Technical Reports

L2/25-206

Proposed Update Unicode® Standard Annex #38

UNICODE HAN DATABASE (UNIHAN)

Version	Unicode 17.0.0
Editors	Ken Lunde 小林劍剣 ~~Richard Cook 曲理査~~
Date	2025-05-18
This Version	https://www.unicode.org/reports/tr38/tr38-38.html
Previous Version	https://www.unicode.org/reports/tr38/tr38-37.html
Latest Version	https://www.unicode.org/reports/tr38/
Latest Proposed Update	https://www.unicode.org/reports/tr38/proposed.html
Database Lookup	https://www.unicode.org/charts/unihan.html
Revision	38

Summary

This document describes the organization and content of the Unihan database.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***A Unicode Standard Annex (UAX)** forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document corresponds to the version of the Unicode Standard of which it forms a part.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "Common References for Unicode Standard Annexes." For the latest version of the Unicode Standard, see [Unicode]. For a list of current Unicode Technical Reports, see [Reports]. For more information about versions of the Unicode Standard, see [Versions]. For any errata which may apply to this annex, see [Errata].

Contents

1 Introduction

The Unihan database is the repository for the Unicode Consortium's collective knowledge regarding the Han ideographs contained in the Unicode Standard. It contains mapping data to allow conversion to and from other coded character sets and additional information to help implement support for the various languages which use the Han script.

Formally, ideographs are defined within the Unicode Standard via their mappings. That is, the Unicode Standard does not formally define what the ideograph `U+4E00` 一 is; rather, it defines it as being the equivalent of, say, `0x523B` in GB/T 2312, `0x14421` in CNS 11643, `0x306C` in JIS X 0208, and so on.

In practice, implementation of Han ideographs requires large amounts of ancillary data. Input methods require information such as readings, as do collation algorithms. Data in character sets not included in the world of international standards bodies needs to be converted. Relationships between ideographs need to be defined to allow for fuzzy string matching. Beyond all this, it's important to track not only what properties a given ideograph has, but who claims it has those properties.

Unlike characters in Western scripts such as Latin and Greek, whose basic property is their sound, which stays largely constant across languages, the basic property for Han ideographs is their meaning. This isn't to say that ideographs are truly ideographic, in that they represent abstract ideas; but they generally have one root meaning from which the others derive, and generally retain the bulk of their semantic content across linguistic boundaries. Most ideographs are divided into a determinative, which gives a vague sense of meaning, and a phonetic, which gives a vague sense of pronunciation. The Unihan database therefore includes structural analyses and definitions for Han ideographs.

This document is a guide to that data, describing the mechanics of the Unihan database, the nature of its contents, and the status of the various properties.

2 Mechanics

2.1 Database Design

The database consists of a number of fields containing data for each Han ideograph in the Unicode Standard. The fields, all of which correspond to properties, have names that consist entirely of ASCII letters and digits with no spaces or other punctuation except for underscore. For historical reasons, they all start with a lowercase `k`.

All data in the Unihan database is stored in UTF-8 using Normalization Form C (NFC). Note, however, that the "Syntax" descriptions below, used for validation of property values, operate on Normalization Form D (NFD), primarily because that makes the regular expressions simpler.

2.1.1 Extension of Unihan Properties to Non-Unihan Characters

Some characters which are not unified ideographs are considered equivalent to unified ideographs. As such, some of the properties defined in this document are applicable to these characters as well, where appropriate. For example, U+2F8D 虫 KANGXI RADICAL INSECT is equivalent to U+866B; therefore, properties such as `kCantonese` (*cung4*), or `kCangjie` (*LMI*) may be inferred as needed for U+2F8D 虫 KANGXI RADICAL INSECT.

This extension process is particularly useful for the `kRSUnicode` and `kTotalStrokes` properties.

The `Equivalent_Unified_Ideograph` property in the Unicode Character Database [UCD] is used to indicate which non-ideographs and unified ideographs are considered equivalent for these purposes. It is explicitly intended to enable `kRSUnicode` and `kTotalStrokes` property values for non-ideographs to be derived from their equivalent unified ideographs. See Unicode Standard Annex #44, "Unicode Character Database" [UAX44], for more information.

2.1.2 Sorting Algorithm Used by the Radical-Stroke Indexes

The Unicode Standard includes a set of radical-stroke indexes for ease in determining the code point of encoded ideographs. Each Han ideograph will occur one or more times in the radical-stroke indexes, with one occurrence per value of its `kRSUnicode` property. Entries in the radical-stroke indexes are ordered using a 64-bit collation key calculated as follows:

Bits 0–19 represent the ideograph's code point. This is more space than is actually needed, but it has the advantage of aligning the code point along a four-bit boundary.

