Chapter 1

DRAFT

Introduction

The Unicode Standard is the universal character encoding scheme for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software. As the default encoding of HTML and XML, the Unicode Standard provides a sound underpinning for the World Wide Web and new methods of business in a networked world. Required in new Internet protocols and implemented in all modern operating systems and computer languages such as Java and C#, Unicode is the basis of software that must function all around the world.

With Unicode, the information technology industry has gained data stability instead of proliferating character sets; greater global interoperability and data interchange; and simplified software and reduced development costs.

While modeled on the ASCII character set, the Unicode Standard goes far beyond ASCII's limited ability to encode only the upper- and lowercase letters A through Z. It provides the capacity to encode all characters used for the written languages of the world—more than 1 million characters can be encoded. No escape sequence or control code is required to specify any character in any language. The Unicode character encoding treats alphabetic characters, ideographic characters, and symbols equivalently, which means they can be used in any mixture and with equal facility (see *Figure 1-1*).

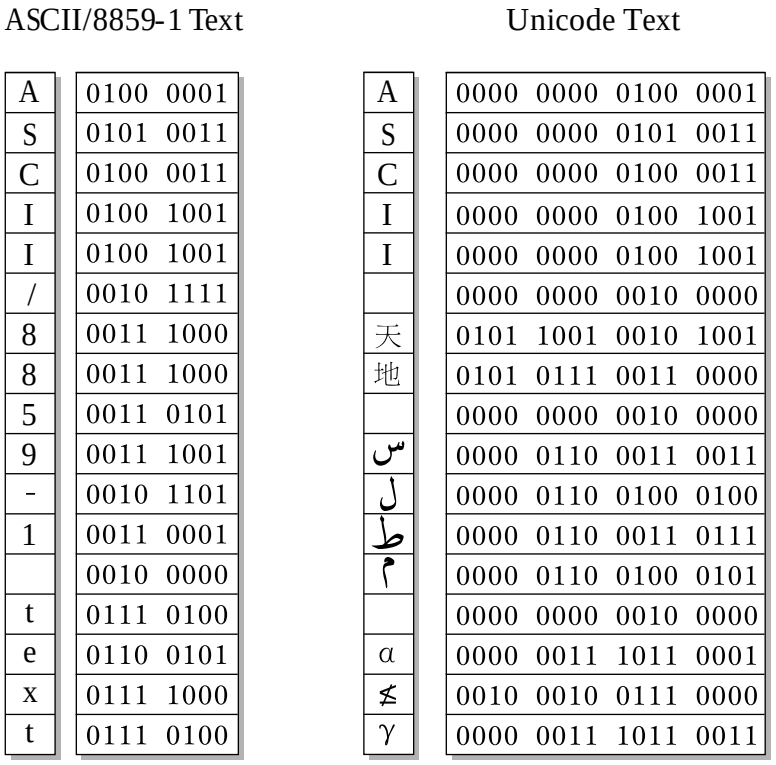
The Unicode Standard specifies a numeric value (code point) and a name for each of its characters. In this respect, it is similar to other character encoding standards from ASCII onward. In addition to character codes and names, other information is crucial to ensure legible text: a character's case, directionality, and alphabetic properties must be well defined. The Unicode Standard defines this and other semantic information, and includes application data such as case mapping tables, character property tables, and mappings to the repertoires of international, national, and industry character sets. The Unicode Consortium provides this additional information to ensure consistency in the implementation and interchange of Unicode data.

Unicode provides for three encoding forms: a 32-bit form (UTF-32), a 16-bit form (UTF-16), and an 8-bit form (UTF-8). The 8-bit, byte-oriented form, UTF-8 has been designed for ease of use with existing ASCII-based systems.

The Unicode Standard, Version 4.0, is code-for-code identical with International Standard ISO/IEC 10646. Any implementation that is conformant to Unicode is therefore conformant to ISO/IEC 10646.

The Unicode Standard provides 1,114,112 code points, most of which are available for encoding of characters. The majority of the common characters used in the major languages of the world are encoded in the first 65,536 code points, also known as the Basic Multilingual Plane (BMP). The overall capacity for more than a million characters is more than sufficient for all known character encoding requirements, including full coverage of all minority and historic scripts of the world.

Figure 1-1. Wide ASCII



[Error in Fig 1-1 in third column, six up from bottom. See Joe's 2-29-00 or 3-7-00 email to errata. Need two changes: 0627 ALEF rather than have 0637; also in third column 0011 should go to 0010]

1.1 Coverage

The Unicode Standard, Version 4.0, contains 95,xxx characters from the world's scripts. These characters are more than sufficient not only for modern communication, but also for the classical forms of many languages. The Standard includes the European alphabetic scripts, Middle Eastern right-to-left scripts, and scripts of Asia, as well as many others. The unified Han subset contains 70,207 ideographic characters defined by national and industry standards of China, Japan, Korea, Taiwan, Vietnam, and Singapore. In addition, the Unicode Standard includes punctuation marks, mathematical symbols, technical symbols, geometric shapes, and dingbats.

