

New Ideographs in Unicode 3.0 and Beyond

John H. Jenkins

International and Text Group

Apple Computer, Inc.

1) Background

The Unicode Standard, version 2.1, contains a total of 21,204 East Asian ideographs. More than half (nearly 55%) of the encoded characters in the standard are ideographs.

This ideographic repertoire, commonly referred to as “Unihan,” is already larger than the ideographic repertoires of most other major character set standards. The exceptions, however, use different unification rules than those used in Unihan, so although they provide more glyphic variants for characters than does Unihan, they actually encode about the same number of characters as Unihan. Nonetheless, Unihan is far from being an exhaustive set of ideographs—tens of thousands more remain unencoded. As a result, additions and extensions to Unihan will continue to be made as the Unicode Standard develops.

The history of East Asian ideographs can be reliably traced back to the second millennium BCE, and all the major features of the current system were in place by the Zhou dynasty (ca. 1100 BCE). The shapes of the ideographs have altered over the centuries, and the Chinese language has continued to develop with new words coming into existence and old ones being dropped, but the writing system has endured. Chinese ideographs constitute the oldest writing system in the world still in common use.

This long history is one of the major reasons why the collection of ideographs is so vast. Roughly speaking, ideographs correspond to English words with each ideograph standing for a monosyllabic unit of meaning in a sentence.¹ New ideographs have been created, and old ones have fallen out of use.

“Misspellings” of ideographs have been deliberately made (such as to avoid using an Emperor’s name as an ordinary word), and other misspellings have been made accidentally. Doubt frequently exists as to whether some of these misspellings are deliberate and should be preserved or accidental and should be suppressed. Even now, new ideographs continue to be coined either to cover new concepts (such as newly discovered chemical elements) or by proud parents who want to give their children unique names borne by nobody else—a practice also not unheard of in the West.

Further complicating matters is the adaptation of Chinese ideographs to the use of other languages, the major instances being Japanese, Korean, and Vietnamese.

In Japan, simplified versions of the ideographs came into use to indicate the grammatical affixes Japanese requires and Chinese lacks, ultimately developing into the hiragana and katakana syllabaries used today.

Korea developed an alphabetic script, but under the influence of Chinese writing, groups the phonetic elements of the script into ideograph-like syllabic blocks. Despite efforts to minimize the use of ideographs in Korean writing, they persist in common use in South Korea.

As for Vietnamese, until this century the language was written using Chinese ideographs adapted for Vietnamese as well as new ideographs coined specifically for use with Vietnamese. It was only comparatively recently that the current Latin-based script became commonly used for Vietnamese.

¹This is a good first approximation but not entirely true. In actual practice, many words in Chinese are actually polysyllabic and a small number of ideographs can never be used as entirely independent entities. This is even more the case for languages such as Japanese and Korean which lack the basically monosyllabic, uninflected nature of Sino-Tibetan languages in general. And, of course, many ideographs have more than one distinct meaning. A good overall introduction to Chinese and the relationship between the written and spoken languages is *The Chinese Language: Fact and Fancy* by John DeFrancis (1984, University of Hawaii Press).

In all three cases, new ideographs unique to the native language were created and old ideographs came to be written in novel ways. Even in China, there have been attempts to standardize and alter the exact glyphs used. The net result is that a complete survey of the full set of ideographs ever used anywhere in East Asia over the course of the past three thousand years has never been made and is probably impossible. Common estimates as to the total count for such a collection range from about 80,000 to over 100,000.

Structurally, most ideographs can be broken down into a “radical” and a “phonetic.” The former provides a broad sense of the ideograph’s meaning, and the latter a broad sense of the ideograph’s pronunciation. The exact phonetic used depends on where and when the ideograph was coined and is also, to some extent, arbitrary. Moreover, two words with the same phonetic element will have independent histories of pronunciation. The net result is the phonetic rarely gives exact information as to how to pronounce a character.

So common is the radical/phonetic structure of ideographs that it is frequently used in dictionaries, with characters not naturally having a radical acquiring one in order to be indexed.

There is no fixed set of radicals in use throughout East Asia. By far the most common and best known, however, is that used by the definitive eighteenth century Chinese dictionary, the *KangXi*, named for the then-reigning Emperor. The *KangXi* dictionary includes a set of 214 radicals which have remained in common use ever since.

It should be noted that, on average, a literate Japanese will be able to read some 2,000 ideographs and a literate Chinese perhaps 5,000. With nearly 28,000 ideographs available, Unicode 3.0 is more than adequate to represent the majority of current East Asian texts.

