Recent Trends in Standardization of Japanese Character Codes

Taichi Kawabata

Abstract

Character encodings are a basic and fundamental layer of digital text that are necessary for exchanging information over the Internet. Character codes are used in plain text files as well as in structured text files such as XML (extensible markup language), MIME (multipurpose Internet mail extensions), and HTML (hypertext markup language).

This article explains the basic architecture of character encodings and reviews Japanese and International standardization activities.

1. Overview of character code standardization history

Character codes are used everywhere on the Internet. They form the most basic layer of the digital text hierarchy. This means that above the plain raw character codes lie layers of variant glyph encodings, text encoding formats such as HTML (hypertext markup language) and XML (extensible markup language), and instructions for the placement of such character codes, such as those given in CSS (cascading style sheets), as shown in **Fig. 1**.

The first attempt to standardize character codes for electronic information interchange was initiated in the early 1960s by the International Organization for Standardization (ISO), followed by the American Standards Association (ASA, now called the ANSI (American National Standards Institute)) and the European Computer Manufacturer's Association (ECMA) as 6-bit and 7-bit character codes. Since then, character codes have been standardized by individual countries or organizations[*1]. The first 2-byte (14-bit) character coding system was standardized by Japan in 1978. Later, China, Korea, and Taiwan also standardized their own 2-byte character codes. These various character codes were not compatible with each other. As a result, some characters were doubly encoded in different character encodings, thus making their combined use practically difficult.

In the late 1980s, some software companies began an attempt to unify these various character codes into a single coding system, which was called Unicode. Similar attempts were also made within the ISO, and with some vicissitudes, these two attempts achieved a *pari passu* (literal meaning: on equal footing) advance, resulting in two technically identical standards, namely, ISO 10646 and Unicode. Since then, both ISO/IEC (International Electrotechnical Commission) and Unicode have continued to publish technically identical standards to this day.

The first version of ISO 10646, whose official name is the Universal Coded Character Set (UCS), was published in 1993. Since then, it has gradually been disseminated all over the world. In Japan, JIS (Japanese Industrial Standards) and its variant character encoding systems (such as Shift_JIS) have been widely adopted, but now UCS prevails on the Japanese Internet[*2]. Today, Japan has already ceased creating its own standards and now proposes new characters to UCS[*3]. A brief history of the standardization achievements over the years is shown in **Fig. 2**.

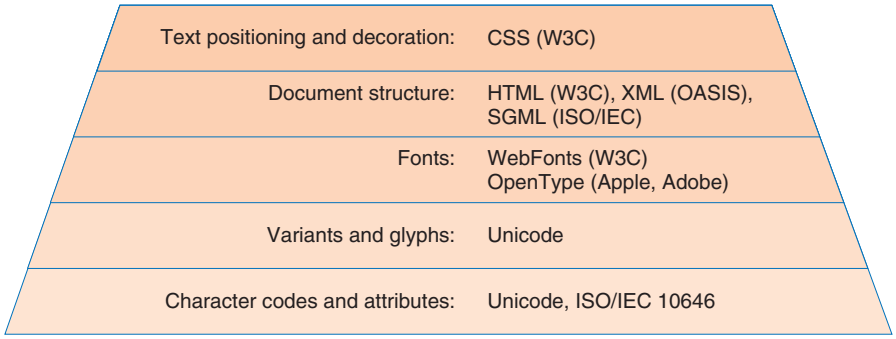
2. Architecture of UCS

The UCS character coding system consists of 17 planes; each plane has 65,534 character code points, which are numerical values representing the code

*1 Basic character encoding frameworks were standardized internationally as ISO 646 or ISO 2022.