[Number of characters above needs updating for 4.0]

Many new scripts have been added between Version 3.0 and Version 4.0, including Old Italic, Gothic, Deseret, Shavian, four Philippine scripts, Osmanya, Linear B, Cypriot, Ugaritic, Limbu, and Tai Le. Many other characters have been added as well. The number of Han ideographs in the standard has more than doubled. A large collection of phonetic characters has been included to support phonetic transcriptions other than IPA, and hundreds of mathematical symbols have been added to fill out the repertoire for mathematical and other technical publishing. Overall character allocation and code ranges are detailed in *Chapter 2, General Structure*.

Note, however, that the Unicode Standard does not encode idiosyncratic, personal, novel, rarely exchanged, or private-use characters, nor does it encode logos or graphics. Graphologies unrelated to text, such as dance notations, are likewise outside the scope of the Unicode Standard. Font variants are explicitly not encoded. The Unicode Standard reserves 6,400 code points in the Basic Multilingual Plane (BMP) for private use, which may be used to assign codes to characters not included in the repertoire of the Unicode Standard. Another 131,068 private-use code points are available outside the BMP, should 6,400 prove insufficient for particular applications.

Standards Coverage

The Unicode Standard is a superset of all characters in widespread use today. It contains the characters from major international and national standards as well as prominent industry character sets. For example, Unicode incorporates the ISO/IEC 6937 and ISO/IEC 8859 families of standards, the SGML standard ISO/IEC 8879, and bibliographic standards such as ISO 5426. Important national standards contained within Unicode include ANSI Z39.64, KS X 1001, JIS X 0208, JIS X 0212, JIS X 0213, GB 2312, GB 18030, and CNS 11643. Industry code pages and character sets from Adobe, Apple, Fujitsu, Hewlett-Packard, IBM, Lotus, Microsoft, NEC, and Xerox are fully represented as well.

For a complete list of ISO and national standards used as sources, see *References*.

New Characters

The Unicode Standard continues to respond to new and changing industry demands by encoding important new characters. As an example, when the need to support the euro sign arose, *The Unicode Standard, Version 2.1*, with euro support was issued to ensure a conforming version.

As the universal character encoding scheme, the Unicode Standard must also respond to scholarly needs. To preserve world cultural heritage, important archaic scripts are encoded as proposals are developed.

For information on how to submit proposals for new characters to the Unicode Consortium, see “Submitting New Characters” in *Section 1.5, The Unicode Consortium*.

1.2 Design Basis

The primary goal of the development effort for the Unicode Standard was to remedy two serious problems common to most multilingual computer programs. The first problem was the overloading of the font mechanism when encoding characters. Fonts have often been indiscriminately mapped to the same set of bytes. For example, the bytes 0x00 to 0xFF are often used for both characters and dingbats. The second major problem was the use of multiple, inconsistent character codes because of conflicting national and industry character standards. In Western European software environments, for example, one often finds confusion between the Windows Latin 1 code page 1252 and ISO/IEC 8859-1. In software for East Asian ideographs, the same set of bytes used for ASCII may also be used as the second byte of a double-byte character. In these situations, software must be able to distinguish between ASCII and double-byte characters.

The ASCII 7-bit code space and its 8-bit extensions, although used in most computing systems, are limited to 128 and 256 code positions, respectively. These 7- and 8-bit code spaces are completely inadequate in the global computing environment.

When the Unicode project began in 1988, the groups most affected by the lack of a consistent international character standard included publishers of scientific and mathematical software, newspaper and book publishers, bibliographic information services, and academic researchers. More recently, the computer industry has adopted an increasingly global outlook, building international software that can be easily adapted to meet the needs of particular locations and cultures. The explosive growth of the Internet has merely added to the demand for a character set standard that can be used all over the world.

The designers of the Unicode Standard envisioned a uniform method of character identification that would be more efficient and flexible than previous encoding systems. The new system would satisfy the needs of technical and multilingual computing and would encode a broad range of characters for professional-quality typesetting and desktop publishing worldwide.

The Unicode Standard was designed to be:

- *Universal.* The repertoire must be large enough to encompass all characters that are likely to be used in general text interchange, including those in major international, national, and industry character sets.
- *Efficient.* Plain text is simple to parse: software does not have to maintain state or look for special escape sequences, and character synchronization from any point in a character stream is quick and unambiguous. A fixed character code allows for efficient sorting, searching, display, and editing of text.
- *Unambiguous.* Any given Unicode code point always represents the same character.