Bits 20–27 represent the ideograph's block. This block value is 0 for ideographs in the CJK Unified Ideographs block, 1 for ideographs in the CJK Unified Ideographs Extension A, 2 for ideographs in the CJK Unified Ideographs Extension B block, and so on. The special values 254 (`0xFE`) and 255 (`0xFF`) are used for ideographs in the CJK Compatibility Ideographs and CJK Compatibility Ideographs Supplement blocks, respectively. This allows accommodation for future CJK Unified Ideograph Extension blocks and guarantees that compatibility ideographs always follow unified ideographs. Note that additional compatibility ideograph blocks will not be encoded in the future.

Bits 28–31 are used to indicate whether the entry has a simplified form for the radical or not. A value of 0 indicates the traditional form for the radical (for example, U+9F8D 龍); a value of 1 indicates the Chinese simplified form of the radical (for example, U+9F99 龙); a value of 2 indicates a non-Chinese simplified form of the radical (for example, U+7ADC 竜); and a value of 3 indicates a second non-Chinese simplified form of the radical (for example, U+31DE5 𱾥).

Bits 32–35 are reserved to hold the entry's first residual stroke, as defined by the Ideographic Research Group (IRG), a subgroup of ISO/IEC JTC 1/SC 2/WG 2. Data for the first residual stroke is currently unavailable. Therefore, these bits are set to 0 in the current data.

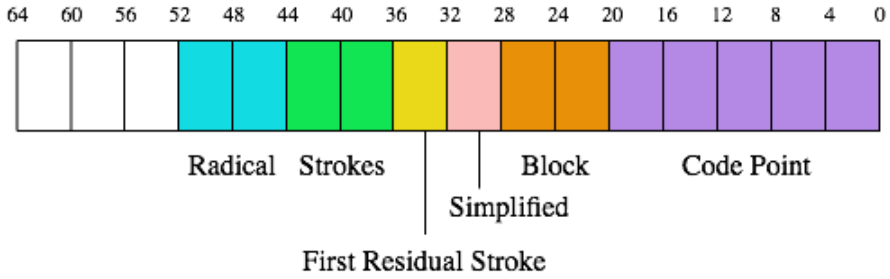
Bits 36–43 are used for the entry's residual stroke count. If the residual stroke count is negative, 0 is substituted.

Bits 44–51 are used for the entry's Kangxi radical.

Bits 52–63 are unused.

This collation key is defined in such a fashion that it can easily be parsed by eye. Figure 1 illustrates its overall structure.

Figure 1. Radical-Stroke Index Collation Key Schema



Examples:

- U+4E95 井 is assigned the collation key 0x0000702000004E95.
- U+3687 㚇 has two values in its kRSUnicode property value and therefore two entries in the radical-stroke indexes. These two entries are assigned the collation keys 0x0002306000103687 and 0x0004206000103687.
- U+F936 虜 is assigned the collation key 0x0008D0600FE0F936.
- U+21FEB 𡿫 is assigned the collation key 0x0002F03000221FEB.
- U+9F8D 龍, U+9F99 龙, U+7ADC 竜, and U+31DE5 𱷥 are assigned the collation keys 0x000D400000009F8D, 0x000D400010009F99, 0x0007505000007ADC/0x000D400020007ADC, and 0x0007506000831DE5/0x000D400030831DE5, respectively.

In addition to the radical-stroke indexes that are distributed as PDF files, the Unicode Standard also includes a “plain text” data file representation of the main radical-stroke index that includes all Han ideographs in that particular version of the standard. This data file is intended for developers who would benefit from a machine-readable implementation of the collation algorithm as applied to the current repertoire of Han ideographs, and for CJK specialists who prefer to work with data files. The format of the data file consists of the following two tab-delimited fields: 1) a unique radical-stroke value pair that is separated by a period; and 2) one or more Unicode Scalar Values for Han ideographs in collation order per this section. The actual Han ideographs that correspond to the Unicode Scalar Values in the second field are provided as comments. An example line from the data file is shown below:

```
211.0 U+9F52 U+2398A U+2398B U+2EBBD U+9F7F U+6B6F # 齒𣪊𣪋𣫽齿歯
```

2.2 Unihan.zip

Included with the [UCD] is a file called Unihan.zip. This is a snapshot of the public contents of the Unihan database as of the release date for this version of the Unicode Standard.

The zip file is an archive of eight text files, each in UTF-8, NFC, and using Unix line endings. Each file contains the values for some of the properties in the Unihan database.

Each file contains those properties which belong to one of the general categories described below; that is, Unihan_Readings.txt contains all the data for all the properties in the Readings category, and so on.

The grouping of properties into categories, and of their data into files, is based on both principled and practical considerations, and it changes over time. For mechanical parsing of Unihan data, it should not be assumed that the data for a particular property is in a particular file. One approach to parsing data for certain properties is to concatenate all of the Unihan*.txt files together (or act as if they were) and extract the desired properties from the whole (for example, using grep). This avoids the need to track which file has a given property across Unicode versions, and therefore avoids the need to adjust parsing code.

