definition D3 in *Section 3.3, Characters and Coded Representations*.)

Accent Mark. A mark placed above, below, or to the side of a character to alter its phonetic value. (See also *diacritic*.)

Alphabet. A collection of symbols that, in the context of a particular written language, represent the sounds of that language. The correspondence between symbols and sounds may be either more or less exact; most alphabets do not exhibit a one-to-one correspondence between distinct sounds (phonemes) and distinct symbols (graphemes).

Alphabetic Property. Informative property of the primary units of alphabets and/or syllabaries. (See *Section 4.10, Letters and Other Useful Properties*.)

Alphabetic Sorting. (See *collation*.)

ANSI. (1) The American National Standards Institute. (2) The Microsoft collective name for all Windows code pages. Sometimes used specifically for code page 1252, which is a superset of ISO/IEC 8859-1.

Arabic Digits. Forms of decimal digits used in most parts of the Arabic world (for instance, U+0660 ٠, U+0661 ١, U+0662 ٢, U+0663 ٣). Although *European digits* (1, 2, 3…) derive historically from these forms, they are visually distinct and are coded separately. (Arabic digits are sometimes called Indic numerals; however, this nomenclature leads to confusion with the digits currently used with the scripts of India.) Arabic digits are referred to as *Arabic-Indic digits* in the Unicode Standard. Variant forms of Arabic digits used chiefly in Iran and Pakistan are referred to as *Eastern Arabic-Indic digits*.

ASCII. Acronym for American Standard Code for Information Interchange, a 7-bit code that is the U.S. national variant of ISO/IEC 646. Formally, the U.S. standard ANSI X3.4.

Assigned Character. Synonym for *encoded character.*

Assigned Code Value. A code value that has defined, interoperable semantics.

Base Character. A character that does not graphically combine with preceding characters, and that is neither a control nor a format character. (See Definition D13 in *Section 3.5, Combination*.)

Basic Multilingual Plane. As defined by International Standard ISO/IEC 10646, code values 0000 through FFFF.

Bicameral. A script that has *case* distinctions. Most often used in the context of European alphabets.

BIDI. Abbreviation of bidirectional, in reference to mixed left-to-right and right-to-left text.

Bidirectional Display. The process or result of mixing left-to-right oriented text and right-to-left oriented text in a single line. (See *Section 3.12, Bidirectional Behavior*.)

Big-endian. A computer architecture that stores multiple-byte numerical values with the most significant byte (MSB) values first.

Binary Files. Files containing nontextual information.

Block. A grouping of related characters within the Unicode encoding space. A block may contain unassigned positions, which are reserved.

BMP. Abbreviation for *Basic Multilingual Plane*.

BNF. Abbreviation for *Backus-Naur Form*, a formal meta-syntax for describing content-free syntaxes. (For details, see *Section 0.2, Notational Conventions*.)

BOM. Acronym for *byte order mark*.

Bopomofo. An alphabetic script used primarily in the Republic of China (Taiwan) to write the sounds of Mandarin Chinese and some other dialects. Each symbol corresponds to either the syllable initial or syllable final sounds; it is therefore a subsyllabic script in its primary usage. The name is derived from the names of its first four elements. More properly known as *zhuyin zimu* or *zhuyin fuhao* (in Mandarin Chinese).

Boustrophedon. A pattern of writing seen in some ancient manuscripts and inscriptions, where alternate lines of text are laid out in opposite directions, and where right-to-left lines generally use glyphs mirrored from their left-to-right forms. Literally, "as the ox turns," referring to the plowing of a field.

Braille. A writing system using a series of raised dots to be read with the fingers by people who are blind or whose eyesight is not sufficient for reading printed material. (See *Section 12.9, Braille*.)

Braille Pattern. One of the 64 (for 6-dot Braille) or 256 (for 8-dot Braille) possible tangible dot combinations.

Byte Order Mark. The Unicode character U+FEFF ZERO WIDTH NO-BREAK SPACE when used to indicate the byte order of a text. (See *Section 2.7, Special Character and Noncharacter Values*, and *Section 13.6, Specials*.)

Byte Serialization. The order of a series of bytes determined by a computer architecture.

Byte-Swapped. Reversal of the order of a sequence of bytes.

Canonical. (1) Conforming to the general rules for encoding—that is, not compressed, compacted, or in any other form specified by a higher protocol. (2) Characteristic of a normative mapping and form of equivalence specified in *Chapter 3, Conformance*.

Canonical Decomposition. See Definition D23 in *Section 3.6, Decomposition*.

*Canonical Equivalen*t. Two character sequences are said to be canonical equivalents if their full canonical decompositions are identical. (See Definition D24 in *Section 3.6, Decomposition*.)