2) Ideographs in the Unicode Standard

The Unicode Standard has always adopted the philosophy that the ideographic entities to be encoded are the abstract characters themselves, and not their

specific graphic representation. This means in particular that where standard writing practice varies from place to place—from mainland China, to Taiwan, to Japan, and to Korea—the various written forms will all be unified into a single character. This process is referred to as “Han unification.” “Han” is a word referring to ethnic Chinese and is used in China, Japan, and Korea to refer to ideographs—*hanzi* in Mandarin, *kanji* in Japanese, and *hanja* in Korean.

It’s important to bear in mind that Han unification is a sort of typographic ecumenism—bringing together what had once been one but had broken up. The ultimate *identity* of the ideographs is never in question; a Japanese book quoting a Chinese author may well use “Japanese glyphs” while doing so. The problem is that the common typographic practice of various East Asian countries diverged, and that this typographic practice became embodied in earlier character set standards. Unicode separates the glyphs used to draw the text from the characters used to embody the text. This is no more true for ideographs than for Latin letters.

The original work of Han unification was done in-house by the Unicode Consortium and was based on major national and industrial standards. In 1990 the Unicode Consortium began cooperating with experts from mainland China, and the work was subsequently taken over by the Chinese/Japanese/Korean Joint Research Group (CJK-JRG) a body composed of representatives from mainland China, Taiwan, Japan, Korea, and elsewhere. The CJK-JRG completed its work in 1992 and the work was adopted by the Unicode Technical Committee and included in version 1.0 of the Unicode Standard. The 20,902-ideograph set produced by the CJK-JRG is found in the CJK Unified Ideographs block of the Unicode Standard. This character set is also referred to as the Unified Repertoire and Ordering (URO), version 2.0.

However, the rules governing the production of the URO version 2.0 that were adopted by the CJK-JRG didn’t fully meet the needs of the Unicode Consortium. They didn’t include duplicate *hanja* from the Korean standard, KS C 5601-1987, or some characters from major industrial standards such as the Big Five. An

additional 302 ideographs were therefore included in the CJK Compatibility Ideographs block.

The CJK-JRG became a formal subgroup of ISO/IEC JTC1/SC2/WG2 in October 1993 and was renamed the Ideographic Rapporteur Group (IRG). The IRG remains the body carrying the burden of correlating the existing ideographic repertoire with new character standards and proposing additional characters to be encoded in order to more fully cover the needs of users.

As such, the IRG provides input to WG2 and the Unicode Technical Committee. This input is in the process of being adopted into future versions of both ISO/IEC 10646 and the Unicode Standard and may be classified as follows:

- 1) Additional mappings for existing ideographs;
- 2) Additional ideographs for encoding; and
- 3) Ancillary characters, primarily to allow some form of representation of unencoded ideographs.

In addition, research is underway into the encoding of characters in a fourth category:

- 4) Characters to provide for locale-sensitive Unihan glyph selection and provide exact information on the glyph to be used for a particular character where possible.

We will consider these groups of characters in that order.

3) Additional Mappings for Existing Ideographs

The characters in the CJK Unified Ideographs block of the Unicode Standard are formally defined to be the targets for mappings from major national standards.² Since the completion of the CJK-JRG's original work, however, new standards have been promulgated by various governments: mainland China, Taiwan, and

Korea all have new standards which need to have mappings defined. The government of Vietnam has also asked to be included in the process to provide support for historical Vietnamese literature. Japan, Singapore, and the Hong Kong Special Administrative Region have also indicated collections of characters that they feel need to be supported.³

One of the main areas of work for the IRG has been comparing these new standards against the existing URO 2.0 and defining mappings where possible. This work is generally referred to as the “horizontal extension” of the URO as no new characters are defined as part of this process.

The Unicode Standard has always included more extensive mapping tables than the *de jure* ones required by the URO 2.0 and will continue to do so. At the same time, the horizontal extension is expected to be a formal part of the next edition of ISO/IEC 10646 and will accordingly be adopted into the Unicode Standard to maintain parity between the two.

The ideographic data files supplied in conjunction with the next edition of the Unicode Standard are expected to provide a clearer distinction between the formal mappings defined by the IRG and the additional mappings suggested by the Unicode Technical Committee.