Each file uses the same structure. Blank lines may be ignored. Lines beginning with `#` are comment lines used to provide the header and footer. Each of the remaining lines is one entry, with three, tab-separated fields: the Unicode Scalar Value, the property name, and the value for the property for the given Unicode Scalar Value. For most of the properties, if multiple values are possible, the values are separated by spaces. No ideograph may have more than one instance of a given property associated with it, and no empty properties are included in any of the files archived inside `Unihan.zip`.

There is no formal limit on the lengths of any of the property values. Any Unicode characters may be used in the property values except for double quotes and control characters (especially tab, newline, and carriage return). Most properties have a more restricted syntax, such as the `kKangXi` property which consists of multiple, space-separated entries, with each entry consisting of four digits 0 through 9, followed by a period, followed by three more digits.

The data lines are sorted by Unicode Scalar Value and property-type as primary and secondary keys, respectively.

Each file's header includes a summary of the properties the file contains.

2.3 Web Access

The Unihan Database Lookup page provides interactive web access to the contents of the Unihan database. For production reasons, the version available for interactive web access may not be immediately updated to the latest available version of the `Unihan.zip` file.

Links to Chinese and Japanese compound data are presented with this web front end, such as to the online CantoDict, CC-CEDICT, and Jim Breen's WWJDIC projects. These additional data are not available in the other versions.

There are also two indices: a grid index grouping the ideographs in blocks of 256 and a radical-stroke index. A search page is also available. Individual ideographs can be accessed through the index or via the "Lookup" button and text field above. You enter the four- or five-digit hexadecimal identifier for the ideograph, and click "Lookup." You will be taken to an information page for the ideograph. The "Use text, not images" check-box allows you to control whether UTF-8 text or embedded GIFs will be used in to display ideographs. The latter technique is less dependent on your browser and system support for Unicode but is much slower.

3 Property Types

The data in the Unihan database serves a multitude of purposes, and the properties are most conveniently grouped into categories according to the purpose they fulfill. We provide here a general discussion of the various categories, followed by a detailed description of the individual properties, alphabetically arranged.

3.1 IRG Sources

Among the few normative parts of the Unihan database, and the most exhaustively checked properties, are the IRG source properties: `kIRG_GSource` (China and Singapore), `kIRG_HSource` (Hong Kong SAR), `kIRG_JSource` (Japan), `kIRG_KPSource` (North Korea), `kIRG_KSource` (South Korea), `kIRG_MSource` (Macao SAR), `kIRG_SSource` (SAT Daizōkyō Text Database Committee), `kIRG_TSource` (TCA), `kIRG_UKSource` (UK), `kIRG_USource` (UTC), and `kIRG_VSource` (Vietnam).

These represent the official mappings between Unihan and the various encoded character sets or collections which have been submitted by IRG members. The versions of these standards may differ from the published versions generally available, particularly for PRC standards. This is because in the early days of Unicode, the PRC would occasionally add ideographs to their standards on an *ad hoc* basis in order to make sure they were included. The various procedures involved in submitting ideographs to the IRG for consideration no longer make this necessary.

The values for the U-source were, in the past, only references to the Unicode Standard itself and were always equal to the ideograph's Unicode Scalar Value. This changed with the inclusion of Extension C

in Version 5.2.0 of the Unicode Standard. The values now consist of indices as described in Unicode Standard Annex #45, “U-Source Ideographs” [UAX45].

The syntax for the values used in the various IRG source properties matches that found in ISO/IEC 10646:2020 [10646].

Detailed descriptions of the syntax used are to be found in Section 4.1, *Alphabetical Listing*, below.

Note that we do not include the IRG dictionary properties in this category, largely because they are not normative parts of the standard.

The `kIICore` property is also defined by the IRG and normative.

3.2 Other Mappings

The values for the properties in this category consist of mappings to the corresponding ideographs in encoded character sets or character collections *not* used by the IRG in its unification work, although some of the character sets covered do mirror official IRG sources. For example, data for mapping GB/T 12345 is included, even though GB/T 12345 is a part of the IRG’s G-source. The difference between the two is that the `kGB1` property maps all of GB/T 12345 to Unicode, and not just that portion included in the G-source, and it doesn’t map any of the informal extensions to GB/T 12345.

3.3 Dictionary Indices

There are three main reasons for providing indices into standard dictionaries.

First, standard dictionaries provide a “paper trail” for properties such as the English gloss (`kDefinition`) and the various pronunciations or readings, as well as variant data.

Second, standard dictionaries provide a reference for scholars or students who wish more information about an ideograph.