Cantillation Mark. A mark that is used to indicate how a text is to be chanted or sung.

Capital. Synonym for uppercase. (See *case*.)

Case. (1) Feature of certain alphabets where the letters have two distinct forms. These variants, which may differ markedly in shape and size, are called the *uppercase* letter (also known as *capital* or *majuscule*) and the *lowercase* letter (also known as *small* or *minuscule*). (2) Normative property of characters, consisting of uppercase, lowercase, and titlecase (Lu, Ll, and Lt). (See *Section 4.1, Case—Normative*.)

Case Mapping. The association of the uppercase, lowercase, and titlecase forms of a letter. (See *Section 5.18, Case Mappings*.)

Cedilla. A mark originally placed beneath the letter *c* in French, Portuguese, and Spanish to indicate that the letter is to be pronounced as an *s*, as in *façade*. Obsolete Spanish diminutive of *ceda*, the letter *z*.

Character. (1) The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape (see also *glyph*), though in code tables some form of visual representation is essential for the reader's understanding. (2) Synonym for *abstract character*. (See Definition D3 in *Section 3.3, Characters and Coded Representations*.) (3) The basic unit of encoding for the Unicode character encoding. (4) The English name for the ideographic written elements of Chinese origin. (See *ideograph* (2).)

Character Block. (See *block*.)

Character Class. A set of characters sharing a particular set of properties.

Character Encoding Form. Mapping from a character set definition to the actual bits used to represent the data.

Character Encoding Scheme. A *character encoding form* plus byte serialization. There are four character encoding schemes in Unicode: UTF-8, UTF-16, UTF-16BE, and UTF-16LE.

Character Properties. A set of property names and property values associated with individual characters. (See *Chapter 4, Character Properties*.)

Character Repertoire. The collection of characters included in a character set.

Character Sequence. See Definitions D4 (*abstract character sequence*) and D7 (*coded character sequence*) in *Section 3.3, Characters and Coded Representations*.

Character Set. A collection of elements used to represent textual information.

Chữ Hán. The name for Han characters used in Vietnam; derived from *Hanzi*.

Chữ Nôm. A demotic script of Vietnam developed from components of Han characters. Its creators used methods similar to those used by the Chinese in creating Han characters.

CJK. Abbreviation for Chinese, Japanese, and Korean. A variant, *CJKV*, means Chinese, Japanese, Korean, and Vietnamese.

Coded Character Representation. An ordered sequence of one or more code values that is associated with an abstract character in a given character repertoire. (See Definition D6 in *Section 3.3, Characters and Coded Representations*.)

Coded Character Sequence. An ordered sequence of coded character representations. (See Definition D7 in *Section 3.3, Characters and Coded Representations*.)

Coded Character Set. A character set in which each character is assigned a numeric code value. Frequently abbreviated as *character set*, *charset*, or *code set*.

Code Page. A coded character set, often referring to a coded character set used by a personal computer—for example, PC code page 437, the default coded character set used by the U.S. English version of the DOS operating system.

Code Point. (1) A numerical index (or position) in an encoding table used for encoding characters. (2) Synonym for *Unicode scalar value*.

Codespace. A range of numerical values available for encoding characters.

Code Value. The minimal bit combination that can represent a unit of encoded text for processing or interchange. (See Definition D5 in *Section 3.3, Characters and Coded Representations*.)

Collation. The process of ordering units of textual information. Collation is usually specific to a particular language. Also known as *alphabetizing* or *alphabetic sorting*. Unicode Technical Report #10, "Unicode Collation Algorithm," defines a complete, unambiguous, specified ordering for all characters in the Unicode Standard.

Combining Character. A character that graphically combines with a preceding base character. The combining character is said to *apply* to that base character. (See Definition D14 in *Section 3.5, Combination*.) (See also *nonspacing mark*.)

Combining Character Sequence. (See Definition D17 in *Section 3.5, Combination*.)

Combining Class. A numeric value given to each combining Unicode character that determines which other combining characters it typographically interacts with. (See Definition D37 in *Section 3.10, Canonical Ordering Behavior*.)

Compatibility. (1) Consistency with existing practice or preexisting character encoding standards. (2) Characteristic of a normative mapping and form of equivalence specified in *Section 3.6, Decomposition*.

Compatibility Character. (1) A character encoded only for compatibility with preexisting character encoding standards to support transcoding. (2) A character that has a compatibility decomposition. (See Definition D21 in *Section 3.6, Decomposition*.)

Compatibility Decomposition. (See Definition D20 in *Section 3.6, Decomposition*.)

