Appendix F: Glossary

Abstract character: A character as a semantic entity as opposed to a particular representation or shape of the character. *See also* character, 1.

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

ANSI. (1) The American National Standards Institute. (2) the Microsoft Windows ANSI character set, essentially ISO 8859-1 plus two characters, so named because it was originally based on an ANSI draft standard.

Arabic digits. Forms of decimal digits used in most parts of the Arabic world (for instance, U+0660, U+0661, U+0662, U+0663 ٠١٢٣). Although European digits derive historically from these forms, they are visually distinct and 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.

ASCII. Acronym for "American Standard Code for Information Interchange," a 7-bit code that is the US national variant of ISO 646. Formally, the U.S. standard ANSI X3.4-1986.

AX. Refers to a character encoding specified in the AX Technical Reference Guide, published by the AX Committee, based in Tokyo.

Backing store. Character storage in memory or on disk, as opposed to characters displayed or printed.

Base character. (1) A character which is neither non-spacing nor characterized by obligatory overlapping with adjacent characters, nor by dependence upon other characters—that is, one that stands on its own. (2) Any graphic character which is not a non-spacing character. Base characters are also known as "spacing" characters, because they generally have non-zero width (in horizontal contexts).

BIDI. Short for "bidirectional," in reference to mixed left-to-right and right-to-left text.

Big-endian. Referring to a computer architecture which stores multiple-byte numerical values with the most significant byte (MSB) values first.

Binary files. Files containing non-textual information.

Block. A convenient unit for grouping characters within the Unicode encoding space, based on a multiple of 16. A typical block may contain unencoded positions, which are reserved.

BOM. Acronym for "Byte Order Mark." The Unicode character (U+FEFF) 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* Appendix B, Implementation Guidelines.

Canonical. Conforming to the general rules for encoding, that is, not compressed, compacted or in any other form specified by a higher protocol, and also not consisting of a Compatibility Zone encoding.

Character. (1) The smallest component of written language that has semantic value. Character refers to the abstract idea, 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) The basic unit of encoding for the Unicode character encoding, 16 bits of information. (3) Synonym for "code element." (4) The English name for the ideographic written elements of Chinese origin.

Character encoding. Association of a unique number with each character in a set of characters. The distinction between characters and glyphs is not absolutely clear in all cases, and so most large character encodings also encode some set of glyphs.

CJK. Acronym for "Chinese, Japanese, and Korean."

Code element. The minimal bit combination that can represent a unit of encoded text for processing or exchange.

Code page. An ordered character set with a numerical index (*code point*) associated with each character.

Code point. A numerical index (or position) in an encoding table used for encoding characters.

Code set. A character encoding; this term is widely used by programmers.

Codespace. The range of numerical values available for encoding characters, given the set of principles for how numerical values are to be used for encoding.

Compatibility Zone. A section of the Unicode codespace, included for compatibility with pre-existing character encoding standards. The Compatibility Zone contains variant characters that can be mapped to Unicode canonical equivalents.

Composed character sequence. A sequence of characters consisting of a base character followed by one or more non-spacing marks.

Conjunct consonant. (1) The juxtaposition of two or more consonants in Indic scripts. "Conjunct" refers to the pronounciation of two consonants with no intervening vowel. (2) The term is also used to refer to the specific typographical forms themselves, where the consonants may be written together in a single typographical form (ligature) or may be written with a *virama* below one or more of the consonants.

DBCS. Double Byte Character Set. Any 2-byte form of Multi-Byte encoding. (*See* MBCS)

Decomposition. (1) The act of mapping a precomposed character into the corresponding composed character sequence. (2) The resulting composed character sequence.

Diacritic. Any character (spacing or non-spacing) used with the Latin script or related scripts such as IPA, typically indicating that a phonetic value is different from the unmarked state. This broad definition is intended to include accents. See also *non-spacing diacritic.*

Diaeresis. (plural *-eses*) 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, Indic digits.*

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 re-coded to pass through narrow transmission channels or to match communication protocols.

European digits. Forms of decimal digits used in Europe (for instance, 0, 1, 2, 3). Historically, these derive from the Arabic digits, but are visually distinct and coded separately. Many countries outside of Europe have also adopted these forms for decimal digits. (Although European digits are sometimes called "Arabic numerals," this leads to confusion with the real Arabic digits.)

Fancy text. A data or interchange format containing both plain text and additional formatting information such as font, style, margins, etc.

Floating (diacritic, accent, mark). See non-spacing mark.

Formatted text. See fancy text.

Formatting codes. Non-spacing characters that are inherently invisible but which have an effect on the surrounding characters. An example is the ZERO WIDTH NON-JOINER.

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) The actual shape (bit pattern, outline) of a character image. For example, an italic "*a*"and a roman "a" are two different glyphs representing the same underlying character. In this strict sense, any two images which differ in shape constitute different glyphs. In this usage, "glyph" is a synonym for "character image" or simply "image." (2) A kind of idealized surface form derived from some combination of underlying characters in some specific context, rather than an actual character image. In this broad usage, two images would constitute the same glyph whenever they have essentially the same topology (as in oblique "*a*" and roman "a"), but *different* glyphs when one is written with a hooked top and the other without (as in italic *a* and roman a). In this usage "glyph" is a synonym for "glyph type," where glyph is defined as in sense 1. (*See also* Text Processes, Section 2.1.)

Han. Generic adjective referring to ideographic characters of Chinese origin.

Han unification. Assignment of the same code point to characters from more than one of the East Asian ideographic character standards in which they may (or may not) be represented by slightly different glyphs (in either of the two senses of "glyph" defined previously). This process actually reunifies elements which are historically used as, or have been historically perceived as being, the same "character." *See also the block description for Han characters.*

Hangul. The Korean syllabic writing system.

