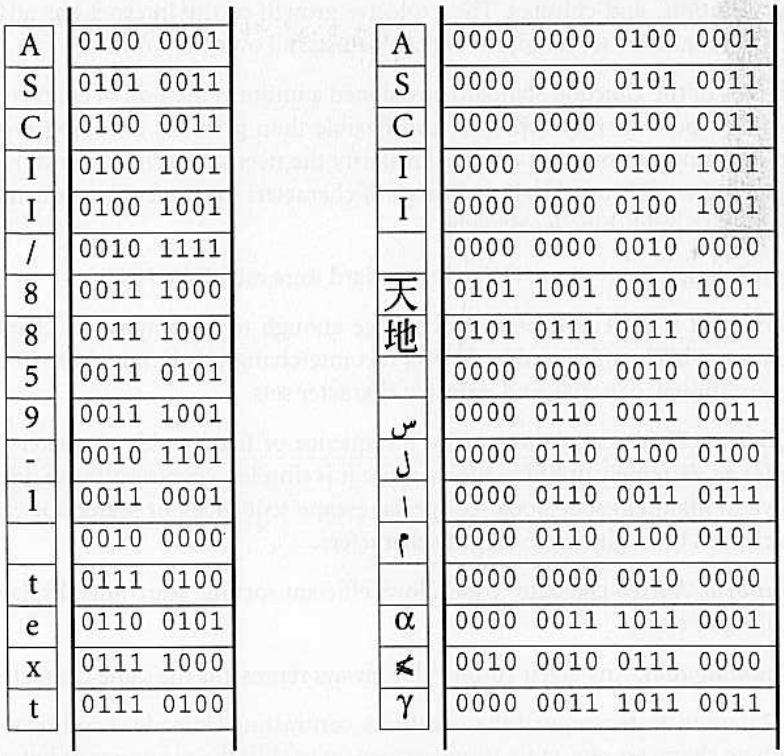
Chapter 1

Introduction

The Unicode Standard is a fixed-width, uniform encoding scheme for written characters and text. The repertoire of this international character code for information processing includes characters for the major scripts of the world, as well as technical symbols in common use. The Unicode character encoding treats alphabetic characters, ideographic characters, and symbols identically, which means that they can be used in any mixture and with equal facility. The Unicode Standard is modeled on the ASCII character set, but uses a 16-bit encoding to support full multilingual text. No escape sequence or control code is required to specify any character in any language.

Figure 1-1. Wide ASCII



The Unicode Standard specifies a numerical value and a name for each of its characters; in this respect, it is similar to other character encoding standards from ASCII onwards (see Figure 1-1). The Unicode Standard is code-for-code identical with International Standard ISO/IEC 10646-1:1993, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)—Part 1: Architecture and Basic Multilingual Plane.*

As well as assigning character codes and names, the Unicode Standard provides other information not found in conventional character set standards, but crucial for using character encoding in implementations. The Unicode Standard defines properties for characters and includes application data such as case mapping tables and mappings to the repertoires of international, national, and industry character sets. The Unicode Consortium provides this additional information to ensure consistency in interchange of Unicode data.

1.1 Design Goals

The primary goal of the development effort for the Unicode Standard was to remedy two serious problems common to most multilingual computer programs: overloading of the font mechanism when encoding characters, and use of multiple, inconsistent character codes due to conflicting national and industry character standards. 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 inadequate in the global computing environment.