Compatibility Equivalent. Two character sequences are said to be compatibility equivalents if their full compatibility decompositions are identical. (See Definition D22 in *Section 3.6, Decomposition*.)

Compatibility Variant. A character that generally can be remapped to another character without loss of information other than formatting.

Composed Character Sequence. (See Definition D17 in *Section 3.5, Combination*.)

Composite Character. (See *decomposable character*.)

Composite Character Sequence. (See *combining character sequence*.)

Conformance. Adherence to a specified set of criteria for use of a standard. (See *Chapter 3, Conformance*.)

Conjunct Form. A type of ligature that appears in most scripts based on the Brahmi family of Indic scripts. (See *Section 9.1, Devanagari*.)

Consonant Cluster. A sequence of characters that represents one or more consonants.

Consonant Conjunct. Typically, a ligated presentation form of a consonant cluster. This term is mostly applied to Brahmi-derived (Indic) scripts.

Contextual Variant. A text element can have a presentation form that depends upon textual context in which it is rendered. This presentation form is known as a *contextual variant*.

Control Codes. The 65 characters in the ranges U+0000..U+001F and U+007F..U+009F. Also known as *control characters*.

Cursive. Writing where the letters of a word are connected.

DBCS. Abbreviation for *double-byte character set*.

Dead Consonant. An Indic consonant character followed by a *virama* character. This sequence indicates that the consonant has lost its inherent vowel. (See *Section 9.1, Devanagari*.)

Decimal Digits. Digits that can be used to form decimal-radix numbers.

Decomposable Character. A character that is equivalent to a sequence of one or more other characters, according to the decomposition mappings found in the names list of *Section 14.1, Character Names List*. It may also be known as a *precomposed character* or a *composite character.* (See Definition D18 in *Section 3.6, Decomposition*.)

Decomposition. (1) The process of separating or analyzing a text element into component units. These component units may not have any functional status, but may be simply formal units—that is, abstract shapes. (2) (See Definition D19 in *Section 3.6, Decomposition*.)

Defective Combining Character Sequence. A combining character sequence that does not start with a base character. (See Definition D17a in *Section 3.5, Combination*.)

Demotic Script. (1) A script or a form of a script used to write the vernacular or common speech of some language community. (2) A simplified form of the ancient Egyptian hieratic writing.

Dependent Vowel. A symbol or sign that represents a vowel and that is attached or combined with another symbol, usually one that represents a consonant. For example, in writing systems based on Arabic, Hebrew, and Indic scripts, vowels are normally represented as dependent vowel signs.

Deprecated. A coded character whose use is strongly discouraged. Such characters are retained in the standard, but should not be used. (See Definition D7a in *Section 3.3, Characters and Coded Representations*.) (Not the same as *obsolete*.)

Diacritic. (1) A mark applied or attached to a symbol to create a new symbol that represents a modified or new value. (2) A mark applied to a symbol irrespective of whether it changes the value of that symbol. In the latter case, the diacritic usually represents an independent value (for example, an accent, tone, or some other linguistic information). Also called *diacritical mark* or *diacritical*. (See also *combining character* and *nonspacing mark*.)

Diaeresis. Two horizontal dots over a letter, as in *naïve*. The diaeresis is not distinguished from the *umlaut* in the Unicode character encoding. (See *umlaut.*)

Digits. (See *Arabic digits*, *European digits*, and *Indic digits.*)

Digraph. A pair of signs or symbols (two graphs), which together represent a single sound or a single linguistic unit. The English writing system employs many digraphs (for example, *th*, *ch*, *sh*, *qu*, and so on). The same two symbols may not always be interpreted as a digraph (for example, *cathode* versus *cathouse*). When three signs are so combined, they are called a *trigraph*. More than three are usually called an *n-graph*.

Dingbats. Typographical symbols and ornaments.

Diphthong. A pair of vowels that are considered a single vowel for the purpose of phonemic distinction. One of the two vowels is more prominent than the other. In writing systems, diphthongs are sometimes written with one symbol, and sometimes with more than one symbol (for example, with a *digraph*).

Directionality Property. A property of every graphic character that determines its horizontal ordering as specified in *Section 3.12, Bidirectional Behavior*. (See Definition D9 in *Section 3.4, Simple Properties.*)

Display Cell. A rectangular region on a display device within which one or more glyphs are imaged.

Display Order. The order of glyphs presented in text rendering.

Double-Byte Character Set. One of a number of character sets defined for representing Chinese, Japanese, or Korean text (for example, JIS X 0208-1990). These character sets are

often encoded in such a way as to allow double-byte character encodings to be mixed with single-byte character encodings. Abbreviated DBCS. (See also *multibyte character set*.)

Ductility. The ability of a cursive font to stretch or compress the connective baseline to effect text justification.