Third, standard dictionaries are a source for unencoded ideographs. This is particularly important for Cantonese, where the Cantonese lexicon is not standardized and has been neglected by the authors and architects of previous character set encodings other than HKSCS.

Three of the dictionary properties represent official IRG indices for the dictionaries used in the four dictionary sorting algorithm. Two (`kIRGHanyuDaZidian` and `kIRGKangXi`) are still being used by the IRG, but the other one (`kIRGDaeJaweon`) is not. We have, nonetheless, retained its data for reference purposes. The property that is associated with the fourth dictionary, `kIRGDaiKanwaZiten`, was removed from the Unihan database.

The remaining dictionaries can be grouped into three categories: general-purpose Chinese (including classical Chinese), Cantonese, ~~and~~ Japanese, and Korean.

- The general-purpose Chinese dictionary properties are: `kCihaiT`, `kFennIndex`, `kGSR`, `kHanYu`, `kKangXi`, `kKarlgren`, `kMatthews`, `kSBGY`, `kSMSZD2003Index`, `kTGHZ2013`, and `kXHC1983`. These represent large, standard Chinese-Chinese and Chinese-English dictionaries, or definitive sinological studies.
- The Cantonese dictionary properties are `kCheungBauerIndex`, `kCowles`, `kLau`, and `kMeyerWempe`. All but Cheung-Bauer are large character-based Cantonese-English dictionaries. The `kSMSZD2003Index` property represents a general-purpose Chinese dictionary which also includes Cantonese data.
- ~~At present, the only Japanese property is~~ The Japanese dictionary properties are `kMorohashi` and `kNelson`~~, the ideograph’s index in the first edition of Andrew N. Nelson’s *Modern Reader’s Japanese-English Character Dictionary*~~.
- The only Korean dictionary property is `kDaeJaweon`.

3.4 Readings

We include in this category the pronunciations for a given ideograph in Mandarin, Cantonese, Tang-dynasty Chinese, Japanese, Sino-Japanese, Korean, and Vietnamese. We also include here the English gloss for a given ideograph.

Any attempt at providing a reading or set of readings for an ideograph is bound to be fraught with difficulty, because the readings will vary over time and from place to place, even within a language. Mandarin is the official language of both the PRC and Taiwan (with some differences between the two) and is the primary language over much of northern and central China, with vast differences from place to place. Even Cantonese, the modern language covered by the Unihan database with the least geographical range, is spoken throughout Guangdong Province and in much of neighboring Guangxi Zhuang Autonomous Region, and covers four large urban centers (Guangzhou, Shenzhen, Macao, and Hong Kong). There are therefore distinct regional variations in pronunciation and vocabulary.

Indeed, even the same speaker will pronounce the same word differently depending on the speaker or even the social context. This is particularly true for languages such as Cantonese, where there has been comparatively little government effort to standardize the language.

Add to this the fact that in none of these languages—the various forms of Chinese, Japanese, Korean, Vietnamese—is the syllable the fundamental unit of the language. As in the West, it's the word, and the pronunciation of an ideograph is tied to the word of which it is a part. In Chinese (followed by Vietnamese and Korean), the rule is one ideograph/one syllable, with most words written using multiple ideographs. In most cases, an ideograph has only one reading (or only one important reading), but there are numerous exceptions.

In Japanese, the situation is enormously more complex. Japanese has two pronunciation systems, one derived from Chinese (the *on* pronunciation, or Sino-Japanese), and the other from Japanese (the *kun* pronunciation).

The *on* readings derive from Chinese loan-words. They depend on factors such as when (and from which part of China) the loan-word was borrowed, and changes to Japanese since then. *On* readings can therefore have little obvious relationship to modern Chinese readings, and the same Chinese reading for a given *kanji* can be reflected in multiple *on* readings in Japanese. Contrary to Chinese practice, *on* readings may be polysyllabic.

Kun readings, on the other hand, derive from native Japanese words for which either existing *kanji* were adopted or new *kanji* coined.

The net result is that multiple readings are the rule for Japanese *kanji*. These multiple readings may bear no relationship to one another and are highly context-sensitive. Even a native Japanese reader may not know the correct pronunciation of a proper noun if it is written only in *kanji*.

Finally, some ideographs have rare pronunciations known only to a minority of native speakers, or are so rare themselves that few, if any, native speakers know how to pronounce them (for example, U+40DF 磩, which is used in a few Hong Kong place names). In many cases, the pronunciations given by professional lexicographers are little more than educated guesses.

Thus, unlike mappings between Unicode and other character sets, providing definitive data on pronunciations or, similarly, providing a definitive English gloss is impossible, and not something which has been achieved. While we make every effort to use our sources judiciously, we are aware of the fact that this data can always be improved and extended. Users should not naïvely assume that learning to pronounce an East Asian language is all about learning to pronounce the individual ideographs, or that reading is done by parsing the ideographs, one at a time.