4) Additional Ideographs for Encoding

CJK Unified Ideographs Extension A

In 1998, the IRG completed work on the first “vertical extension” to the URO 2.0. This is a collection of an additional 6582 ideographs derived from IRG source sets which have been subject to all of the same unification rules as the original 20,902 with one exception: the so-called “source separation” rule is no longer applied.

² In some cases, the editions of these standards used by the CJK-JRG and IRG in their work differ from the editions generally available.

³ The United States also has a character set standard with an extensive ideograph collection, ANSI Z39.64-1989 (EACC). There has been some interest in including EACC as a source set for the IRG, and this may yet be done at some future date. Japan is currently working on a new character set standard with additional ideographs, JIS X 0213.

The source separation rule required characters that were found twice within one of the CJK-JRG's source sets to be encoded twice in the URO 2.0 even if they would otherwise be unified. A standard example of this is U+8AAA, which ordinarily would have been unified with U+8AAC had they not both been included in those portions of CNS 11643-1986 used by the CJK-JRG as a formal source.

The sources for the vertical extension are the same as those used in the horizontal extension discussed above. The set is currently targeted for encoding in the next edition of the Unicode Standard and ISO/IEC 10646 beginning at U+3400 (the CJK Unified Ideographs Extension A block).⁴

Note that this block of characters had been originally used for precomposed Korean hangul syllables in Unicode 1.1; the hangul were removed from Unicode 2.0 with the addition of the Hangul Syllables block from U+AC00 to U+D7A3. Implementations of the Unicode standard will need to distinguish clearly between versions 1.1 and 3.0 in order to avoid misinterpreting these characters.

The UniHan database released with the next edition of Unicode is expected to include data for these characters and the printed edition will have a single radical-stroke lookup chart covering both blocks of ideographs.

CJK Unified Ideographs Extension B

The IRG is currently finalizing an additional set of ideographs called "CJK Unified Ideographs B." The current draft, produced at the IRG's 13th meeting, held in Hong Kong in May 1999, consists of 40,759 ideographs from the following sources:

the KangXi dictionary and the Hanyu Da Zidian;

⁴ Nearly two-thirds of the characters currently targeted for addition in Unicode 3.0 are part of the CJK Unified Ideographs Extension A block. The net result is that over half of the Unicode Standard will continue to be taken up by ideographs. The percentage will actually increase to 56%.

the Government Chinese Character Set (GCCS) and Industry Collection, Parts A, B, and C, from the Hong Kong SAR (the H-source, with 866 characters);

PKS 5700-3:1998 from Korea (the K-source, with 169 characters);

characters from planes 4, 5, 6, 7 and 15 of CNS-11643-1992 (the T-source, with 29,795 characters); and

TCVN 5773:1993, VHN 01:1998, and VHN 02:1998 from Vietnam (the V-source).

It is anticipated that before CJK Unified Ideographs B is finalized, it will also include as a J-source new ideographs introduced in 1999 with the Japanese standard, JIS X 0213. Preliminary examinations of JIS X 0213 by the Unicode Consortium indicated several hundred ideographs in JIS X 0213 not included in the URO 2.0 or Extension A. The IRG is investigating JIS X 0213 to make sure its repertoire is fully covered by Extension B.

Notable about Extension B is its use of dictionaries as sources of ideographs. With the adoption of the CJK Unified Ideographs Extension A, the Unicode Standard will provide nearly as thorough of *existing* ideograph character set standards as is possible consistent with its unification principles. The IRG is therefore looking to additional sources to anticipate needs of users which are not met with existing character sets.

As noted, two dictionaries used as sources for Extension B are the KangXi dictionary itself and an exhaustive dictionary of modern Chinese, the *Hanyu Da Zidian*. Extending the ideograph collection of Unicode to cover the entire contents of these two dictionaries should make possible the computer encoding of virtually all extant Chinese literature.

Extension B is expected to be voted on by WG2 as a part of part 2 of ISO-IEC 10646 later in 1999. It is expected to become a formal part of the Unicode standard shortly after its approval by WG2.

Note that Extension B will be encoded using surrogates. WG2 is currently reserving an entire plane, plane 2 of UCS-4, as the CJK Unified Ideographs Supplementary Plane. This corresponds to characters encoded with the high surrogates U+D840 through U+D87F in Unicode. This 65,536 character block will be largely filled by the nearly 41,000 ideographs in Extension B.

Work on Extension B is not complete, however, and there is as yet no “part 2” of ISO/IEC 10646, so Extension B will not be included in the next editions of the Unicode Standard or of ISO/IEC 10646.