When the Unicode project began in 1988, groups most affected by the lack of a consistent international character standard included the 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 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 which would be more efficient and flexible than previous encoding systems. The new system would be complete enough to 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 original design goals of the Unicode Standard were established as:

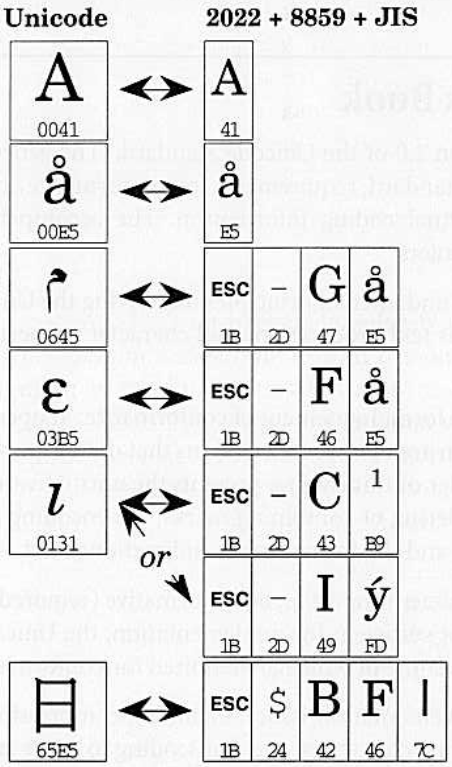
- ***Universal.*** The repertoire must be large enough to encompass all characters that were likely to be used in general text interchange, including those in major international, national, and industry character sets.
- ***Efficient.*** Plain text, composed of a sequence of fixed-width characters, provides an extremely useful model because it is simple to parse: software does not have to maintain state, look for special escape sequences, or search forward or backward through text to identify characters.
- ***Uniform.*** A fixed character code allows efficient sorting, searching, display, and editing of text.
- ***Unambiguous.*** Any given 16-bit value always represents the same character.

Figure 1-2 demonstrates some of these features, contrasting Unicode encoding to mixtures of single-byte character sets, with escape sequences to shift the meanings of bytes.

1.2 Coverage

The Unicode Standard, Version 2.0 contains 38,885 characters from the world's scripts. These characters are more than sufficient not only for modern communication, but also for

Figure 1-2. Universal, Efficient, and Unambiguous



the classical forms of many languages. Languages that can be encoded include Russian, Arabic, Anglo-Saxon, Greek, Hebrew, Thai, and Sanskrit. The unified Han subset contains 20,902 ideographic characters defined by national and industry standards of China, Japan, Korea, and Taiwan. In addition, the Unicode Standard includes mathematical operators and technical symbols, geometric shapes, and dingbats. Overall character allocation and the code ranges are detailed in *Chapter 2, General Structure.*

Included in the Unicode Standard are characters from all major international standards approved and published before December 31, 1990, in particular, the ISO International Register of Character Sets, the ISO/IEC 6937 and ISO/IEC 8859 families of standards, as well as ISO/IEC 8879 (SGML). Other primary sources included bibliographic standards used in libraries (such as ISO/IEC 5426 and ANSI Z39.64), the most prominent national standards, and various industry standards in very common use (including code pages and character sets from Adobe, Apple, Fujitsu, Hewlett-Packard, IBM, Lotus, Microsoft, NEC, WordPerfect, and Xerox). The complete Hangul repertoire of Korean National Standard KS C 5601 was added in Version 2.0. For a complete list of ISO and national standards used as sources, see the bibliography.

The Unicode Standard does not encode idiosyncratic, personal, novel, rarely exchanged, or private-use characters, nor does it encode logos or graphics. Artificial entities, whose sole function is to serve transiently in the input of text, are excluded. Graphologies unrelated to text, such as musical and dance notations, are outside the scope of the Unicode Standard. Font variants are explicitly not encoded. The Unicode Standard includes a *Private Use Area*, which may be used to assign codes to characters not included in the repertoire of the Unicode Standard.

The Unicode Consortium (see *Section 1.4, The Unicode Consortium*) periodically develops proposals for new scripts. The Consortium welcomes the submission of new characters for

possible inclusion in the Unicode Standard. (For instructions on how to submit characters to the Unicode Consortium, see *Appendix B, Submitting New Characters.*)

1.3 About This Book

This book defines Version 2.0 of the Unicode Standard. The general principles and architecture of the Unicode Standard, requirements for conformance, and guidelines for implementers precede the actual coding information. The accompanying CD-ROM carries tables of use to implementers.