Dynamic Composition. Creation of composite forms such as accented letters or Hangul syllables from a sequence of characters.

EBCDIC. Acronym for Extended Binary-Coded Decimal Interchange Code. A group of coded character sets used on mainframes that consist of 8-bit coded characters. EBCDIC coded character sets reserve the first 64 code positions (x00 to x3F) for control codes, and reserve the range x41 to xFE for graphic characters. The English alphabetic characters are in discontinuous segments with uppercase at xC1 to xC9, xD1 to xD9, xE2 to xE9, and lowercase at x81 to x89, x91 to x99, xA2 to xA9.

Encapsulated Text. (1) Plain text surrounded by formatting information. (2) Text recoded to pass through narrow transmission channels or to match communication protocols.

Encoded Character. An *abstract character* together with its associated *Unicode scalar value*. By itself, an abstract character has no numerical value, but the process of "encoding a character" associates a particular Unicode scalar value with a particular abstract character, thereby resulting in an "encoded character."

Encoding Form. (See *character encoding form*.)

Encoding Scheme. (See *character encoding scheme*.)

Equivalence. In the context of text processing, the process or result of establishing whether two text elements are identical in some respect.

Equivalent Sequence. (See *canonical equivalent*.)

Escape Sequence. A sequence of bytes that is used for code extension. The first byte in the sequence is *escape* (hex 1B).

European Digits. Forms of decimal digits first used in Europe and now used worldwide. Historically, these digits were derived from the Arabic digits; they are sometimes called "Arabic numerals," but this nomenclature leads to confusion with the real *Arabic digits*.

Fancy Text. Also known as *rich text*. The result of adding additional information to plain text. Examples of information that can be added include font data, color, formatting information, phonetic annotations, interlinear text, and so on. The Unicode Standard does not address the representation of fancy text. It is expected that systems and applications will implement proprietary forms of fancy text. Some public forms of fancy text are available (for example, ODA, HTML, and SGML). When everything but primary content is removed from fancy text, only plain text should remain.

Floating (*diacritic, accent, mark*). (See *nonspacing mark*.)

Font. A collection of glyphs used for the visual depiction of character data. A font is often associated with a set of parameters (for example, size, posture, weight, and serifness), which, when set to particular values, generate a collection of imagable glyphs.

Formatted Text. (See *fancy text*.)

Formatting Codes. Characters that are inherently invisible but that have an effect on the surrounding characters.

FSS-UTF. Abbreviation for *File System Safe UCS Transformation Format*, published by the X/Open Company Ltd., and intended for the UNIX environment. Now known as *UTF-8*.

Fullwidth. Characters of East Asian character sets whose glyph image extends across the entire character display cell. In legacy character sets, fullwidth characters are normally encoded in two or three bytes. The Japanese term for fullwidth characters is *zenkaku*.

GCGID. Acronym for Graphic Character Global Identifier. These are listed in the IBM document *Character Data Representation Architecture, Level 1, Registry SC09-1391*.

General Category. Partition of the characters into major classes such as letters, punctuation, and symbols, and further subclasses for each of the major classes. (See *Section 4.5, General Category—Normative in Part.*)

Glyph. (1) An abstract form that represents one or more glyph images. (2) A synonym for *glyph image*. In displaying Unicode character data, one or more glyphs may be selected to depict a particular character. These glyphs are selected by a rendering engine during composition and layout processing. (See also *character.*)

Glyph Code. A code value that refers to a glyph. Usually, the glyphs contained in a font are referenced by their glyph code. Glyph codes may be local to a particular font; that is, a different font containing the same glyphs may use different codes.

Glyph Identifier. Similar to a glyph code, a glyph identifier is a label used to refer to a glyph within a font. A font may employ both local and global glyph identifiers. A collection of global or universal glyph identifiers is defined by the Association for Font Information and Interchange (AFII).

Glyph Image. The actual, concrete image of a glyph representation having been rasterized or otherwise imaged onto some display surface.

Glyph Metrics. A collection of properties that specify the relative size and positioning along with other features of a glyph.

Grapheme. (1) A minimally distinctive unit of writing in the context of a particular writing system. For example, ‹b› and ‹d› are distinct graphemes in English writing systems because there exist distinct words like big and dig. Conversely, ‹*a*› and ‹a› are not distinct graphemes because no word is distinguished on the basis of these two different forms. A grapheme is for a writing system what a phoneme is for a phonology. (2) What a user thinks of as a character.

Graphic Character. (1) A character typically associated with a visible display representation. (See also *glyph*.) (2) Any character that is not primarily associated with a control or formatting function.