The full collection of ideographs included in the URO 2.0 and Extensions A and B are referred to as “SuperCJK” by the IRG. With the adoption of Extension B, the Unicode standard will include the full repertoire of SuperCJK and contain some 68,000 ideographs.

5) Ancillary Characters

There are three groups of these.


A) The Ideographic Variation Indicator (U+303E)


The Ideographic Variation Indicator is currently targeted for encoding in the next editions of Unicode and ISO/IEC 10646. It is a graphic character that is to be rendered visibly, and alerts the user that the intended character is similar to, but not equal to, the character that follows. Its use is similar to the existing character U+3013 GETA MARK, which is used in the printing industry to reserve space on the line for ideographic characters that cannot be rendered. A geta mark substitutes for the unknown or unavailable character but does not identify it. The Ideographic Variation Indicator is the head of a two-character sequence that gives some indication about the intended glyph or intended character, but its use disrupts the spacing on the line. Ultimately, the Ideographic Variation Character and the character following it are intended to be replaced by the correct character, once it has been encoded or identified, or a font resource or input resource has been provided for it.

The idea of substituting a similar character for an unrepresentable one is not unique to East Asia, of course. It accounts for the common “archaic” English spelling of “the” as “ye.” In this case, the “y” is actually a substitute for the letter eth, which looks similar to a “y” and had once been a part of the English alphabet. The Ideographic Variation Indicator is merely a visible marker that such is being done. It otherwise provides no information about the intended character.

To illustrate the use of this character, we use the following passage from the *Mencius* (7B14.4):⁵

犧牲既成，粢盛既潔，祭祀以時；
然而旱乾水溢，則變置社。

An older way of writing the ideograph 乾 is 𤱔, and a scholar transcribing an ancient copy of the *Mencius* may well have to deal with the ideograph 𤱔 occurring in the text. But 𤱔 isn’t encoded in Unicode 3.0⁶. Our putative scholar could use the Ideographic Variation Indicator in their text. The glyph used for the Ideographic Variation Indicator in the Unicode Standard will be . We might see, therefore, something like:

犧牲既成，粢盛既潔，祭祀以時；
然而旱乾水溢，則變置社。

in the printed edition. In this case, the Ideographic Variation Indicator would indicate that the meaning of the text is preserved, even if its exact spelling isn’t.

B. KangXi Radicals and the CJK Radicals Supplement

⁵ D.C. Lau’s English translation is: “When the sacrificial animals are sleek, the offerings are clean and the sacrifices are observed at due times, and yet floods and droughts come, then the altars should be replaced.”

⁶ It is found in plane 7 of CNS 11643-1992 and the KangXi dictionary, and so is a part of the current CJK Unified Ideographs Extension B proposal.

Also targeted for inclusion in the next edition of the Unicode Standard and ISO/IEC 10646 are two blocks of radicals: the KangXi radicals block (U+2F00 through U+2FD5), which contains the base forms for the 214 KangXi radicals, and the CJK Radicals Extension block (U+2E80 through U+2EF3), which contains a set of variant shapes taken by these radicals either when they occur as parts of characters or when used for simplified Chinese. These variant shapes are commonly found as independent and distinct characters in dictionary indices—such as the radical-stroke charts in the next edition of *The Unicode Standard*. As such, they have not been subject to the usual unification rules used for other characters in the standard.

Most radicals are ideographs in their own right. Even those that are not might be found in larger dictionaries (such as the KangXi dictionary) with their own entries. This is analogous to an English dictionary having an entry for the letter "E," even though that isn't an English word. All of the characters in the KangXi radicals block and most of those in the CJK Radicals Extension block are already encoded in the CJK Unified Ideographs block of Unicode. Compatibility mappings will be provided in future versions of the Unicode properties database.

The primary reason for including a separate block of radicals is for enhanced round-trip compatibility with CNS 11643-1992, which includes 212 of the 214 radicals separately from its encoded ideographs. There is also some feeling that it might be useful in some contexts to distinguish an ideograph used as a word and an ideograph used as a letter. (This would be analogous to making a distinction between the letter "A" and the word "A" in English.)

If it should be desirable to make a distinction between an ideograph and its radical equivalent, it might be best to use distinct typefaces to reflect this distinction. In any event, the properties of the two are very different. Ideographs are categorized as letters within the Unicode Standard; radicals are categorized as symbols. In particular, it is *never* proper to treat the radicals as equivalent to

their ideographic counterparts in such operations as searching, although they might be treated as equivalent in other operations such as collation.