Despite these caveats, the reading and definition data is very useful both for the student attempting to learn these languages, and for the professional attempting to use them, and so the data is included in the Unihan database.

3.5 Dictionary-like Data

This category is something of a hodge-podge, consisting of various properties including information one might find in a dictionary (such as an ideograph's *cangjie* input code), or data useful in determining levels of support (such as frequency), or structural analyses which can be helpful in lookup systems (such as the ideograph's phonetic).

As with the readings and English gloss, this data does not cover as much of Unihan as is theoretically possible, although it does cover the bulk of what is used day-to-day.

3.6 Radical-Stroke Counts

We include two radical-stroke counts for Unihan: `kRSAdobe_Japan1_6` and `kRSUnicode`.

All the radical-stroke properties are based on the radical system introduced by the 18th-century *Kangxi Dictionary* (康熙字典 *Kāngxī Zìdiǎn*). Each ideograph is assigned one of 214 radicals. In most cases, the radical assigned is the natural radical, giving a clue as to the ideograph's meaning; in the rest, the radical is arbitrary, based on the ideograph's structure. One also counts the ideograph's residual strokes, that is, the number of brush strokes required to write everything in the ideograph except the radical.

To find an ideograph using the radical-stroke system, one determines its radical and the number of residual strokes, then looks through the list of ideographs with those characteristics. This is a clumsy system compared to alphabetical lookup, but is one of the most widespread systems throughout East Asia. Unfortunately, it is also ambiguous.

First of all, if an ideograph does not have a natural radical, it can sometimes be hard to tell what the radical ought to be. For example, `U+4E95` 井 being arbitrarily assigned Radical 7 (二). Even if the ideograph naturally falls into radical-like pieces, it can be hard to tell which is the radical and which is the phonetic. For example, `U+548C` 和, which looks like it belongs to Radical 115 (禾), actually belongs to Radical 30 (口). Moreover, since Unicode encodes characters, not glyphs, two different glyphs for the same ideograph may have different residual strokes (such as `U+8005` 者, which can be written either with or without a dot, altering its stroke count between nine and eight, respectively).

The primary use for the `kRSUnicode` property is to cover the normative radical-stroke value defined by [10646]. However, it is also used for cases where there is sufficient ambiguity that a reasonable person might look for an ideograph in multiple places, particularly where one of our source dictionaries categorizes an ideograph under a different radical or with a different stroke count.

The `kRSUnicode` property also uses apostrophes after the radical number to indicate that the ideograph uses a standard simplification. A single apostrophe indicates the Chinese simplified form of the radical (for example, `U+9F99` 龙 for `U+9F8D` 龍), two apostrophes indicate a non-Chinese simplified form of the radical (for example, `U+7ADC` 竜 for `U+9F8D` 龍), and three apostrophes indicate a second non-Chinese simplified form of the radical (for example, `U+31DE5` 𱾥 for `U+9F8D` 龍).

There is, by the way, no standard way of ordering ideographs within a given radical-stroke group. The Unicode Standard's radical-stroke indexes order ideographs with the same radical-stroke count by the Unicode block in which they occur. If looking for an ideograph with Radical 64 (手) and ten residual strokes, one knows that of the hundreds of candidates in the Unicode Standard, the most common ones come towards the head of the list and the less common ones later.

The IRG is in the process of adopting a common system of assigning the first stroke of the phonetic element, more commonly referred to as the "first residual stroke," to one of five categories, and sorting by those categories. This data is now required for all IRG submissions; see Section 2.4 of Unicode Standard Annex #45, "U-Source Ideographs" [UAX45]. When this data is available for all of Unihan, it will be added to the Unihan database as a new property, and will simplify the process of finding an ideograph within a particular radical-stroke block.

3.7 Variants

Although Unicode encodes characters and not glyphs, the line between the two can sometimes be hard to draw, particularly in East Asia. There, thousands of years worth of writing have produced thousands of pairs which can be used more-or-less interchangeably.

To deal with this situation, the Unicode Standard has adopted a three-dimensional model for determining the relationship between ideographs, and has formal rules for when two forms may be unified, which includes the now-abolished Source Separation Rule. Both are described in some detail

in the Unicode Standard. Briefly, however, the three-dimensional model uses the x-axis to represent meaning, the y-axis to represent abstract shape, and the z-axis for stylistic variations.

To illustrate, U+8AAA 說 and U+8C93 貓 have different positions along the x-axis, because they mean two entirely different things (*to speak* and *cat*, respectively). U+8C93 貓 and U+732B 猫 mean the same thing and are pronounced the same way, but have different abstract shapes, so they have the same position on the x-axis (semantics), but different positions on the y-axis (abstract shape). They are said to be y-variants of one another. On the other hand, U+8AAA 說 and U+8AAC 説 have the same meaning and pronunciation, and the same abstract shape, and so have the same positions on both the x- and y-axes, but different positions on the z-axis. They are z-variants of one another.

