Glossary

Abstract Character. A unit of information used for the organization, control, or representation of textual data. (See also *character* (1, 2) and *surrogate pair.*)

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 which, 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).

ANSI. (1) The American National Standards Institute. (2) The Microsoft collective name for all the 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 derive historically from these forms, they are visually distinct and are coded separately. (Arabic digits are sometimes called Indic numerals; however, this 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.

Area. An organizational unit of the Unicode Standard larger than the *block*. The areas are enumerated in *Chapter 2, General Structure,* and are further described in *Chapter 6, Character Block Descriptions.*

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

Base Character. A character that does not graphically combine with preceding characters.

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.

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

Binary Files. Files containing non-textual information.

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

BOM. Acronym for *byte order mark.* The Unicode character U+FEFF ZERO WIDTH NO-BREAK SPACE is used to indicate the byte order of a text. The BOM allows a receiver of Unicode text to distinguish between text arriving in big-endian order from text arriving in little-endian order *in the absence of a higher level protocol.* (See *Chapter 5, Implementation Guidelines* and the Specials subsection in *Section 6.8, Compatibility Area and Specials.*)

Bopomofo. An alphabetic script used primarily in the Republic of China (Taiwan) to write the sounds of Mandarin Chinese. Each symbol corresponds to either the syllable initial or

syllable final sounds; it is therefore a sub-syllabic 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).

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 the Conformance chapter of this standard.

Canonical Decomposition. The decomposition of a character which results from applying canonical mappings and then reordering non-spacing marks according to the Canonical Ordering Algorithm. (See *Chapter 3, Conformance.*)

Canonical Equivalent. Two character sequences are said to be canonical equivalents if their canonical decompositions are identical. (See *Chapter 3, Conformance.*)

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

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.* (3) Loosely, the basic unit of encoding for the Unicode character encoding, a 16-bit unit of textual information. (4) Synonym for *code value.* (5) The English name for the ideographic written elements of Chinese origin. (See *ideograph (2).*)

Character Properties. An unordered list of property names and property values associated with individual character code elements. (See *Chapter 4, Character Properties.*)

Character Repertoire. (See *repertoire.*)

CJK. Abbreviation for Chinese, Japanese, and Korean.

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.

Code Element. (See *code point.*)

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 *code value.*

Code Space. A range of numerical values available for encoding characters.

Code Value. A minimal bit combination that can represent a unit of encoded text for processing or interchange. Also known as *code point.*

Coded Character Representation. An ordered sequence of one or more code values associated with an abstract character in a given character repertoire. (See *Chapter 3, Conformance.*)

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.*

Collation. The process of ordering units of textual information. Collation is usually specific to a particular language. Also known as *alphabetizing* or *alphabetic sorting.*

Combining Character. A character that graphically combines with a preceding base character. (See also *non-spacing mark.*)

Combining Character Sequence. A character sequence consisting of a base character followed by one or more combining characters. Also called *composed character sequence.*

Compatibility. (1) Consistency with existing practice or pre-existing character encoding standards. (2) Characteristic of a normative mapping and form of equivalence specified in the Conformance chapter of this standard.

Compatibility Area. An area of the Unicode Standard, which contains characters included only for compatibility with pre-existing character encoding standards.

Compatibility Character. (1) A character encoded only for compatibility. (2) A character which has a compatibility decomposition.

Compatibility Decomposition. The decomposition of a character which results from applying the compatibility mappings and then reordering non-spacing marks according to the Canonical Ordering Algorithm. (See *Chapter 3, Conformance.*)

Compatibility Equivalent. Two character sequences are said to be compatibility equivalents if their compatibility decompositions are identical.

Composite Character. (See *decomposable character.*)

Conjunct Form. A type of ligature that appears in most scripts based on the Brahmi family of Indic scripts. (See the Devanagari character block description.)

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.*

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 the Devanagari character block description.)

Decomposable Character. A character that is equivalent to a sequence of one or more other characters. Also known as a *precomposed character* or a *composite character.*

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) A sequence of one or more characters that is equivalent to a decomposable character. (See *decomposable character* and *Chapter 3, Conformance.*)

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. (See cover.)

Dependent Vowel. A symbol or sign that represents a vowel and which 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.