Within Unicode and 10646, ideographs are given names that are generated algorithmically. This is not true for the radicals. The most commonly used KangXi radicals are frequently referred to by Western sinologists by meaning (e.g., the *water* radical, the *hand* radical). English dictionaries of Chinese frequently will include charts or tables of the KangXi radicals giving them English names. While there is no universally-used set of English names for all 214 KangXi radicals, it has proven practical to provide a set of names which would generally be understood. Accordingly, the characters in the two radicals blocks have been assigned English, and not algorithmic, names.

C. Ideographic Description Characters

This set of twelve characters, targeted for encoding at U+2FF0 through U+2FFB in the next editions of the Unicode Standard and ISO/IEC 10646, represents the most flexible solution to the perennial problem of unencoded ideographs which will be a formal part of the standard. It is, however, a solution that should be used cautiously.

The characters in the Ideographic Description block provide a mechanism for the standard interchange of data including unencoded ideographs. Unencoded ideographs can be described using these characters and encoded ideographs; the reader can then from the description create a mental picture of the ideographs so described.

This is different from a formal *encoding* of an ideograph. There is no canonical description of unencoded ideographs; there is no semantic assigned to described ideographs; there is no equivalence defined for described ideographs.

Conceptually, ideograph descriptions are more akin to the English phrase, “An e with an acute accent on it,” than to the character sequence “U+006E U+0301.”

In particular, support for the characters in the Ideographic Description block does *not* require that the rendering engine recreate the graphic appearance of the described character.

Note also that many of the ideographs which users might represent using the Ideographic Description characters will be formally encoded in future versions of the Unicode Standard (such as ones including CJK Unified Ideographs Extension B).

The Ideographic Description algorithm depends on the fact that virtually all CJK Ideographs can be broken down into smaller pieces which are themselves ideographs. The broad coverage of the ideographs already encoded in the Unicode Standard implies that the vast majority of unencoded ideographs can be represented using the Ideographic Description characters.

	U+2FF0	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT
	U+2FF1	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW
	U+2FF2	IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT
	U+2FF3	IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW
	U+2FF4	IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND
	U+2FF5	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE
	U+2FF6	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW
	U+2FF7	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT
	U+2FF8	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT
	U+2FF9	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT
	U+2FFA	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT
	U+2FFB	IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

Figure 1. The Ideographic Description Characters

The twelve Ideographic Description characters are listed in Figure 1. Each one stands at the head of an “Ideographic Description Sequence” and indicates the geometric relationship of pieces of the ideograph being described.

Ideographic Description Sequences are defined by the following grammar in Backus Naur Form:

```
IDS ::= Ideograph |
      Radical |
      BinaryDescriptionOperator IDS IDS |
      TrinaryDescriptionOperator IDS IDS IDS

BinaryDescriptionOperator ::= U+2FF0 | U+2FF1 | U+2FF4 |
      U+2FF5 | U+2FF6 | U+2FF7 | U+2FF8 | U+2FF9 | U+2FFA |
      U+2FFB

TrinaryDescriptionOperator ::= U+2FF2 | U+2FF3

Radical ::= U+2E80 | U+2E81 | ... | U+2EF2 | U+2EF3 |
      U+2F00 | U+2F01 | ... | U+2FD4 | U+2FD5

Ideograph ::= U+3400 | U+3401 | ... | U+4DB4 | U+4DB5 |
      U+4E00 | U+4E01 | ... | U+9FA4 | U+9FA5 | U+FA0E |
      U+FA0F | U+FA11 | U+FA13 | U+FA14 | U+FA1F |
      U+FA21 | U+FA23 | U+FA24 | U+FA27 | U+FA28 |
      U+FA29
```

The meaning of a sequence of characters that includes Ideographic Description characters but does not conform to the above grammar is undefined.

The operators are to be taken as indicating the relative graphic positions of the operands running from left-to-right and from top-to-bottom.

Note that non-unique compatibility ideographs (U+F900 through U+FA2D except for U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, and U+FA29) are not counted as ideographs for the purposes of the grammar.

Figure 2 illustrates the use of this grammar to provide descriptions of unencoded ideographs.

