2.0 Unification of the Han Characters

This chapter explains the background, repertoire, and ordering rationale for the unification of the Chinese, Japanese, and Korean ideographic characters. (For an introduction to the Han character set, see Volume I, Section 3.4 of *The Unicode Standard*, CJK Ideographs.)

Terminological note: There are several standard romanizations of the term used to refer to East Asian ideographic characters. These include *hanzi* (Chinese), *kanzi* (Japanese), *kanji* (colloquial Japanese), and *hanja* (Korean). The standard English translations for these terms are the following: Han character, Han ideographic character, East Asian ideographic character, or CJK ideographic character. For the purpose of clarity, The Unicode Standard uses some subset of the English terms when referring to these characters. The term *KangXi* is used only when referring to the dictionary on which the *Unified Repertoire and Ordering, Version 1.0* was based (see Section 2.3.) The term *Kanzi* is used in Chapter 3 in reference to a specific Japanese government publication.

2.1 Han Character Encoding Standards

A number of different Han character encoding standards are currently in use in several nations. These character collections overlap greatly in content, in the eighty percent range between most pairs of standards. The goal of Han Unification is to assign only one code point to each Han character, even though it might appear differently encoded in several of the source character sets. The most obvious benefit from this approach is a dramatic savings in character code space. The 21,000 characters in the Unicode Han set represent at least 121,000 code points in the source character sets from which the repertoire was drawn. In addition, the possibility of confusion between apparently identical characters is reduced.

Thousands of Han characters, such as 一 (one), have counterparts identical in form and similar in meaning in all Chinese, Japanese, and Korean standards; in these cases unification presents no problem. For others, though, the relationships are somewhat more complex; in these cases unification is handled according to the set of rules presented in this chapter.

2.2 Background: Unicode Han and the CJK-JRG

The notion that the Han characters can be encoded in a unified set to allow for more efficient and consistent data interchange has a long history. The Chinese Character Code for Information Interchange (CCCII) developed in Taiwan has been in use since 1980. It contains characters for use in China, Taiwan, and Japan. In 1981, Takahashi Tokutaro of Japan's National Diet Library proposed standardization of a character set for common use among East Asian countries. The East Asian Character Code (EACC) based on CCCII became an American national standard (ANSI Z39.64) in 1989. These sets were all designed primarily to meet the needs of bibliographers working with Chinese, Japanese, and Korean data.

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 Research Libraries Information Network'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 TC1/SC2/WG2 DIS 10646 (a universal multi-octet character code set, hereafter referred to as ISO 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 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 10646, which led to the first ISO *ad hoc* meeting to discuss the issue of unification, held in Beijing in October 1989. The October 1989 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 ISO *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 in that it used the principle of *KangXi* radical/stroke ordering of the characters. In order 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 below. Both proposals benefited greatly from programmatic comparisons of the two databases.

As a result of the agreement to merge the Unicode standard and ISO 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) 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, Version 2.0.*

The following discussion of the unification rules and ordering is taken largely from CJK-JRG Document 3-28, *Explanatory Notes for the Unified Ideographic CJK Characters Repertoire and Ordering, Version 1.0.*

2.3 Principles Underlying the Unified Han Character Set

Several principles were adopted by the CJK-JRG for the unification of the Han characters: unification of characters based on a set of rules, separation of characters based on the origin of the character, and sequence of characters based on their order in four source dictionaries.

Source Character Sets

The Unified Repertoire and Ordering, Version 2.0 includes all characters from the following standards:

GB 2312-80
GB 12345-90[1]
GB 8565-89, CNS 11643 (1st plane)
CNS 11643 (2nd plane)

JIS X 0208-1990
JIS X 0212-1990
KS C 5601-1989
KS C 5657-1991

[1] This includes some characters used to write Korean in China (GB 12052-89) and colloquial Cantonese characters used in Hong Kong.

In addition, some characters of the unsimplified form of GB 7589-87, the unsimplified form of GB 7590-87, CNS 11643/14th plane (including some characters used for colloquial Cantonese in Hong Kong), characters from the old Chinese telegraph code, and unique characters from ANSI Z39.64-1989 were included.