Hanja. The Korean name for the ideographic characters of Chinese origin.

Hanzi (~Han tsu). The Chinese name for the ideographic characters of Chinese origin.

Ideographic character. A character that generally stands for a word or a morpheme, rather than a sound. In this context the term is applied to characters of the Chinese script used for writing various languages. The Chinese script can be more accurately described as "logographic," but the Unicode standard uses the more widely-known terminology. The term "ideographic" is often also applied to various hieroglyphic writing systems.

Indic digits. Forms of decimal digits used in various Indic scripts (for instance, Devanagari: U+0966, U+0967, U+0968, U+0969 ०१२३). Although Arabic digits (and, eventually, European digits) derive historically from these forms, they are visually distinct and coded separately.

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

Kanji. The Japanese name for the ideographic characters of Chinese origin.

Keyboard language. Synonym for "keyboard layout," when it refers to accepted national arrangements of characters.

Keyboard layout. The designation of which keys produce which characters (or scan-codes).

Letter. (1) Basic element of a script as understood by the end user. (2) A higher level of abstraction than "character." *See* text element.

Ligature. Two (or more) characters combined into a single typographical form. 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. Some languages have mandatory ligatures; other languages have text elements that historically were derived from ligatures, but are now characters. Examples of the latter are the German *eszet* (the ligature of long and short “s”) and the ampersand (&) which originated as a contracted form of the Latin word “et” (which derivation can still be discerned in many fonts).

Little-endian. Referring to a computer architecture which stores multiple-byte numerical values with the least significant byte (LSB) values first.

Locale. The national and/or cultural environment in which a system or program is running. Specifically, a software implementation of this environment. The locale determines sort order, language of messages, keyboard layout, date and time formatting conventions, etc. It is sometimes, but not necessarily, coincident with a language or country boundary. For example, French Swiss and German Swiss constitute different locales.

Localization. The process of adapting a program for a different international market. A program which requires only translation of the program’s messages is said to be *locale independent;* the operating system provides the parameters for the program to reconfigure itself. The usual examples are different date and time formatting.

Logographic. See Ideographic Character. Chinese characters are more properly termed logographic than ideographic.

LSB. Acronym for *least significant byte.*

MBCS. Acronym for *multi byte character set.* This implies more than one byte per character and often, that the number of bytes may be different for different characters in the character set. Many large character sets have been defined as MBCS in order to keep strict compatibility with the ASCII subset and/or ISO 2022.

MSB. Acronym for *most significant byte.*

Multibyte. The term used for character encodings that employ more than one byte per character; some such encodings may have a variable number of bytes per character. These encodings typically follow the ISO 2022 model, avoiding use of byte values in the so-called control ranges to encode graphic characters. The essential factor which distinguishes the Unicode character encoding from a typical multibyte code set is that the Unicode encoding is *defined* not as a 2-byte standard, but as a 16-bit standard.

Multilingual. Referring to many languages. A multilingual program strives to handle data in a way that is not dependent on a particular language or writing system. Multilingual documents combine text which is written in different languages. Multilingual may refer to many languages which all use the same script (such as English, French, and German), or to many languages which use distinct scripts (such as German, Hebrew, and Korean). The latter case is also referred to as *multiscript.*

National convention. A convention which is specific to a particular nation or district, and which may refer to everything from the currency symbol to the names of the weekdays in any language or local dialect. Somewhat synonymous with "locale," but the emphasis is not on the computer representation.

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

Non-spacing character. A character with character width (advance width) equal to zero. There are two types of non-spacing characters: formatting codes and non-spacing marks. *See* formatting codes and non-spacing marks.

Non-spacing diacritic. A diacritic which is a non-spacing mark.

Non-spacing mark. A non-spacing graphic character which is positioned with reference to a preceding base character.

Plain text. Computer encoded text which consists *only* of a sequence of code elements from a given standard, with no other formatting or structural information to indicate a specific interpretation. Plain text interchange is commonly used between computer systems which may have no other factors in common.

Points. The non-spacing vowels and other signs of written Hebrew.

Precomposed character. A single Unicode character which represents a composed character sequence, usually a combination of one or more diacritic marks with a base character.

Radical. Parts of Chinese characters used collectively as a method of indexing them. A character may contain more than one element which is recognized as being a Radical, but each contains only *one* element that is actually *used* as the indexing Radical for the character. Many radicals can exist as stand-alone Characters.

Rendering. The text process of displaying or printing a sequence of characters in a visible form.

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

Replacement glyph. (Synonym: missing glyph) A glyph used to render a character which cannot be rendered with the correct appearance in a particular font. It often is shown as an open box □ or as a black rectangle ▮.

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

Text Element. A minimal unit of text relative to some specific language and some specific process performed on the text.

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 folding together of elements that need not be distinguished, or the recognition that two things which are distinguished in some places should not, in general, be distinguished as encoded characters because their use is disjoint enough to be apparent from context. Two ends of the spectrum are "radical" unification, which might fold together the upper and lowercase forms of Greek A, Cyrillic A and Latin A into a single code point; and the more conservative unification of the Unicode standard, which maintains boundaries between scripts in the broadest sense and, for compatibility, preserves many distinctions made in other computer character encodings.

Vowel sign. In many scripts, a mark used to indicate a vowel or vowel quality. Typical vowel signs are either partially overlapping or non-spacing marks, but those are not necessary criteria. Vowel signs are characterized generally by *functional* dependence upon an adjacent or nearby consonant, the inherent vowel of which they typically modify.

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). A frequent assumption is that the source and target code sets are different, and that the size of the "char" data type is insufficiently "wide" to hold a character of the target code set.

Zero width. *See* non-spacing.