Guillemet. Punctuation marks resembling small less-than and greater-than signs, used as quotation marks in French and other languages. (See “Language-Based Usage of Quotation Marks” in *Section 6.1, General Punctuation*.)

Halant. A synonym for the *virama* character. It literally means *killer*, referring to its function of *killing* the inherent vowel of a consonant letter. (See *virama*.)

Half-Consonant Form. In the Devanagari script, and certain other scripts of the Brahmi family of Indic scripts, a dead consonant may be depicted in the so-called half-form. This form is composed of the distinctive part of a consonant letter symbol without its vertical stem. It may be used to create conjunct forms that follow a horizontal layout pattern. Also known as *half-form.*

Halfwidth. Characters of East Asian character sets whose glyph image occupies half of the character display cell. In legacy character sets, halfwidth characters are normally encoded in a single byte. The Japanese term for halfwidth characters is *hankaku*.

Han Characters. Ideographic characters of Chinese origin. (See *Section 10.1, Han*.)

Hangul. The name of the script used to write the Korean language.

Hanja. The Korean name for Han characters; derived from the Chinese word *hanzi*.

Hankaku. (See *halfwidth*.)

Han Unification. The process of identifying Han characters that are in common among the writing systems of Chinese, Japanese, Korean, and Vietnamese.

Hanzi. The Mandarin Chinese name for Han characters.

Harakat. Marks that indicate vowels or other modifications of consonant letters in Arabic script.

Higher-Level Protocol. Any agreement on the interpretation of Unicode characters that extends beyond the scope of this standard. Such an agreement need not be formally announced in data; it may be implicit in the context. (See Definition D8 in *Section 3.3, Characters and Coded Representations*.)

High-Surrogate. A Unicode code value in the range U+D800 through U+DBFF. (See Definition D25 in *Section 3.7, Surrogates*.)

Hiragana. One of two standard syllabaries associated with the Japanese writing system. Hiragana syllables are typically used in representation of native Japanese words and grammatical particles.

HTML. HyperText Markup Language. A text description language related to SGML; it mixes text format markup with plain text content to describe formatted text. HTML is ubiquitous as the source language for Web pages on the Internet. Starting with HTML 4.0, the Unicode Standard functions as the reference character set for HTML content. (See also *SGML*.)

IANA. Internet Assigned Numbers Authority.

Ideograph. (1) Any symbol that primarily denotes an idea (or meaning) in contrast to a sound (or pronunciation)—for example, ☎ and 💣. (2) A common term used to refer to Han characters.

Ideographic Property. Informative property of characters that are ideographs. (See *Section 4.10, Letters and Other Useful Properties*.)

Illegal Code Value Sequence. (See Definition D31 in *Section 3.8, Transformations*.)

Ill-Formed Code Value Sequence. (See Definition D30 in *Section 3.8, Transformations*.)

Independent Vowel. In Indic scripts, certain vowels are depicted using independent letter symbols that stand on their own. This is often true when a word starts with a vowel or a word consists only of a vowel.

Indic Digits. Forms of decimal digits used in various Indic scripts (for example, Devanagari: U+0966 ०, U+0967 १, U+0968 २, U+0969 ३). Arabic digits (and, eventually, European digits) derive historically from these forms.

Informative. Information in this standard that is not normative but that contributes to the correct use and implementation of the standard.

Inherent Vowel. In writing systems based on a script in the Brahmi family of Indic scripts, a consonant letter symbol normally has an inherent vowel, unless otherwise indicated. The phonetic value of this vowel differs among the various languages written with these writing systems. An inherent vowel is overridden either by indicating another vowel with an explicit vowel sign or by using *virama* to create a dead consonant.

Inner Caps. Mixed case format where an uppercase letter is in a position other than first in the word—for example, "G" in the Name "McGowan."

IPA. (1) The International Phonetic Alphabet. (2) The International Phonetic Association, which defines and maintains the International Phonetic Alphabet.

IRG. Abbreviation for Ideographic Rapporteur Group, a subgroup of ISO/IEC JTC1/SC2/WG2. See *Appendix A, Han Unification History.*

Irregular Code Value Sequence. (See Definition D32 in *Section 3.8, Transformations*.)

ISCII. Acronym for Indian Standard Code for Information Interchange.

Jamo. The Korean name for a single letter of the Hangul script. Jamos are used to form Hangul syllables.

Joiner. An invisible character that affects the joining behavior of surrounding characters. (See *Section 8.2, Arabic*, and "Cursive Connection" in *Section 13.2, Layout Controls*.)

JTC1. The Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission responsible for information technology standardization.

Kana. The name of a primarily syllabic script used by the Japanese writing system. It comes in two forms, *hiragana* and *katakana*. The former is used to write particles, grammatical affixes, and words that have no *kanji* form; the latter is used primarily to write foreign words.