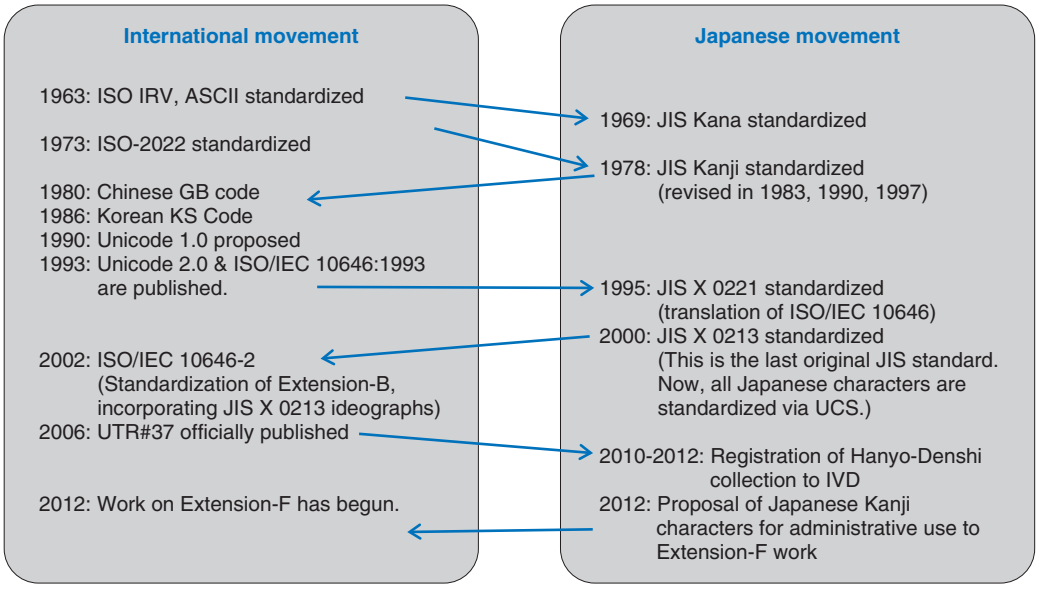
*2 Except for email, in which the JIS coding system is still widely employed.

*3 The JIS translation of UCS is referred to as JIS X 0221.



ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
OASIS: Organization for the Advancement of Structured Information Standards
SGML: standard generalized markup language
W3C: World Wide Web Consortium

Fig. 1. Digital text hierarchy showing character/coding characteristics and details related to their standardization.



ASCII: American Standard Code for Information Interchange
IRV: International Reference Version
IVD: Ideographic Variation Database
UTR: Unicode Technical Report

Fig. 2. Brief history of character encoding standardization.

space. This results in nearly 1.1 million code points for all 17 planes, as shown in **Fig. 3**. In the UCS, unique code points are assigned to characters from all over the world. Today, approximately 100,000 character code points have been assigned, as shown in the figure[*4].

The UCS code points are represented in the form of U+XXXX, where XXXX is a 4 to 6 hexadecimal digit. For example, the Latin character *a* is assigned the code point U+0061, and the kanji character 漢 is assigned the code point U+6F22. Since its standardization in 1993, most characters in modern scripts have been encoded; however, efforts to assign archaic

*4 About 70% of them are CJK (Chinese/Japanese/Korean) Unified Ideographs.

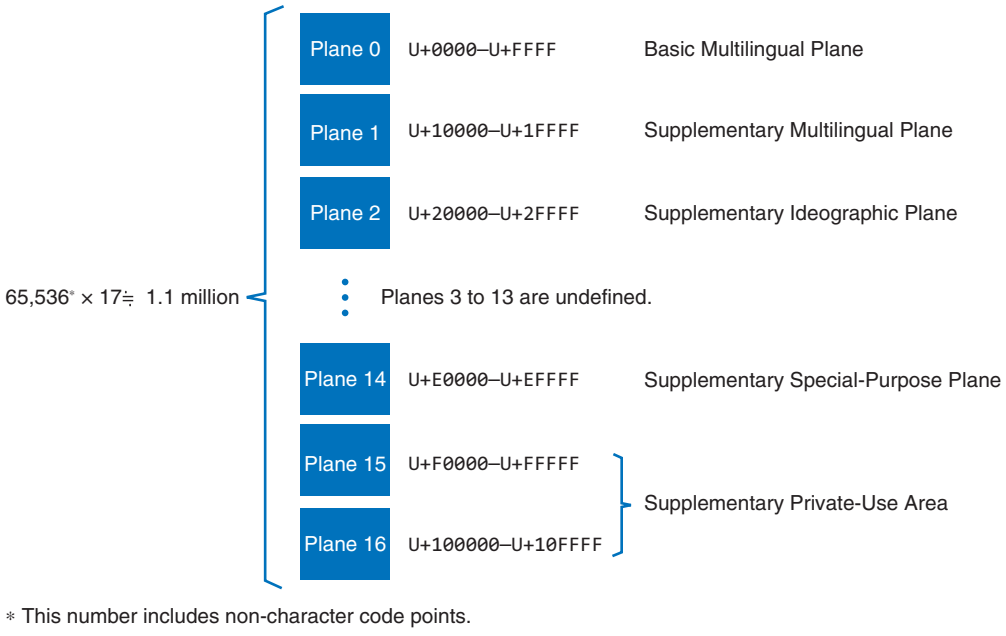


Fig. 3. Planes of UCS character coding system.

