The Unicode® Standard
Version 10.0 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 10.0.
Includes bibliographical references and index.
ISBN 978-1-936213-16-0 (http://www.unicode.org/versions/Unicode10.0.0/)
1. Unicode (Computer character set) I. Unicode Consortium.
QA268.U545 2017

ISBN 978-1-936213-16-0
Published in Mountain View, CA
June 2017

Appendix E

Han Unification History

Efforts to standardize a comprehensive Han character repertoire go back at least as far as the Eastern Han dynasty, when the important dictionary *Shuowen Jiezi* (121 CE) codified a set of some 10,000 characters and variants, crystallizing earlier Qin dynasty initiatives at orthographic reform. Subsequent dictionaries in China grew larger as each generation recombined the *Shuowen* script elements to create new characters. By the time the Qing dynasty *Kang Xi* dictionary was completed in the 18th century, the character set had grown to include more than 40,000 characters and variants. In relatively recent times many more characters and variants have been created and catalogued, reflecting modern PRC simplification and standardization initiatives, as well as ongoing inventories of legacy printed texts.

The effort to create a unified Han character encoding was guided by the developing national standards, driven by offshoots of the dictionary traditions just mentioned, and focused on modern bibliographic and pedagogical lists of characters in common use in various genres. Much of the early work to create national and transnational encoding standards was published in China and Japan in the late 1970s and early 1980s.

The Chinese Character Code for Information Interchange (CCCII), first published in Taiwan in 1980, identified a set of some 5,000 characters in frequent use in China, Taiwan, and Japan. (Subsequent revisions of CCCII considerably expanded the set.) In somewhat modified form, CCCII was adopted for use in the United States as ANSI Z39.64-1989, also known as EACC, the *East Asian Character Code For Bibliographic Use*. EACC encoded some 16,000 characters and variants, organized using a twelve-layer variant mapping mechanism.

In 1980, Takahashi Tokutaro of Japan's National Diet Library proposed ISO standardization of a character set for common use among East Asian countries. This proposal included a report on the first *Japanese Industrial Standard* for *kanji* coding (JIS C 6226-1978). Published in January 1978, JIS C 6226-1978 was growing in influence: it encoded a total of 6,349 *kanji* arranged in two levels according to frequency of use, and approximately 500 other characters, including Greek and Cyrillic.

E.1 Development of the URO

The Unicode Han character set began with a project to create a Han character cross-reference database at Xerox in 1986. In 1988, a parallel effort began at Apple based on the RLG's CJK Thesaurus, which is used to maintain EACC. The merger of the Apple and Xerox databases in 1989 led to the first draft of the Unicode Han character set. At the September 1989 meeting of X3L2 (an accredited standards committee for codes and character sets operating under the procedures of the American National Standards Institute), the Unicode Working Group proposed this set for inclusion in ISO/IEC 10646.

The primary difference between the Unicode Han character repertoire and earlier efforts was that the Unicode Han character set extended the bibliographic sets to guarantee complete coverage of industry and newer national standards. The unification criteria employed in this original Unicode Han character repertoire were based on rules used by JIS and on a set of Han character identity principles (*rentong yuanze*) being developed in China by experts working with the Association for a Common Chinese Code (ACCC). An important principle was to preserve all character distinctions within existing and proposed national and industry standards.

The Unicode Han proposal stimulated interest in a unified Han set for inclusion in ISO/IEC 10646, which led to an ad hoc meeting to discuss the issue of unification. Held in Beijing in October 1989, this meeting was the beginning of informal cooperation between the Unicode Working Group and the ACCC to exchange information on each group's proposals for Han unification.

A second ad hoc meeting on Han unification was held in Seoul in February 1990. At this meeting, the Korean delegation proposed the establishment of a group composed of the East Asian countries and other interested organizations to study a unified Han encoding. From this informal meeting emerged the Chinese/Japanese/Korean Joint Research Group (hereafter referred to as the CJK-JRG).

A second draft of the Unicode Han character repertoire was sent out for widespread review in December 1990 to coincide with the announcement of the formation of the Unicode Consortium. The December 1990 draft of the Unicode Han character set differed from the first draft in that it used the principle of *KangXi* radical-stroke ordering of the characters. To verify independently the soundness and accuracy of the unification, the Consortium arranged to have this draft reviewed in detail by East Asian scholars at the University of Toronto.

In the meantime, China announced that it was about to complete its own proposal for a Han Character Set, GB 13000. Concluding that the two drafts were similar in content and philosophy, the Unicode Consortium and the Center for Computer and Information Development Research, Ministry of Machinery and Electronic Industry (CCID, China's computer standards body), agreed to merge the two efforts into a single proposal. Each added missing characters from the other set and agreed upon a method for ordering the characters using the four-dictionary ordering scheme described in *Section 18.1, Han*. Both proposals benefited greatly from programmatic comparisons of the two databases.

As a result of the agreement to merge the Unicode Standard and ISO/IEC 10646, the Unicode Consortium agreed to adopt the unified Han character repertoire that was to be developed by the CJK-JRG.

The first CJK-JRG meeting was held in Tokyo in July 1991. The group recognized that there was a compelling requirement for unification of the existing CJK ideographic characters into one coherent coding standard. Two basic decisions were made: to use GB 13000 (previously merged with the Unicode Han repertoire) as the basis for what would be termed "The Unified Repertoire and Ordering," and to verify the unification results based on rules that had been developed by Professor Miyazawa Akira and other members of the Japanese delegation.