Figure 1-2 demonstrates some of these features, contrasting the Unicode encoding with mixtures of single-byte character sets with escape sequences to shift the meanings of bytes in the ISO 2022 framework using multiple character encoding standards.

1.3 Text Handling

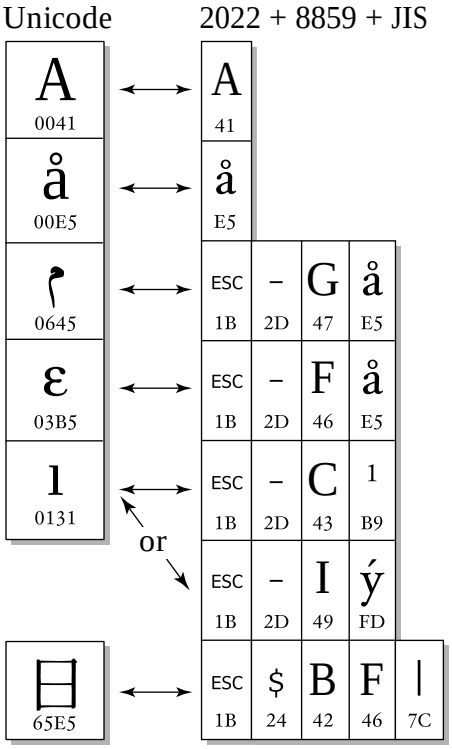
Computer text handling involves processing and encoding. When a word processor user types in the letter “T” via a keyboard, the computer’s system software receives a message that the user pressed a key combination for “T”, which it encodes as U+0054. The word processor stores the number in memory and also passes it on to the display software responsible for putting the character on the screen. This display software, which may be a windows manager or part of the word processor itself, then uses the number as an index to find an image of a “T”, which it draws on the monitor screen. The process continues as the user types in more characters.

The Unicode Standard directly addresses only the encoding and semantics of text and not any other actions performed on the text. In the preceding scenario, the word processor might check the typist’s input after it has been encoded to look for misspelled words, and then highlight any errors it finds. Alternatively, the word processor might insert line breaks when it counts a certain number of characters entered since the last line break. An important principle of the Unicode Standard is that the standard does not specify how to carry out these processes as long as the character encoding and decoding is performed properly and the character semantics are maintained.

Interpreting Characters

The difference between identifying a character and rendering it on screen or paper is crucial to understanding the Unicode Standard’s role in text processing. The character identi-

Figure 1-2. Universal, Efficient, and Unambiguous



fied by a Unicode code point is an abstract entity, such as "LATIN CAPITAL LETTER A" or "BENGALI DIGIT FIVE". The mark made on screen or paper, called a glyph, is a visual representation of the character.

The Unicode Standard does not define glyph images. That is, the standard defines how characters are interpreted, not how glyphs are rendered. Ultimately, the software or hardware rendering engine of a computer is responsible for the appearance of the characters on the screen. The Unicode Standard does not specify the precise shape, size, or orientation of on-screen characters.

Text Elements

The successful encoding, processing, and interpretation of text requires appropriate definition of useful elements of text and the basic rules for interpreting text. The definition of text elements often changes depending on the process handling the text. For example, when searching for a particular word or character written with the Latin script, one often wishes to ignore differences of case. However, correct spelling within a document requires case sensitivity.

The Unicode Standard does not define what is and is not a text element in different processes; instead, it defines elements called *encoded characters*. An encoded character is represented by a number from 0 to $10FFFF_{16}$, called a code point. A text element, in turn, is represented by a sequence of one or more encoded characters.

1.4 The Unicode Standard and ISO/IEC 10646

The Unicode Standard is fully compatible with the International Standard ISO/IEC 10646-1:2000, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)—Part 1: Architecture and Basic Multilingual Plane,* and ISO/IEC 10646-2:2001, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)—Part 2: Supplementary Planes,* which are together known as the Universal Character Set (UCS). Note that future editions of ISO/IEC 10646 may merge Part 1 and Part 2 into a single publication.

During 1991, the Unicode Consortium and the International Organization for Standardization (ISO) recognized that a single, universal character code was highly desirable. A formal convergence of the two standards was negotiated, and their repertoires were merged into a single character encoding in January 1992. Since then, close cooperation and formal liaison between the committees have ensured that all additions to either standard are coordinated and kept synchronized, so that the two standards maintain exactly the same character repertoire and encoding.