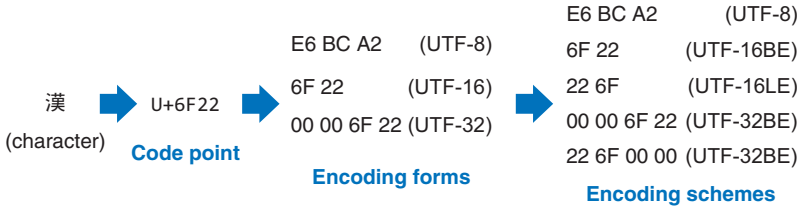


Fig. 4. Example of a character and its codepoints and encoding forms/schemes.

characters and scholarly symbols are still continuing.

For actual communication or memory storage of characters, code points must be *encoded* and *serialized*. This means that these character code points will be converted to one or more 8-bit or 16-bit code unit sequences. UCS defines three encoding forms, namely UTF (UCS Transformation Format)-8, UTF-16, and UTF-32, and seven encoding schemes, namely, UTF-8, UTF-16, UTF-16BE, UTF-16LE, UTF-32, UTF-32BE, and UTF-32LE. An example of their use is shown in **Fig. 4**.

Among these, UTF-8 is upward compatible with ASCII (American Standard Code for Information Interchange) character codes and is widely used in UNIX-based operating systems (OSs) and in communication over the Internet. Windows OS and Java VM (Virtual Machine) have adopted UTF-16 for their internal representation and storage.

UTF-8 will encode UCS code points to one to four 8-bit code units. UTF-16 will encode UCS code points to one or two 16-bit code units. The correspondence between UCS code points and their UTF-8 encodings are shown in **Table 1**. As shown in this table, the first unit of UTF-8 code units can be easily distinguished among the remaining code units.

3. Ideographic Variation Database and its selectors

UCS is used to encode abstract characters. For example, the ideograph 葛 would have the same character code regardless of its shape or font. However, for some special scholarly, publishing, or administra-

Table 1. UTF-8 codepoints.

Codepoints	Bit arrays	Byte sequences
U+0000–U+007F	0xxx xxxx	00-7F
U+0080–U+07FF	110yyyyx 10xxxxxx	C2-DF 80-BF
U+0800–U+FFFF	1110zzzz 10yyyyyy 10xxxxxx	E0-EF 80-BF 80-BF
U+10000–U+10FFFF	11110uuu 10uuzzzz 10yyyyyy 10xx xxxx	F0-F7 80-BF 80-BF 80-BF

tive usages, slight differences in shape that would otherwise be ignored during encoding have sometimes resulted in significant differences.

For example, in Japanese administrative systems, Katsushika Ward in Tokyo (葛飾区) and Katsuragi City (葛城市) in Nara Prefecture would have the same character "葛" with ordinary character encoding, yet they might be displayed differently in official administrative documents.

UCS provides a method to differentiate such variants that would otherwise be unified with the Ideographic Variation Database (IVD) and Ideographic Variation Sequence (IVS). IVD was introduced in 2006 as the Unicode Technical Standard (UTS) #37.

This scheme involves attaching the variation selector after the character code, which makes it possible to represent ideographic variations.

In the above examples of 葛 (Katsushika) and 葛 (Katsuragi), the different variations can be specified as U+845B U+E0103 and U+845B U+E0102, respectively, as shown in **Fig. 5**. There are 240 variation selectors for ideographs assigned from U+E0100 to U+E01EF of the UCS 14th plane.

Currently, the IVD is maintained by the Unicode Consortium, a nonprofit organization that develops the Unicode Standard.

According to UTS #37, when someone wants to register a variation of an ideograph, he/she applies for registration of variations along with its name to the Unicode Consortium (along with the requested variants).

Then the Unicode Consortium opens the application to public review for three months. After the public review has closed and comments have been reflected in the application, the Unicode Consortium will assign an IVS to each glyph.

Currently, two collections are registered: Adobe-Japan1 and Hanyo-Denshi. The latter collection of variants is registered by the Hanyo-Denshi committee of Japan for the administration systems of national and local Japanese governments.

Fig. 5. Registered Hanyo-Denshi variation of "葛".

4. Relationship between domestic and international standardization committees

The character coding system is one of the standards in the information technology (IT) field. International standards in the IT field are currently established by Joint Technical Committee One (JTC 1) of the ISO and the IEC, which is referred to as ISO/IEC JTC 1. There are about 20 subcommittees in JTC 1, and the subcommittee working on character encoding standardization is subcommittee 2 (SC 2). SC 2 has one working group, called WG 2, which is working on ISO/IEC 10646. WG 2 has entrusted the Ideographic Rapporteur Group (IRG) with specifically creating the repertoire of CJK Unified Ideographs for standardization. These committees and subcommittees are often referred to by their abbreviations. For example, WG 2 is officially known as ISO/IEC JTC 1/SC 2/WG 2.