Ideally, there would be no pairs of z-variants in the Unicode Standard; however, the need to provide for round-trip compatibility with earlier standards, and some out-and-out mistakes along the way, mean that there are some. These are marked using the kZVariant property.

The remaining variant properties are used to mark different types of y-variation.

3.7.1 Simplified and Traditional Chinese Variants

The kTraditionalVariant and kSimplifiedVariant properties are used in character-by-character conversions between simplified and traditional Chinese (abbreviated as SC and TC, respectively). For any ideograph *X*, when converting between SC and TC, there are four possible cases:

1. *X* is used in both SC and TC and is unchanged when mapping between them. An example would be U+4E95 井. This is the most common case, and is indicated by both the kSimplifiedVariant and kTraditionalVariant properties being empty.
2. *X* is used in TC but not SC, that is, it is changed when converting from TC to SC, but not vice versa. In this case, the kSimplifiedVariant property lists the ideograph(s) to which it is mapped and the kTraditionalVariant property is empty. An example would be U+66F8 書 whose kSimplifiedVariant property is U+4E66 书.
3. *X* is used in SC but not TC, that is, it is changed when converting from SC to TC, but not vice versa. In this case, the kTraditionalVariant property lists the ideograph(s) to which it is mapped and the kSimplifiedVariant property is empty. An example would be U+5B66 学 whose kTraditionalVariant property is U+5B78 學.
4. *X* is used in both SC and TC and may be changed when mapping between them. This is the most complex case, because there are two distinct sub-cases:
    1. *X* may be mapped to itself or to another ideograph when converting between SC and TC. In this case, the ideograph is its own simplification as well as the simplification for other ideographs. An example would be U+540E 后, which is the simplification for itself and for U+5F8C 後. When mapping TC to SC, it is left alone, but when mapping SC to TC it may or may not be changed, depending on context. In this case, both kTraditionalVariant and kSimplifiedVariant properties are defined and *X* is included among the values for both.
    2. *X* is used for different words in SC and TC. When converting between the two, it is always changed. An example would be U+82E7 苧. In traditional Chinese, it is pronounced *zhù* and refers to a kind of nettle. In simplified Chinese, it is pronounced *níng* and means limonene (a chemical found in the rinds of lemons and other citrus fruits). When converting TC to SC it is mapped to U+82CE 苎, and when converting SC to TC it is mapped to U+85B4 薴. In this case, both kTraditionalVariant and kSimplifiedVariant properties are defined but *X* is not included in the values for either.

In practice, conversion between simplified and traditional Chinese is complicated by several factors:

1. The conversion is almost always one-to-one, but in some cases may be one-to-many, and context may need to be evaluated to determine which specific mapping to use. When converting SC to TC, U+810F 脏 is mapped to U+81DF 臟 when it means “viscera” and to U+9AD2 髒 when it means “dirty.”

2. Simplified/traditional pairs may be affected by the now-abolished Source Separation Rule, such as 1) when the traditional variant that is common in TW/HK is not identical to the official traditional variant as defined by CN standards, such as U+4E3A 为 versus U+70BA 爲 (HK, TW) / U+7232 爲 (CN); and 2) when the traditional variant that is common in TW/HK diverges, whereby the preferred HK traditional variant is identical to the official traditional variant as defined by CN standards, such as U+8BF4 说 versus U+8AAC 說 (HK, CN) / U+8AAA 說 (TW).
3. An SC ideograph may be used in actual TC text and, more rarely, vice versa. This is particularly true in handwritten and ancient texts. Indeed, many SC forms originated as handwritten forms or ancient synonyms. It also occurs when one of a number of synonymous TC ideographs is identified as the preferred or correct ideograph to use in SC. For example, both U+732B 猫 and U+8C93 貓 are acceptable TC ideographs meaning “cat,” but only U+732B 猫 should be used in SC.
4. The mappings defined within the Unihan database are informal and based on actual practice. Different authorities may not agree on the simplified or traditional form of a specific ideograph, and either might be at odds with official, formal definitions, such as those in the 通用规范汉字表 (*Tōngyòng Guīfàn Hànzì Biǎo*, *Table of General Standard Chinese Characters*).
5. Political divisions within the Chinese-speaking community have resulted in different coinages in different locales for various modern terms, and so actual conversion between SC and TC is ideally done on a word-by-word basis, not a character-by-character basis. A hard disk, for example, is called 硬盘 in the PRC, and 硬碟 in Taiwan.