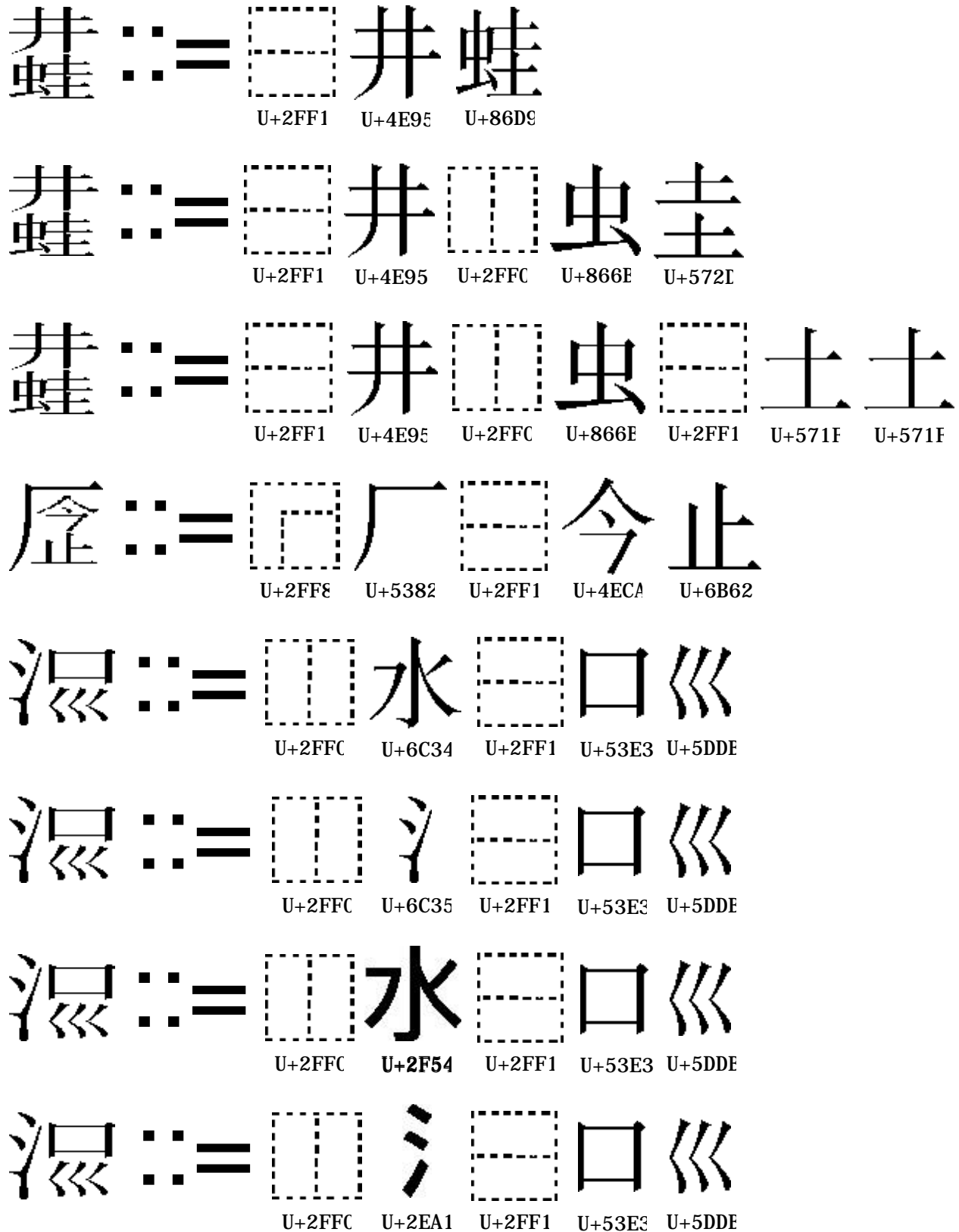


Figure 2. Using the Ideographic Description Characters

A user wishing to represent an unencoded ideograph will need to analyze its

structure to determine how to describe it using an Ideographic Description Sequence. As a rule, it is best to use the natural radical-phonetic cleavage for an ideograph if it has one and to use as short a description sequence as possible, but there is no requirement that these rules be followed.

The fact that Ideographic Description Sequences can contain other Ideographic Description Sequences means that implementations may need to be aware of the *recursion depth* of a sequence and its *back-scan length*.

The recursion depth of an Ideographic Description Sequence is the maximum number of pending operations encountered in the process of parsing an Ideographic Description Sequence. In Figure 2, the maximum recursion depth is on the third line, where three operations are still pending when you come to the end of the IDS.