Kanji. The Japanese name for Han characters; derived from the Chinese word *hanzi*. Also romanized as *kanzi*.

Katakana. One of two standard syllabaries associated with the Japanese writing system. Katakana syllables are typically used in representation of borrowed vocabulary (other than that of Chinese origin), sound-symbolic interjections, or phonetic representation of "difficult" kanji characters in Japanese.

Kerning. (1) Changing the space between certain pairs of letters to improve the appearance of the text. (2) Process of mapping from pairs of glyphs to a positioning offset used to change the space between letters.

Letter. (1) An element of an alphabet. In a broad sense, includes elements of syllabaries and ideographs. (2) Informative property of characters that are used to write words.

Ligature. A glyph representing a combination of two or more characters. In the Latin script, there are only a few in modern use, such as the ligatures between "f" and "i" (= fi) or "f and l" (= fl). Other scripts make use of many ligatures, depending on the font and style.

Little-endian. A computer architecture that stores multiple-byte numerical values with the least significant byte (LSB) values first.

Logical Order. The order in which text is typed on a keyboard. For the most part, logical order corresponds to phonetic order. (See *Section 2.2, Unicode Design Principles.*)

Logical Store. Memory representation.

Lowercase. (See *case*.)

Low-Surrogate. A Unicode code value in the range U+DC00 through U+DFFF. (See Definition D26 in *Section 3.7, Surrogates*.)

LSB. Abbreviation for *least significant byte*.

LZW. Abbreviation for *Lempel-Ziv-Welch*, a standard algorithm widely used for compression of data.

Majuscule. Synonym for *uppercase*. (See *case*.)

Mathematical Property. Informative property of characters that are used as operators in mathematical formulae.

Matra. A dependent vowel in an Indic script. It is the name for vowel letters that follow consonant letters in logical order. A matra often has a completely different letter form from that for the same phonological vowel used as an independent letter.

MBCS. Abbreviation for *multibyte character set*.

MIME. Multipurpose Internet Mail Extensions. MIME is a standard that allows the embedding of arbitrary documents and other binary data of known types (images, sound, video, and so on) into e-mail handled by ordinary Internet electronic mail interchange protocols.

Minuscule. Synonym for *lowercase*. (See *case*.)

Mirrored Property. The property of characters whose images are mirrored horizontally in text that is laid out from right to left (versus left to right). (See Definition D10 in *Section 3.4, Simple Properties*.) (See also *Section 4.7, Mirrored—Normative*.)

Missing Glyph. (See *replacement glyph*.)

Modifier Letter. (1) Lm category in the Unicode Character Database. (2) Collection in the Modifier Letters block. Look like letters or punctuation and modify the pronunciation of other letters (similar to diacritics). (See *Section 7.8, Modifier Letters*.)

Monotonic. Modern Greek written with the basic accent, the *tonos*.

MSB. Abbreviation for *most significant byte*.

Multibyte Character Set. A character set encoded with a variable number of bytes per character. Many large character sets have been defined as MBCS so as to keep strict compatibility with the ASCII subset and/or ISO/IEC 2022. Abbreviated as MBCS.

Nekudot. Marks that indicate vowels or other modifications of consonantal letters in Hebrew.

Neutral Character. A character that can be written either right to left or left to right, depending on context. (See *Section 3.12, Bidirectional Behavior.)*

Non-joiner. An invisible character that affects the joining behavior of surrounding characters. (See *Section 8.2, Arabic*, and "Cursive Connection" in *Section 13.2, Layout Controls*.)

Nonspacing Diacritic. A diacritic that is a nonspacing mark.

Nonspacing Mark. A combining character whose positioning in presentation is dependent on its base character. It generally does not consume space along the visual baseline in and of itself. (See Definition D15 in *Section 3.5, Combination*.) (See also *combining character.*)

Normalization. Transformation of data to a normal form—for example, to unify spelling. (See *Section 5.7, Normalization*.)

Normative. Required for conformance with the Unicode Standard.

NSM. Abbreviation for *nonspacing mark*.

Numeric Value Property. A property of characters used to represent numbers. (See Definition D10b in *Section 3.4, Simple Properties.*)

Obsolete. Applies to a character no longer in current use, but which has been used historically. Whether a character is obsolete depends on context: for example, the Cyrillic letter *big yus* is obsolete for Russian, but is used in modern Bulgarian. (Not the same as *deprecated*.)

Phoneme. A minimally distinct sound in the context of a particular spoken language. For example, in American English, /p/ and /b/ are distinct phonemes because pat and bat are distinct; however, the two different sounds of /t/ in tick and stick are not distinct in English, even though they are distinct in other languages such as Thai.