3.7.2 Semantic Variants

Two variation properties, `kSemanticVariant` and `kSpecializedSemanticVariant`, are used to mark cases where two ideographs have identical and overlapping meanings, respectively.

Thus U+514E 兎 and U+5154 兔 are y-variants of one another; both mean *rabbit*. U+4E3C 丼 and U+4E95 井 are not pure y-variants of one another. U+4E95 井 means *a well*, and although U+4E3C 丼 can also mean *a well* and be used for U+4E95 井, it can also mean *a bowl of food*. We use `kSemanticVariant`, then, for the former pair, and `kSpecializedSemanticVariant` for the latter. In many cases, data is provided listing the Unihan sources which indicate the variant relationship. The syntax is described in detail below, but as an example, U+792E 礮 has the `kSemanticVariant` value `U+70AE<kMeyerWempe U+7832<kLau,kMatthews,kMeyerWempe U+791F<kLau,kMatthews`. This means that the Mathews, Lau, and Meyer-Wempe dictionaries all say that it is a y-variant of U+7832 砲, whereas only Mathews and Lau identify it as a variant of U+791F 礮 and only Meyer-Wempe identifies it as a variant of U+70AE 炮.

3.7.3 Spoofing Variants

The `kSpoofingVariant` property is used to denote a special class of variant, a *spoofing variant*. Spoofing variants are potentially used in bad faith to direct users to unexpected URLs, evade email filters, or otherwise deceive end-users. Determining whether or not two ideographs are spoofing variants is based entirely on the glyph shape, without regard for semantics. Etymologically unrelated pairs such as U+571F 土 and U+58EB 士 or U+672A 未 and U+672B 末 are considered spoofing variants. A common source of spoofing variants is deliberate confusion between Radicals 74 (月) and 130 (肉). These two radicals, when used in Han ideographs, look very similar or identical (for example, in U+3B35 胶 and U+80F6 胶). Similarly, even if the visual appearance of two radicals is distinct, they may be similar enough that a user might overlook the distinction (for example, 氵 and 氺), especially in a spoofing context such as `https://清水.org/` versus `https://清水.org/`. Spoofing variants also include instances where two highly similar shapes are separately encoded because of source code separation, without regard to other considerations. Cases include the following pairs: U+672C 本 and U+5932 夲; U+520A 刊 and U+520B 刋.

Some spoofing variants might be sufficiently dissimilar in shape that they can be distinguished at large point sizes. Others are dissimilar in meaning so that they can be distinguished in running text. They might also be visually distinct in one font but not another, due to the language or region that the font supports. These considerations are irrelevant to their status; even dissimilar pairs can be used to misdirect users (particularly when URLs are displayed at small point sizes).

Because z-variant pairs are, by definition, either identical or unifiable, they should all be considered spoofing variants as well. The same is true of compatibility variants. Because of these considerations, the `kSpoofingVariant` property only includes spoofing variants which are *not* also z-variants or compatibility variants.

The `kSpoofingVariant` property is symmetric (if *A* is a spoofing variant of *B*, then *B* is a spoofing variant of *A*) and transitive (if *A* is a spoofing variant of *B* and *B* is a spoofing variant of *C*, then *A* is a spoofing variant of *C*).

The `kSpoofingVariant` property only covers ideographs in the CJK Unified Ideographs blocks. Other CJK-related spoofing data is found in the `EquivalentUnifiedIdeographs.txt` file in the [UCD].

3.8 Numeric Values

There are ~~five~~six properties—`kAccountingNumeric`, `kOtherNumeric`, `kPrimaryNumeric`, `kTayNumeric`, `kVietnameseNumeric`, and `kZhuangNumeric`—that indicate the numerical values an ideograph may have. Traditionally, ideographs were used both for numbers and words, and so many ideographs have (or can have) numeric values. The various kinds of numeric values are specified by these ~~five~~six properties.

3.9 Source References

A number of properties in the Unihan database indicate the source from which the data is taken. As noted above, this includes but is not limited to the `kSemanticVariant` and `kSpecializedSemanticVariant` properties.

These source references are of two kinds:

1. Sources corresponding to one or more properties in the Unihan database. These are identified by their property identifiers. In cases where multiple properties correspond to one source (such as `kHanYu` and `kHanyuPinyin`), only one is used.
2. Other sources or standard references. These are identified by a lower case `s` (for “source”) followed by a series of ASCII letters, numerals, and underscores. The overall syntax, except for the first letter, is the same as property identifiers used by the Unihan database.

Indices to or data from some of these additional sources may at some future point be added to the Unihan database as a new property. In that case, the initial `s` will be changed to a `k`.

A complete list of these additional sources with bibliographic information is found in Section 4.5 below.

3.10 IRG Source Specifiers

A number of properties in the Unihan database indicate IRG sources with which the data is associated. This includes but is not limited to the `kAlternateTotalStrokes`, `kIICore`, and `kUnihanCore2020` properties.