The back-scan length is the maximum number ideographs unbroken by Ideographic Description Characters in the IDS. None of the examples in Figure 2 have more than two ideographs in a row; for all, the back-scan depth is two. If you access the middle of a text stream and encounter an ideograph, the back-scan length tells you how far backwards you have to go to know whether or not the ideograph is part of an Ideographic Description Sequence. In the examples, you never have to go back more than two characters.

The Unicode Standard places no limits on the recursion depth of Ideographic Description Sequences. It does, however, say that (1) Ideographic Description Sequences be as short as possible; (2) Ideographic Description Sequences must not have a back-scan depth greater than six; and (3) Ideographic Description Sequences must not be greater than sixteen characters in total length. This is to simplify the work done by Unicode implementations that parse Ideographic Description Sequences.

Many unencoded ideographs can be described in more than one way using this algorithm, either because the pieces of a description can themselves be broken down further (lines one through three in Figure 2), or because of duplications within the Unicode Standard (lines five through eight in Figure 2).

The Unicode Standard does not define equivalence for two Ideographic Description Sequences that are not identical. Although they are both valid descriptions of the same unencoded ideograph, it is not required that U+2FF1 U+4E95 U+86D9 and U+2FF1 U+4E95 U+2FF0 U+866B U+2FF1 U+571F U+571F be drawn the same or in any way treated as equivalent. Nor may it be assumed that the distinction between any of the four potential representations of the *water radical* (U+2EA1, U+2F54, U+6C34, and U+6C35) is necessarily meaningful. On the other hand, there is no guarantee that using different forms for the same radical is *not* meaningful; U+716E and U+7151 are two Unicode characters which differ *only* in the shape taken by the *fire radical* (and which mean the same thing).

In particular, ideographic Description Sequences must *not* be used to provide alternate graphic representations of encoded ideographs. Searching, collation, and other content-based text operations would then fail.

As with ideographs proper, the Ideographic Variation Indicator (U+303E) may be placed before an Ideographic Description Sequence to indicate that the description is only an approximation of the original ideograph desired. Note, however, that the use of the Ideographic Variation Indicator *inside* of a Ideographic Description Sequence is undefined.

Ideographic Description characters are visible characters. They are not to be treated as invisible control characters. The sequence U+2FF1 U+4E95 U+86D9 (with the ideographic description character at the front) must have a distinct appearance from U+4E95 U+86D9 (without it).

An implementation may render a valid Ideographic Description Sequence either by rendering the individual characters individually, or by parsing the Ideographic Description Sequence and drawing the ideograph so described. In the latter case, it may be desired to treat the Ideographic Description Sequence as a ligature of the individual characters for purposes of hit testing, cursor movement, and other user interface operations.

Other descriptions are, of course, possible for this character, such as the trivial reversal 𠄎 𠄎 𠄎 工 𠄎 日 日 日 𠄎 乞. The final character, 乞, could also itself be further described if one chose, although it is already encoded.

Note, however, that this particular description is a particularly involved one. It has a recursion depth of four and a back-scan length of six, and it includes a total of ten characters. Treating the fully expanded series as a single word may not be a good idea, as it might create an awkward line-break:

犧牲既成，粢盛既潔，祭祀以
時；然而旱
𠄎 𠄎 𠄎 𠄎 日 日 工 日 𠄎 乞 水 溢，
則變置社。

Moreover, simply inserting the sequence into text as such makes it difficult for human eyes to parse. It might be better to allow line-breaks within the sequence and use brackets or some other visual hint as to the limits of the Ideographic Description Sequence:

犧牲既成，粢盛既潔，祭祀以
時；然而旱 [𠄎 𠄎 𠄎 𠄎 日 日 工 日
𠄎 乞] 水 溢，則變置社。

It is, of course, possible with an advanced layout system like those provided by technologies such as Apple Advanced Typography, OpenType, or T_EX, to create a single ligature out of the sequence 𠄎 𠄎 𠄎 𠄎 日 日 工 日 𠄎 乞 and draw it with the glyph 𠄎. This would leave the Ideographic Description Sequence intact in the underlying text—which avoids the need of using the Private Use Area for the unencoded ideograph—but make a printed text better suited for human reading purposes:

犧牲既成，粢盛既潔，祭祀以時；
然而旱乾水溢，則變置社。

6) Unihan glyph selection and Itaiji

In the early days of the Unicode Standard, there was a considerable amount of heated debate between opponents of Han unification and its proponents. Opponents treated unification as a disaster in the making, something which would “force” Japanese readers to see their language written with the graphically unacceptable Chinese glyphs, and proponents tended to minimize this as “merely” a font problem.