Version 4.0 of the Unicode Standard is code-for-code identical to ISO/IEC 10646-1:2000, ISO/IEC 10646-2:2001, and their published amendments. This code-for-code identity holds true for all encoded characters in the two standards, including the East Asian (Han) ideographic characters. ISO/IEC 10646 provides character names and code points; the Unicode Standard provides the same names and code points plus important implementation algorithms, properties, and other useful semantic information.

For details about the Unicode Standard and ISO/IEC 10646, see *Appendix C, Relationship to ISO/IEC 10646,* and *Appendix D, Changes from Unicode Version 3.0.*

1.5 The Unicode Consortium

The Unicode Consortium was incorporated in January 1991, under the name Unicode, Inc., to promote the Unicode Standard as an international encoding system for information interchange, to aid in its implementation, and to maintain quality control over future revisions.

To further these goals, the Unicode Consortium cooperates with the Joint Technical Committee 1 of the International Organization for Standardization (ISO/IEC JTC1). It holds a Class C liaison membership with ISO/IEC JTC1/SC2; it participates in the work of both JTC1/SC2/WG2 (the technical working group for the subcommittee within JTC1 responsible for character set encoding) and the Ideographic Rapporteur Group (IRG) of WG2. The Consortium is a member company of the National Committee for Information Technology Standards, Technical Committee L2 (NCITS/L2), an accredited U.S. standards organization. In addition, Full Member companies of the Unicode Consortium have representatives in many countries who also work with other national standards bodies.

A number of organizations are Liaison Members of the Unicode Consortium: the Center for Computer and Information Development (CCID, China), European Computer Manufacturers Association (ECMA), High Council of Informatics of Iran (HCI), the International Forum for Information Technology in Tamil (INFITT), the Internet Engineering Task Force (IETF), ISO/IEC JTC1/SC2/WG2, ISO/IEC JTC1/SC22/WG20, the Kongju National Library (Chung-nam, Korea), SC2 (the international standardization subcommittee for coded character sets), the Technical Committee on Information Technology (TCVN/TC1, Viet Nam), and the World Wide Web Consortium (W3C) I18N Working Group.

[Update liaisons above again later; current as of April, 2001; check SC2 liaison status--Ken's Q. 8-15-01]

Membership in the Unicode Consortium is open to organizations and individuals anywhere in the world who support the Unicode Standard and who would like to assist in its extension and widespread implementation. Full and Associate Members represent a broad spectrum of corporations and organizations in the computer and information processing industry. The Consortium is supported financially solely through membership dues.

The Unicode Technical Committee

The Unicode Technical Committee (UTC) is the working group within the Consortium responsible for the creation, maintenance, and quality of the Unicode Standard. The UTC follows an open process in developing the Unicode Standard and its other technical publications. It coordinates and reviews all technical input to these documents and decides their contents. Full Members of the Consortium vote on UTC decisions. Associate and Specialist Members and Officers of the Unicode Consortium are nonvoting UTC participants. Other attendees may participate in UTC discussions at the discretion of the Chair, as the intent of the UTC is to act as an open forum for the free exchange of technical ideas. For more information on the UTC and the process by which the Unicode Standard and the other technical publications are developed, see:

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

Submitting New Characters

The Unicode Consortium accepts proposals for inclusion of new characters and scripts in the Unicode Standard. All proposals must be in writing, must include at least one picture of each proposed character (normally from a printed source), and must include significant documentation justifying the proposal. Those considering submitting a proposal should first determine whether a particular script or character has already been proposed. The identification of the sponsor(s) must be included, along with postal address and an electronic mail address or phone number. Please consult the Unicode Consortium's Web site (*http://www.unicode.org*) for the most current guidelines. The Web site also provides information on proposed new scripts and characters, which may help to determine whether a script or character has already been proposed.

Before preparing a proposal, sponsors should note in particular the distinction between the terms *character* and *glyph* as defined in this standard. Because of this distinction, graphics such as ligatures, conjunct consonants, minor variant written forms, or abbreviations of longer forms are generally not acceptable as Unicode characters. For further information about how to find out whether a given character is already covered by the Unicode Standard, see the "Where is my Character?" page on the Unicode Web site at:

http://www.unicode.org/unicode/standard/where/

Experience has shown that it is often helpful to discuss preliminary proposals before submitting a detailed proposal. One open forum for such discussion is the Unicode mailing list. Please see the Unicode Web site for instructions on how to subscribe to the mailing list. Sponsors are urged to send a message of inquiry or a preliminary proposal there before formal submission. Many problems and questions can be dealt with there.

Each proposal received will be evaluated initially by technical officers of the Unicode Consortium and the result of this evaluation will be communicated to the sponsor(s) of the proposal. All proposals, whether successful or not, will be retained by the Unicode Consortium as a matter of record.