Chapter 2 sets forth the fundamental principles underlying the Unicode Standard and covers specific topics such as text processes, overall character properties, and the use of nonspacing marks.

Chapter 3 constitutes the formal statement of conformance. It opens with the conformance clauses themselves, which are followed by sections that define more precisely terms used in the clauses. The remainder of this chapter presents the normative algorithms for three processes: the canonical ordering of combining marks, the encoding of Korean Hangul syllables by conjoining *jamo*, and the formatting of bidirectional text.

Chapter 4 describes character properties, both normative (required) and informative. Since code charts alone are not sufficient for implementation, the Unicode Standard also specifies character properties, some of which are required for conformance.

Chapter 5 discusses implementation issues, including compression, strategies for dealing with unknown and missing characters, and transcoding to other standards.

Chapter 6 contains character block descriptions, part of the coding information in the Unicode Standard. A character block generally contains characters from a single script (for example, *Tibetan*) or is a collection of a particular type of character (for example, *Mathematical Operators*). A character block description gives basic information about the script or collection and may discuss specific characters.

Chapter 7 presents the individual characters, arranged by character block. An overview of a particular character block is given by means of a code chart. With the exception of the blocks for East Asian ideographs and Korean hangul syllables, the individual characters of a block are identified in the accompanying names list.

Chapter 8 provides a radical/stroke index to East Asian ideographs.

Appendix A describes the various encoding forms that may be applied to Unicode character data to meet particular needs, for example, UTF-7 to facilitate the exchange of Unicode data in 7-bit environments.

Appendix B gives instructions on how to submit characters for consideration as additions to the Unicode Standard.

Appendix C gives the details of the merger of the Unicode Standard and ISO/IEC 10646, which occurred in 1991.

Appendix D lists the changes to the Unicode Standard since Version 1.0.

Appendix E describes the history of Han Unification in the Unicode Standard.

The appendices are followed by a glossary of terms, a bibliography, and two indices: an index to Unicode characters and an index to the text of Chapters 1 through 8.

The major table on the CD-ROM is the *Unicode Character Database*, which gives character codes, character names (with Version 1.0 name if different), character properties, and decompositions for decomposable or compatibility characters. The CD-ROM also includes

property-based mapping tables (for example, tables for case) and transcoding tables for international, national, and industry character sets (including the Han cross-reference table). (For the complete contents of the CD-ROM, see its *READ ME* file.)

Notational Conventions

Throughout this book, certain typographic conventions are used. In running text, an individual Unicode value is expressed as *U+nnnn*, where *nnnn* is a four digit number in hexadecimal notation, using the digits 0–9 and the letters A–F (for 10 through 15 respectively). In tables, the *U+* may be omitted for brevity.

- U+0416 is the Unicode value for the character named CYRILLIC CAPITAL LETTER ZHE.

A range of Unicode values is expressed as U+xxxx →U+yyyy or U+xxxx — U+yyyy, where *xxxx* and *yyyy* are the first and last Unicode values in the range, and the arrow or long dash indicates a contiguous range.

- The range U+0900→U+097F contains 128 character values.

All Unicode characters have unique names, which are identical to those of the English language version of International Standard ISO/IEC 10646. Unicode character names contain only uppercase Latin letters A through Z, space, and hyphen-minus; this convention makes it easy to generate computer language identifiers automatically from the names. Unified East Asian ideographs are named CJK UNIFIED IDEOGRAPH-X, where X is replaced with the hexadecimal Unicode value; for example, CJK UNIFIED IDEOGRAPH-4E00. The names of Hangul syllables are generated algorithmically; for details, see Hangul Syllable Names in *Section 3.10, Combining Jamo Behavior.*

In running text, a formal Unicode name is shown in small capitals (for example, GREEK SMALL LETTER MU), and alternative names (aliases) appear in italics (for example, *umlaut*). Italics are also used to refer to a text element that is not explicitly encoded (for example, *pasekh alef*), or to set off a foreign word (for example, the Welsh word *ynghyd*).