Ordering

The character order follows the index (page and position) of the dictionaries listed here with their priorities:

Priority	*Dictionary*	*City*	*Publisher*	*Version*
1	*KangXi Zidian*	Beijing	Zhonghua Bookstore, 1989	7th edition
2	*Dai Kanwa Ziten*	Tokyo	Taisyuukan Syoten, 1986	Revised edition
3	*Hanyu Da Zidian*	Chengdu	Sichuan Cishu Publishing, 1986	1st edition
4	*Dae Jaweon*	Seoul	Samseong Publishing Co. Ltd, 1988	1st edition

When a character is found in the *KangXi Zidian*, it follows the *KangXi Zidian* order. When it is not found in the *KangXi Zidian* and it is found in *Dai Kanwa Ziten*, it is given a position extrapolated from the *KangXi* position of the preceding character in *Dai Kanwa Ziten.* When it is not found in either *KangXi* or *Dai Kanwa*, *Hanyu Da Zidian* and *Dae Jaweon* dictionaries are consulted in a similar manner.

The Three-Dimensional Conceptual Model

The following figure shows the conceptual model on which the unification was done.

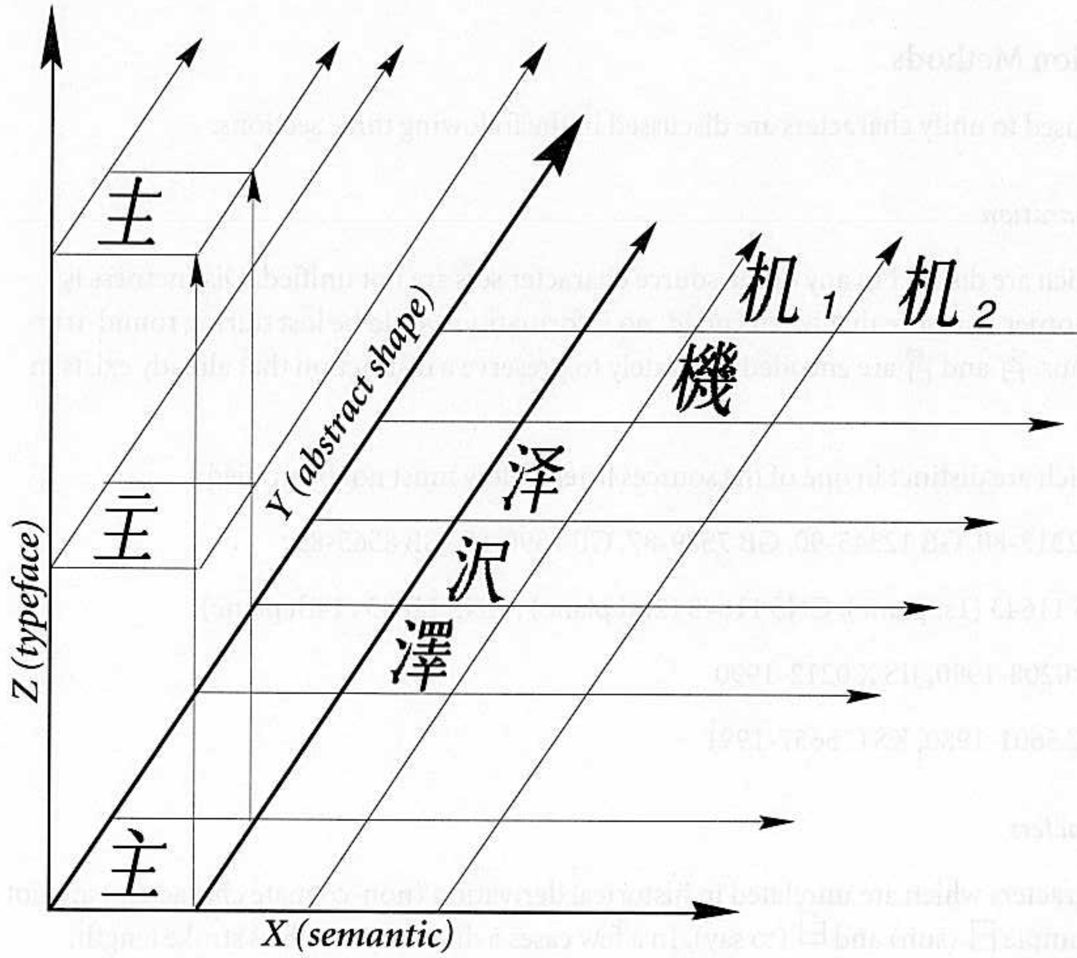

Figure 2-1. Three-Dimensional Conceptual Model

In this model, three attributes of a written form of a character are considered.