The formal review of GB 13000 began immediately. Subsequent meetings were held in Beijing and Hong Kong. On March 27, 1992, the CJK-JRG completed the *Unified Repertoire and Ordering* (URO)*, Version 2.0*. This repertoire was subsequently published both by the Unicode Consortium in *The Unicode Standard, Version 1.0*, Volume 2, and by ISO in ISO/IEC 10646-1:1993.

E.2 Ideographic Rapporteur Group

In October 1993, the CJK-JRG became a formal subgroup of ISO/IEC JTC1/SC2/WG2 and was renamed the Ideographic Rapporteur Group (IRG). The IRG now has the formal responsibility of developing extensions to the URO 2.0 to expand the encoded repertoire of unified CJK ideographs. The Unicode Consortium participates in this group as a liaison member of ISO.

In its second meeting in Hanoi in February 1994, the IRG agreed to include Vietnamese Chữ Nôm ideographs in a future version of the URO and to add a fifth reference dictionary to the ordering scheme.

In 1998, the IRG completed work on the first ideographic supplement to the URO, CJK Unified Ideographs Extension A. This set of 6,582 characters was culled from national and industrial standards and historical literature and was first encoded in *The Unicode Standard, Version 3.0*. CJK Unified Ideographs Extension A represents the final set of CJK ideographs to be encoded on the BMP.

In 2000, the IRG completed work on the second ideographic supplement to the URO, a very large collection known as CJK Unified Ideographs Extension B. These 42,711 characters were derived from major classical dictionaries and literary sources, and from many additional national standards, as documented in *Section E.3, CJK Sources*. The Extension B collection was first encoded in *The Unicode Standard, Version 3.1*, and is the first collection of unified CJK ideographs to be encoded on Plane 2.

In 2005, the IRG identified a subset of the unified ideographs, called the International Ideograph Core (IICore). This subset is designed to serve as a relatively small collection of around 10,000 ideographs, mainly for use in devices with limited resources, such as mobile phones. The IICore subset is meant to cover the vast majority of modern texts in all locales where ideographs are used. The repertoire of the IICore subset is identified with the kIICore key in the *Unihan Database*.

Also in 2005, a small set of ideographs was encoded to support the complete repertoire of the GB 18030:2000 and HKSCS 2004 standards. In addition, an initial set of CJK strokes was encoded.

In 2008, the IRG completed work on the third ideographic supplement to the URO, a collection of 4,149 characters from various sources. The Extension C collection was first encoded in the Unicode Standard, Version 5.2.

In 2009, the IRG completed work on the fourth ideographic supplement to the URO, a collection of 222 characters from various sources as documented *Section E.3, CJK Sources*. The Extension D collection represents a small number of characters which IRG members felt were urgently needed; this collection was first encoded in the Unicode Standard, Version 6.0.

In 2012, the IRG completed work on the fifth supplement to the URO, a collection of 5,762 characters from various sources. The Extension E collection was first encoded in the Unicode Standard, Version 8.0.

In 2015, the IRG completed work on the sixth supplement to the URO, a collection of approximately 7,500 characters from various sources. The Extension F collection was first encoded in the Unicode Standard, Version 10.0.

The IRG is finishing the work on the seventh supplement to the URO. Current IRG work includes submissions from China, SAT, South Korea, TCA, the United Kingdom, and the United States. Submissions to the eighth supplement are due in late 2017.

E.3 CJK Sources

The Unicode Standard draws its unified Han character repertoire from a number of different character set standards. These standards, dictionaries and other documents are grouped into nine sources. The detailed listing of all of those sources, including bibliographic references for the various standards and other documents involved, can be found in Unicode Standard Annex #38, "Unicode Han Database (Unihan)." The primary work of unifying and ordering the characters from these sources was done by the Ideographic Rapporteur Group (IRG).

The G, T, H, M, J, K, KP, and V sources represent the characters submitted to the IRG by its member bodies. The G source consists of submissions from the People's Republic of China and Singapore. The other seven sources are the submissions from Taiwan, the Hong Kong SAR, the Macao SAR, Japan, South and North Korea, and Vietnam, respectively.

The U source represents character repertoires of three different types. First, it includes character submissions from the Unicode Technical Committee to the IRG. These were used by the IRG in the preparation of Extensions C and D. Second, corrections to IRG data sometimes leave unified ideographs without any official IRG source. Such "orphaned" ideographs are added to the U source to guarantee that each unified ideograph has at least one source listing. Finally, the U source includes ideographs from character set standards that were not submitted to the IRG by any member body, but which were used by the Unicode Technical Committee during the preparation of the initial set of Unified CJK Ideographs included in the Unicode Standard, Version 1.0.1—the set known as the URO.

Omission of Repertoire for Some Sources. In some cases, the entire ideographic repertoire of the original character set standards was *not* included in the corresponding source. Three reasons explain this decision:

1. Where the repertoires of two of the character set standards within a single source have considerable overlap, the characters in the overlap might be included only once in the source. This approach is used, for example, with GB 2312-80 and GB/T 12345-90, which have many ideographs in common. Characters in GB/T 12345-90 that are duplicates of characters in GB 2312-80 are not included in the G source.

2. Where a character set standard is based on unification rules that differ substantially from those used by the IRG, many variant characters found in the character set standard will not be included in the source. This situation is the case with CNS 11643-1992, EACC, and CCCII. It is the only case where full round-trip compatibility with the Han ideograph repertoire of the relevant character set standards is not guaranteed.

3. KS C 5601-1987 contains numerous duplicate ideographs included because they have multiple pronunciations in Korean. These multiply encoded ideographs are not included in the K source but are included in the U source. They are encoded in the CJK Compatibility Ideographs block to provide full round-trip compatibility with KS C 5601-1987 (now known as KS X 1001:2004).