The symbols used in the character names list are described at the beginning of *Chapter 7, Code Charts.*

In the text of this book, the word "Unicode" if used alone as a noun refers to the Unicode Standard or a Unicode character value.

1.4 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. The Unicode Consortium is the central focus and contact point for conducting these activities.

To further these goals, the Unicode Consortium cooperates with the International Organization for Standardization (ISO). The Consortium holds a Class C liaison membership with ISO/IEC JTC1/SC2; it participates both in the work of JTC1/SC2/WG2 (the working group within ISO responsible for computer character sets) and in the work of the Ideographic Rapporteur Group of WG2. The Consortium is a member company of ANSI Subcommittee X3L2. In addition, member representatives in many countries also work with their national standards bodies.

A number of standards organizations are Liaison Members of the Unicode Consortium: ECMA (a European-based organization for standardizing information and communication systems), Association of Common Chinese Code of the Center for Computer & Information Development Research (China), and the Technical Committee on Information Technology of the Viet Nam General Department for Standardization, Metrology, and Quality Control (TCVN/TC), and the WG2 standards committee of Korea.

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 through the volunteer efforts of employees of member companies and individual members, and financially 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 controls all technical input to the standard and makes associated content decisions. UTC members represent the companies that are Full and Associate Members of the Consortium. Observers are welcome to attend UTC meetings and may participate in the discussions, since the intent of the UTC is to act as an open forum for the free exchange of technical ideas.

1.5 The Unicode Standard and ISO/IEC 10646

During 1991, the Unicode Consortium and the International Organization for Standardization (ISO) recognized that a single, universal character code was highly desirable. Mutually acceptable changes were made to Version 1.0 of the Unicode Standard and to the first ISO/IEC Draft International Standard DIS 10646.1, and their repertoires were merged into a single character encoding in January 1992. After international ballot and editorial changes to accommodate comments, the final ISO standard was published in May 1993 as ISO/IEC 10646-1:1993, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)—Part 1: Architecture and Basic Multilingual Plane.*

In accord with the merger agreement, a revision of the Unicode Standard was published in 1993 as *Unicode Technical Report, No. 4*, with the title: *The Unicode Standard, Version 1.1, Prepublication Edition.* Version 1.1 of the Unicode Standard specified a repertoire and set of code assignments identical to those of the new ISO/IEC standard.

After the initial release of ISO/IEC 10646 and the Unicode Standard Version 1.1, both ISO JTC1/SC2/WG2 (the ISO working group responsible for ISO/IEC 10646) and the Unicode Technical Committee continued to develop the merged standard. These developments lead to Version 2.0 of the Unicode Standard, incorporating the first seven amendments made to or proposed for ISO/IEC 10646. (For details, see *Appendix C, Relationship to ISO/IEC 10646*, and *Appendix D, Cumulative Changes.)*

1.6 Resources

On-line Information Sources

The Unicode Consortium provides a number of on-line resources for obtaining information and data about the Unicode Standard. They are

- the World-Wide Web site at URL `http://www.unicode.org`
- the anonymous FTP site at URL `ftp://unicode.org`

Use account name `anonymous`, and specify your electronic mail address as the password

- the mailing list `unicode@unicode.org`

Note that this is a mailing list, not a listserv. To be added or removed, send a request to

- `unicode-request@unicode.org`

How to Contact the Unicode Consortium

Contact the Consortium for membership inquiries and to order publications (including additional copies of this book).

- Electronic mail address: `unicode-inc@unicode.org`
- Postal address:

  P.O. Box 700519
  San Jose, CA 95170-0519
  U.S.A.

- Telephone: +1 (408) 777-3721
- Fax: +1 (408) 777-3784
- Courier deliveries only: 10000 Torre Ave, Cupertino, CA 95014, U.S.A.