The semantic attribute (represented along the X-axis) distinguishes characters by meaning and usage. Distinctions are made between entirely unrelated characters such as 澤 (marsh) and 機 (machine) as well as extensions or borrowings beyond the original semantic cluster such as 机$_1$ (a phonetic borrowing used as a simplified form of 機) and 机$_2$ (table, the original meaning).

The abstract shape attribute (the Y-axis) distinguishes the variant forms of a single character with a single semantic attribute (that is, a character with a single position on the X axis).

The typeface attribute (the Z-axis) is for differences of type design (the actual shape used in imaging) of each variant form.

Only characters that have the same abstract shape (occupy a single point on the Y-axis) are potential candidates for unification. Typeface and semantic differences are generally ignored. The distinction between abstract shape and type-face is further discussed below.

2.4 Unification Methods

The methods used to unify characters are discussed in the following three sections.

Source Set Separation

Characters which are distinct in any of the source character sets are not unified. Distinctness is maintained in order to insure that when coded, no information would be lost during round-trip conversion. Thus, 靑 and 青 are encoded separately to preserve a distinction that already exists in CNS 11643.

Characters which are distinct in one of the sources listed below must not be unified:

G-source: GB 2312-80, GB 12345-90, GB 7589-87, GB 7590-87, GB 8565-89

T-source: CNS 11643 (1st plane), CNS 11643 (2nd plane) , CNS 11643 (14th plane)

J-source: JIS X 0208-1990, JIS X 0212-1990

K-source: KS C 5601-1980, KS C 5657-1991

Cognate Characters

In general, characters which are unrelated in historical derivation (non-cognate characters) are not unified, for example 日 (sun) and 曰 (to say). In a few cases a difference, such as stroke length, which is usually just a typeface (Z-axis) attribute, instead represents a semantic (X-axis) distinction between non-cognate characters, for example 土 (earth) and 士 (scholar, knight).

Two Level Classification of Characters

Characters are analyzed in a two-level classification. The two-level classification distinguishes characters by abstract shape (Y-axis) and actual shape of a particular typeface (Z-axis). Variant forms can be identified based on the difference of abstract shape.

Rules: Application and Examples

The two-level classification is performed with the following algorithm, which is applied to each character under comparison:

Component structure. The component structure of each character is examined. A component is a geometrical combination of primitive elements. Alternate characters can be configured with these components used in conjunction with other components. Some components can be combined to make a component more complicated in its structure. A character can be therefore defined as a component tree with the top node of the character and bottom nodes of primitive elements.

Character features. The following features of each character to be compared are examined:

- Number of components
- Relative position of components in each complete character
- Structure of a corresponding component
- Treatment in a source character set
- Radical contained in a component

Uniqueness. If one or more features are different between the characters compared, the characters are defined as having different abstract shapes and therefore are considered unique characters and are not unified.

Unification. If all the features are identical, the characters are unified.

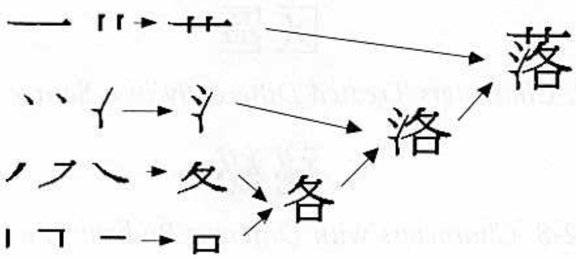

Figure 2-2. Component Structure

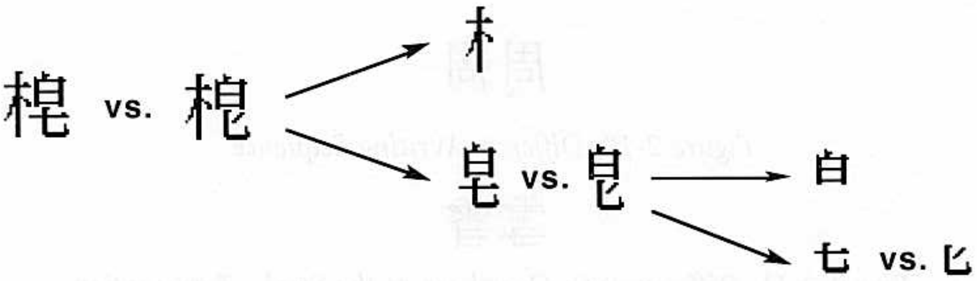

Figure 2-3. The Most Superior Node of a Component

The following examples represent some typical differences in abstract character shape. The characters are therefore not unified.