Diacritic. (1) A mark applied or attached to a symbol in order to create a new symbol that represents an 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 *non-spacing mark.*)

Diaeresis. Two horizontal dots over a letter, as in *naïve.* The same Unicode character is used to represent the *umlaut.* (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.*

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. A property of every graphic character that determines its horizontal ordering as specified in the Unicode Bidirectional Algorithm. (See *Chapter 3, Conformance.*)

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

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 multi-byte character set.)

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

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

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

European Digits. Forms of decimal digits first used in Europe and now used worldwide. Historically, these derive from the Arabic digits; they are sometimes called Arabic numerals, but this 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 *non-spacing 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 which have an effect on the surrounding characters. An example is U+206E NATIONAL DIGIT SHAPES.

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

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. 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 since there exist distinct words like big and dig. Conversely, ‹*a*› and ‹a› are not distinct graphemes since 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.

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.

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-form Consonant. 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.

Han Character. Ideographic characters of Chinese origin.

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

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

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

Hankaku. Japanese for *halfwidth*; refers to glyph images designed to fit half the display space of a Han character.

Hanzi. The Chinese name for Han characters. The Han script was first codified during the Han dynasty (in the 3rd century BC). The name hanzi is derived from two Han characters, signifying Han (that is, "Chinese") and character respectively.

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.

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.

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 which 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.

ISCII. (1) Indian Standard Code for Information Interchange. (2) Iranian Standard Code for Information Interchange

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

Joiner. An invisible character that affects the joining behavior of surrounding characters. (See the General Punctuation character block description.)

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 which have no *kanji* form; the latter is used primarily to write foreign words.

Kanji. The name for Han characters used in Japanese; 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.

Letter. An element of an alphabet. In a broad sense, includes elements of syllabaries and ideographs.

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. (For more information, see *Chapter 2, General Structure.*)

LSB. Abbreviation for *least significant byte.*

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

Mirrored. A property of characters whose images are mirrored horizontally in text that is laid out from right to left (versus left to right). (See *Chapter 3, Conformance.*)

MSB. Abbreviation for *most significant byte.*

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

Neutral character. A character that can be written either right-to-left or left-to-right, depending on context.

Non-Joiner. An invisible character that affects the joining behavior of surrounding characters. (See the General Punctuation character block description.)

Non-spacing Diacritic. A diacritic that is a non-spacing mark.

Non-spacing Mark. A combining character whose positioning in presentation is dependent on its base character. A non-spacing mark generally does not consume space along the visual baseline in and of itself. (See also *combining character.*)

Normative. Required for conformance with the Unicode Standard.

NSM. Abbreviation for *non-spacing mark.*

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 in Thai.

Plain Text. Computer encoded text that consists *only* of a sequence of code elements 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 non-spacing vowels and other signs of written Hebrew. (2) A unit of measurement in typography.

Precomposed Character. (See *decomposable character.*)

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

Private Use. 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. The collection of characters included in a character set.

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 5.4, Unknown and Missing Characters.*)

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.

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

Spacing Mark. A combining character that is not a non-spacing mark. (See *non-spacing mark.*)

Surrogate, High. A Unicode code value in the range U+D800 through U+DBFF.

Surrogate, Low. A Unicode code value in the range U+DC00 through U+DFFF.

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.

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.*)

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

Titlecase. Uppercased initial letters 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 non-spacing mark that represents a phonemic tone. Tone languages are common in Southeast Asia and Africa. Since 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).

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 same Unicode character is used to represent the *diaeresis.* (See *diaeresis.*)

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

UTF-7. Unicode (or UCS) Transformation Format, 7-bit form. (See *Appendix A, Transformation Formats.*)

UTF-8. Unicode (or UCS) Transformation Format, 8-bit form. (See *Appendix A, Transformation Formats.*)

UTF-16. The ISO/IEC 10646 encoding that is equivalent to the Unicode Standard with the use of surrogates as described in *Section 3.7, Surrogates.* (See also *Appendix C, Relationship to ISO/IEC 10646.)*

Virama. The name of a symbol used with Indic scripts to indicate a dead consonant. (See the Devanagari and Tamil character block descriptions.)

Vowel Sign. 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.

Written Language. A non-oral form of language. A writing system is the means by which a language is written.

Zenkaku. Japanese for *fullwidth*; refers to glyph images designed to fit the same display space as a Han character.

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