Pinyin. Standard system for the romanization of Chinese on the basis of Mandarin pronunciation.

Pivot Conversion. The use of a third character encoding to serve as an intermediate step in the conversion between two other character encodings. The Unicode Standard is widely used to support pivot conversion, as its character repertoire is a superset of most other coded character sets.

Plain Text. Computer-encoded text that consists *only* of a sequence of code values from a given standard, with no other formatting or structural information. Plain text interchange is commonly used between computer systems that do not share higher-level protocols. (See also *fancy text*.)

Points. (1) The nonspacing vowels and other signs of written Hebrew. (2) A unit of measurement in typography.

Polytonic. Ancient Greek written with several contrastive accents.

Precomposed Character. (See *decomposable character.*)

Presentation Form. A ligature or variant glyph that has been encoded as a character for compatibility. (See also *compatibility character* (1).)

Private Use. Unicode values from U+E000 to U+F8FF and surrogate pairs (see *Section 3.7, Surrogates)* whose high-surrogate is from U+DB80 to U+DBFF are available for private use. (See Definition D12 in *Section 3.4, Simple Properties*.) Refers to code values and areas of the standard whose interpretation is not specified by the standard and whose use may be determined by private agreement among cooperating users.

Property. (See *character properties*.)

Radical. A structural component of a Han character conventionally used for indexing. The traditional number of such radicals is 214.

Rendering. (1) The process of selecting and laying out glyphs for the purpose of depicting characters. (2) The process of making glyphs visible on a display device.

Repertoire. (See *character repertoire*.)

Replacement Character. Character used as a substitute for an uninterpretable character from another encoding. The Unicode Standard uses U+FFFD REPLACEMENT CHARACTER for this function.

Replacement Glyph. A glyph used to render a character that cannot be rendered with the correct appearance in a particular font. It often is shown as an open □ or black ■ rectangle. Also known as a *missing glyph*. (See *Section 5.3, Unknown and Missing Characters*.)

Reserved. Unassigned code values that are set aside for future standardization by the Unicode Consortium.

Rich Text. (See *fancy text*.)

SBCS. Acronym for *single-byte character set*. Any 1-byte character encoding. This term is generally used in contrast with DBCS and/or MBCS.

Scalar Value. (See *Unicode scalar value*.)

Script. A collection of symbols used to represent textual information in one or more writing systems.

SGML. Structured Graphic Markup Language. A standard framework for defining particular text markup languages. The SGML framework allows for mixing structural tags that describe format with the plain text content of documents, so that fancy text can be fully described in a plain text stream of data. (See also *HTML*, *XML*, and *fancy text*.)

Shaping Characters. Characters that assume different glyphic forms depending on the context.

Small Letter. Synonym for *lowercase.* (See *case*.)

Sorting. (See *collation.)*

Spacing Mark. A combining character that is not a nonspacing mark. (See *nonspacing mark*.)

Static Form. (See *decomposable character.*)

Surrogate Pair. A coded character representation for a single abstract character that consists of a sequence of two Unicode values, where the first value of the pair is a *high-surrogate* and the second is a *low-surrogate*. (See Definition D27 in *Section 3.7, Surrogates*.)

Syllabary. An alphabet whose symbols typically represent multiple phonemes of a language. These multiple phonemes are generally combinations of consonants and vowels.

Syllable. (1) An element of a syllabary. (2) A basic unit of articulation that corresponds to a pulmonary pulse.

Symmetric Swapping. (See *mirrored*.)

TEX. Computer language designed for use in typesetting, in particular for typesetting math and other technical material. (According to Knuth, TEX rhymes with the word *blecchhh*.)

Text Element. A minimum unit of text in relation to a particular text process, in the context of a given writing system. In general, the mapping between text elements and code values is many-to-many. (See *Chapter 2, General Structure*.)

Titlecase. Uppercased initial letter followed by lowercase letters in words. A casing convention often used in titles, headers, and entries, as exemplified in this glossary.

Tone Mark. A diacritic or nonspacing mark that represents a phonemic tone. Tone languages are common in Southeast Asia and Africa. Because tones always accompany vowels (the syllabic nucleus), they are most frequently written using functionally independent marks attached to a vowel symbol. However, some writing systems such as Thai place tone marks on consonant symbols; Chinese does not use tone marks (except when it is written phonemically).

Transcoding. Conversion of character data between different character sets.

Transformation Format. A mapping from a coded character sequence to a unique sequence of code values (typically bytes).

Triangulation. (See *pivot conversion*.)

UCS. Abbreviation for Universal Character Set, which is specified by International Standard ISO/IEC 10646.