崖 厓

Figure 2-4. Different Number of Components

峰 峯

Figure 2-5. Same Number of Components Placed in Different Relative Position

拡 擴

Figure 2-6. Same Number and Same Relative Position of Components, Corresponding Components Structured Differently

区 區

Figure 2-7. Characters Treated Differently in a Source Character Set

祕 秘

Figure 2-8. Characters with Different Radical in a Component

爲 為

Figure 2-9. Same Abstract Shape, Difference in Actual Shape

Differences in actual shape of characters that *have been* unified are illustrated in Figures 2-10 through 2-17:

周 周

Figure 2-10. Different Writing Sequence

雪 雪

Figure 2-11. Differences in Overshoot at the Stroke Termination

西 西

Figure 2-12. Differences in Contact of Strokes

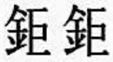

Figure 2-13. Differences in Protrusion at the Folded Corner of Strokes

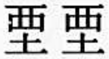

Figure 2-14. Differences in Bent Strokes

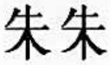

Figure 2-15. Differences in Stroke Termination

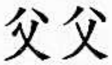

Figure 2-16. Differences in Accent at the Stroke Initiation

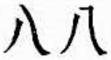

Figure 2-17. Difference in Rooftop Modification

The following figure shows characters with the same abstract shape that would have been unified *except for* the source separation rule.

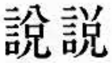

Figure 2-18. Difference in Rotated Strokes/Dots

2.5 Compatibility with Existing Standards

The compatibility of the Unicode Han character set with the repertoire of existing standards is assured by the source separation rule described above. The Unicode standard contains additional Han characters that are not included in the unified repertoire, but that do occur in widely-used corporate character sets. This practice is recognized by CJK-JRG. The following table lists all the standards that comprise the Unicode Han character set, and the number of characters included from each.

Standard	*Number of Characters*
ANSI Z39.64-1989 (EACC)	13,053
Big Five	13,481
CCCII, level 1	4,808
CNS 11643-1986	13,051
CNS 11643-1986 User Characters	3,418
GB 2312-80 (GB_0)	6,763
GB 12345-90 (GB_1)[1]	2,176
GB 7589-87 (GB_3)	7,327
GB 7590-87 (GB_5)	7,039
General Use Characters forModern Chinese (GB_7)[2]	41
GB 8565-89 (GB_8)[3]	287
GB 12052-89 (Korean)	94
JEF (Fujitsu)	3,149
JIS X 0208-1990	6,355
JIS X 0212-1990	5,801
KS C 5601-1989	4,888
KS C 5657-1991	2,856
PRC Telegraph Code	~8,000
Taiwan Telegraph Code	9,040
Xerox Chinese	9,776
Total characters covered	~121,403
Total unique characters	21,001

[1] Characters not already included in GB_0.

[2] Characters not already included in GB_0, GB_1, GB_3, GB_5, or GB_8.

[3] Characters not already included in GB_0, GB_1, GB_3, or GB_5.

Positional compatibility with the existing standards was not feasible given the incompatible ordering principles used by each. The dictionary ordering method used in this standard is an acceptable international ordering for users of Han characters. To ease the conversion of data encoded in a source character encoding to and from the Unicode standard, the Unicode consortium provides machine-readable mapping tables. (See Appendix F for information.) A full radical/stroke index is provided in Section 3.2 of this volume to help users locate characters.

Sorting Han Characters

As in Volume 1 of The Unicode Standard, Volume 2 does not define a method by which characters are sorted; the requirements for sorting differ by locale and application. Possible collating sequences include phonetic, radical-stroke (*KangXi, Xinhua Zidian*, and so on), four-corner, and total stroke count. Raw character codes alone are seldom sufficient to achieve a useable ordering in any of these schemes: ancillary data is usually required.

Character Glyphs

There may be a wide variation in the glyphs used in different countries and for different applications. The most commonly used typefaces in one country may not be used in others.

The types of glyphs used to represent characters in the Unicode Han character set have been constrained by available fonts. Users are advised to consult authoritative sources for the appropriate glyphs for individual markets and applications. It is assumed that most Unicode implementations will provide users with the ability to select the font (or mixture of fonts) that is most appropriate for a given application.