In Japan, the Information Technology Standards Committee of Japan (ITSCJ), a division of the Information Processing Society of Japan (IPSJ) and entrusted by the Ministry of Economy, Trade and Industry, is a corresponding domestic standardization body for most subcommittees of ISO/IEC JTC 1. SC 2 in Japan is working with ISO/IEC JTC 1/SC 2, SC 2/WG 2, and SC 2/WG 2/IRG, to reflect the Japanese concerns and requirements in international character codes. The relationship between the inter-

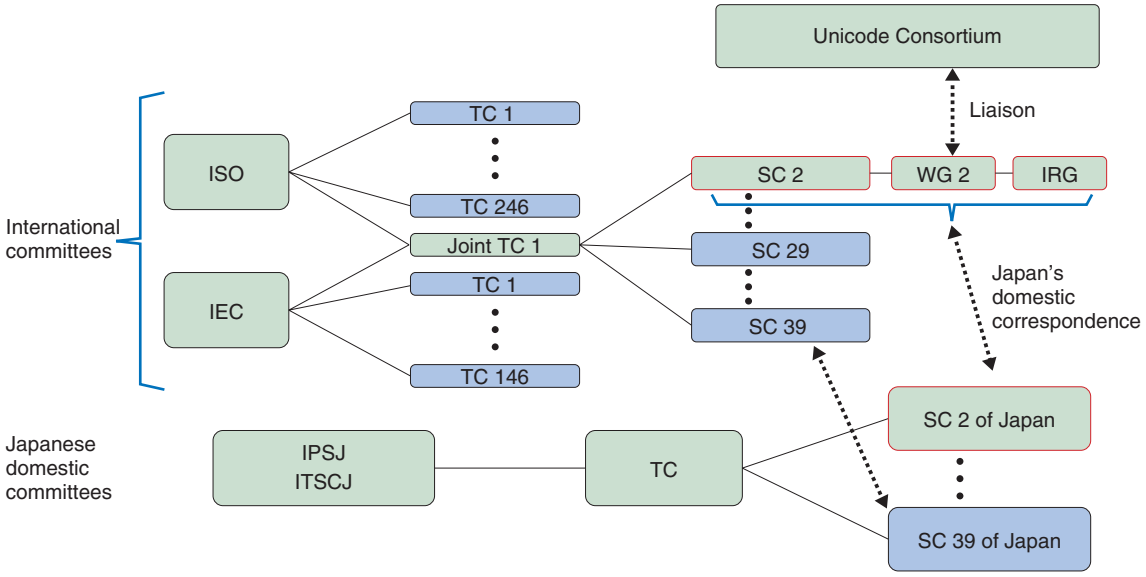


Fig. 6. International and Japanese committees working on character code standardization.

national organizations and the Japanese organizations is shown in **Fig. 6**.

5. Future outlook

Most kanji (CJK Unified) ideographs needed for Japanese governmental administration systems are proposed to the IRG as part of Extension F standardization efforts[*5]. They are expected to be standardized within the next several years.

As devices such as e-books and studies on the digital humanities evolve, the character encodings for kanji ideographs used in classical documents will follow the same path of standardization that ideographs needed for administration systems have. In the current work being done on Extension F, the kanji ideographs that appear in the classical Buddhist canon *Taishō Revised Tripiṭaka* are in the process of being standardized. These efforts mean that many more classical documents will be available on e-books or the Internet in the future.

Emojis (Japanese-style pictorial characters) were also developed by Japanese mobile phone carriers, but their standardization in the UCS was initiated by Apple and Google, who have developed operating systems for cellular phones. Emoji characters also have numerous variations, and to date, nearly 700 characters have been standardized as emoticons and pictographs.

NTT recognizes that standards and technologies related to character encodings will be crucial in applications involving the web, email, and e-books and services, and will therefore stay engaged in this technology.

*5 Standardized CJK Ideographs are currently grouped into Unified Repertoire and Ordering (URO) character sets identified with Extensions. Extensions A to D have already been standardized. Extension E has just passed the vote and will be officially published next year. The work on Extension F has just begun.

Taichi Kawabata
Research Engineer, Ubiquitous Software Group, Ubiquitous Service Systems Laboratory, NTT Network Innovation Laboratories.
He received the M.S. degree in computer science from the University of Tokyo in 1997 and joined NTT in the same year. He has mainly been engaged in researching the usage of personal data on network systems, and has also worked on standardization activities on character, text encoding, and related standards in ISO, IEC, and W3C standardization organizations.