UCS-2. ISO/IEC 10646 encoding form: Universal Character Set coded in 2 octets. (See *Appendix C, Relationship to ISO/IEC 10646.)*

UCS-4. ISO/IEC 10646 encoding form: Universal Character Set coded in 4 octets. (See *Appendix C, Relationship to ISO/IEC 10646.)*

Umlaut. Two horizontal dots over a letter, as in German *Köpfe*. The umlaut is not distinguished from the *diaeresis* in the Unicode character encoding. (See *diaeresis*.)

Unassigned. Code values that either are reserved for future use or are never to be used.

Unicameral. A script that has no *case* distinctions. Most often used in the context of European alphabets.

Unicode Character Database. A collection of files providing normative and informative Unicode character properties and mappings. (See *Chapter 4, Character Properties*, and the UnicodeCharacterDatabase.html on the CD-ROM.)

Unicode Scalar Value. A number N from 0 to $10FFFF_{16}$ defined by application of the algorithm in Definition D28. (See *Section 3.7, Surrogates.*)

Unicode Signature. An implicit marker to identify a file as containing Unicode text in a particular encoding form. An initial *byte order mark* (BOM) may be used as a Unicode signature.

Unicode (or *UCS*) *Transformation Format.* (See Definition D29 in *Section 3.8, Transformations*, see also *Section C.3, UCS Transformation Formats.*)

Unification. The process of identifying characters that are in common among writing systems.

Uppercase. (See *case*.)

URO. Abbreviation for Unified Repertoire and Ordering, the original set of CJK unified ideographs used in the Unicode Standard.

UTF. Abbreviation for *Unicode* (or *UCS*) *Transformation Format*.

UTF-2. Obsolete name for *UTF-8*.

UTF-7. Unicode (or UCS) Transformation Format, 7-bit encoding form, specified by *RFC-2152*.

UTF-8. Unicode (or UCS) Transformation Format, 8-bit encoding form. UTF-8 is the Unicode Transformation Format that serializes a Unicode scalar value as a sequence of one to four bytes, as specified in *Table 3-1, UTF-8 Bit Distribution*. (See Definition D36 in *Section 3.8, Transformations.)*

UTF-16. Unicode (or UCS) Transformation Format, 16-bit encoding form. The UTF-16 is the Unicode Transformation Format that serializes a Unicode value as a sequence of two bytes, in either big-endian or little-endian format. (See Definition D35 in *Section 3.8, Transformations.*)

UTF-16BE. The Unicode Transformation Format that serializes a Unicode value as a sequence of two bytes, in big-endian format. An initial sequence corresponding to U+FEFF is interpreted as a ZERO WIDTH NO-BREAK SPACE. (See Definition D33 in *Section 3.8, Transformations.*)

UTF-16LE. The Unicode Transformation Format that serializes a Unicode value as a sequence of two bytes, in little-endian format. An initial sequence corresponding to U+FEFF is interpreted as a ZERO WIDTH NO-BREAK SPACE. (See Definition D34 in *Section 3.8, Transformations.*)

Virama. The name of a symbol used with Indic scripts to indicate a dead consonant. (See *Section 9.1, Devanagari*, and *Section 9.6, Tamil.*)

Visual Order. Characters ordered as they are presented for reading. (Contrast with *logical order.*)

Vocalization. Marks placed above, below, or within consonants to indicate vowels or other aspects of pronunciation. A feature of Middle Eastern scripts.

Vowel Mark. In many scripts, a mark used to indicate a vowel or vowel quality.

wchar_t. The ANSI C defined *wide character* type, usually implemented as either 16 or 32 bits. ANSI specifies that `wchar_t` be an integral type and that the C language source character set be mappable by simple extension (zero- or sign-extension).

Writing Direction. The direction or orientation of writing characters within lines of text in a writing system. Three directions are common in modern writing systems: left to right, right to left, and top to bottom.

Writing System. A set of rules for using one or more scripts to write a particular language. Examples include the American English writing system, the British English writing system, the French writing system, and the Japanese writing system.

XML. eXtensible Markup Language. A subset of SGML constituting a particular text markup language for interchange of structured data. The Unicode Standard is the reference character set for XML content. (See also *SGML* and *fancy text.*) XML is a trademark of the World Wide Web Consortium.

Zenkaku. (See *fullwidth*).

Zero Width. Characteristic of some spaces or format control characters that do not advance text along the horizontal baseline. (See *nonspacing mark.*)

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, 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 (see http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html). More recent versions of the Unicode standard exist (see http://www.unicode.org/unicode/standard/versions/).

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

ISBN 0-201-61633-5

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: http://www.awl.com/cseng/

First printing, January 2000.