The truth is somewhere in the middle. Most ideographs can be written identically anywhere inside East Asia and no reader would notice the difference between the Chinese and Japanese glyphs. Some, however, are written distinctly differently from place to place—analogous to the differences between American and British spellings of English. An Englishman would be annoyed if his computer disallowed British spellings and always Americanized them; a Japanese would be annoyed if Japanese text would occasionally include unexpected Chinese preferences for how some ideographs are drawn.

It’s possible to select appropriate Unihan glyphs algorithmically. Japanese text, for example, would include a high percentage of kana; Korean text an even higher percentage of hangul. Distinguishing simplified from traditional Chinese is more difficult, but also possible.

Still, burdening operating systems and applications with the responsibility of doing this algorithmic process is untenable, as is expecting them to rely entirely on out-of-band information. This was one of the motivations underlying the so-called “Plane 14 tag” scheme. This scheme has been approved by the Unicode Technical Committee but is not yet a formal part of the Unicode Standard, nor will it be in the next version of the Unicode Standard.

The full text of the Unicode Technical Report explaining the Plane 14 tag scheme is found at the URL <http://www.unicode.org/unicode/reports/tr7.html>. To summarize, however, this mechanism reserves 128 characters from Plane 14 of ISO/IEC 10646 (corresponding to Unicode characters with high surrogate U+DB40 and low surrogates from U+DC00 through U+DC7F). Of these characters, those with low surrogates U+DC20 through U+DC7E are to be thought of as a clone of the visual characters from ASCII.

The sequence U+DB40 U+DC01 marks the beginning of a tag. Coming next, spelled out with the ASCII clones, is a language tag as defined in RFC 1766, making use only of registered tag values or of user-defined language tags starting with the characters “x-.”

This mechanism allows a distinction to be made between Chinese (as written in mainland China), Chinese (as written in Taiwan), Japanese, Korean, and Vietnamese. A system which uses this language tagging scheme or its equivalent could readily select the best font or the best glyph within a font for display.⁷

For example, it is possible to include in a TrueType font multiple character to glyph mapping tables that can be distinguished by language. Language or locale tags could be used to select the proper table and control display of language-specific glyphs thereby.⁸

The Unihan glyph selection problem is, however, only a part of a more general problem which is usually described with the Japanese name of “itaiji.” Although in general the choice of which glyph to use for a particular character is purely typeface-dependent, there are situations where the user may wish to *force* a display of a particular glyph if at all possible. This is analogous to the rare situations where it is desirable to force a Latin “a” or “g” to be displayed in the gothic or the grotesque form. A typical example is personal names. Where an

⁷ Note that the Plane 14 scheme is designed for handling a specific problem and isn't a good general solution to the problem of language-tagging Unicode text, let alone proper locale-sensitive Unihan glyph selection.

ideograph may be written in more than one way, many individuals nonetheless prefer only one form to be used for their name.

The UTC has not adopted a specific solution to the itaiji problem. One solution has been proposed, and the UTC has indicated potential interest in that solution.

Rather like the Plane 14 tag mechanism, it involves tagging characters to define specific glyph selection, using a small set of tag characters. These tag characters are to be thought of as invisible combining marks. An ideograph followed by one of these tag characters specifies a particular glyphic form for the ideograph—if the current font has that specific glyph, it should be used; otherwise, the regular glyph for that character is used.

This mechanism would require the creation of a registration authority that would keep track of the potential glyphic variants for individual ideographs which people may wish to use. It would also require operating systems to honor the tags and font vendors to include multiple glyphs for individual characters according to the registry.

None of these problems are insurmountable, but all of them will take time. Assuming that all goes well, the mechanism could be in place not far into the next millennium.

7) Summary

The Unicode Standard, version 2.1, already provides more extensive support for Han ideographs than most other character set standards in common use. This support will be significantly extended in future versions of the standard, increasing its utility as the character set of choice to represent East Asian text.

⁸ “Language,” of course, is something of a misnomer here. The problem is a locale one, not a language one. Mandarin as spoken in Taiwan is written differently from Mandarin as spoken in mainland China, and Cantonese as spoken in Hong Kong is written the same as Mandarin as spoken in Taiwan.