These sources references consist of a series of one-letter identifiers. These letters match the full IRG source designations (for example, “H” refers to `kIRG_HSource`), except that “B” is used instead of “UK” and “P” instead of “KP.” The order of the letters within the source reference is not specified.

4 The Properties

We now give two listings of the properties in the Unihan database. The first is an alphabetical listing, with information on the property contents and syntax. The second is a listing of the properties by the version of the Unicode Standard in which they were first introduced.

4.1 Alphabetical Listing

For each property we give the following information in the alphabetical listing: its *Property* tag, its Unicode *Status*, its *Category* as defined above, the Unicode version in which it was *Introduced*, its *Delimiter*, its *Syntax*, and its *Description*.

The *Property* name is the tag used in the Unihan database to mark instances of this property.

The Unicode *Status* is either *Normative*, *Informative*, or *Provisional*, depending on whether it is a normative part of the standard, an informative part of the standard, or neither. We may also include *Deprecated* as a Unicode Status if the property is no longer to be used.

Properties which allow multiple property values have a *Delimiter* defined as “space” (U+0020 SPACE). Properties which do not have multiple property values (such as the IRG source properties) have this defined as “N/A.” Some properties do not currently have multiple values in the data but may do so in the future.

For most properties with multiple values, the order of the values is arbitrary and has no particular significance. The most common order in such cases is alphabetical. For example, see the ~~kCantonese~~ kStrange property.

Review note: The kCantonese property no longer applies to the above paragraph, hence the change.

However, for certain properties the ordering of values may be significant; in such cases, the significance is specified in the Description for the property. For example, see the kMandarin property. In later versions of the [UCD], a property may change from arbitrary order to a specified order.

Validation is done as follows: The entry is split into subentries using the *Delimiter* (if defined), and each subentry converted to Normalization Form D (NFD). The value is valid if and only if each normalized subentry matches the property’s *Syntax* regular expression. Note that any given property’s *Syntax* is not guaranteed to be stable and may change in the future.

Finally, the *Description* contains not only a description of what the property contains, but also source information, known limitations, methodology used in deriving the data, and so on.

The properties covered in the table are: kAccountingNumeric, kAlternateTotalStrokes, kBigFive, kCangjie, kCantonese, kCCCII, kCheungBauer, kCheungBauerIndex, kCihaiT, kCNS1986, kCNS1992, kCompatibilityVariant, kCowles, kDaeJaweon, kDefinition, kEACC, kFanqie, kFenn, kFennIndex, kFourCornerCode, kGB0, kGB1, kGB3, kGB5, ~~kGB7,~~ kGB8, kGradeLevel, kGSR, kHangul, kHanYu, kHanyuPinlu, kHanyuPinyin, kHDZRadBreak, kHKGlyph, kIBMJapan, kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_MSource, kIRG_SSource, kIRG_TSource, kIRG_UKSource, kIRG_USource, kIRG_VSource, kIRGDaeJaweon, kIRGHanyuDaZidian, kIRGKangXi, ~~kJa,~~ kJapanese, kJapaneseKun, kJapaneseOn, kJinmeiyoKanji, kJis0, kJis1, kJIS0213, kJoyoKanji, kKangXi, kKarlgren, kKorean, kKoreanEducationHanja, kKoreanName, kLau, kMainlandTelegraph, kMandarin, kMatthews, kMeyerWempe, kMojiJoho, kMorohashi, kNelson, kOtherNumeric, kPhonetic, kPrimaryNumeric, kPseudoGB1, kRSAdobe_Japan1_6, kRSUnicode, kSBGY, kSemanticVariant, kSimplifiedVariant, kSMSZD2003Index, kSMSZD2003Readings, kSpecializedSemanticVariant, kSpoofingVariant, kStrange, kTaiwanTelegraph, kTang, kTayNumeric, kTGH, kTGHZ2013, kTotalStrokes, kTraditionalVariant, kUnihanCore2020, kVietnamese, kVietnameseNumeric, kXerox, kXHC1983, kZhuang, kZhuangNumeric, and kZVariant.

Property	**kAccountingNumeric**
Status	Informative
Category	Numeric Values
Introduced	3.2
Delimiter	N/A
Syntax	~~[0-9]~~\d+
Description	The value of the ideograph when used as an accounting numeral to prevent fraud in Chinese and derivative numeric systems. A numeral such as 十 (ten) is easily transformed into 千 (thousand) by adding a single stroke, so monetary documents often use an accounting form of the numeral, such as 拾 (ten), instead of the more common—and simpler—form. Ideographs with this property will have a single, well-defined value, which a native reader can reasonably be expected to understand.

	The three Chinese numeric-value properties should have no overlap; that is, ideographs with a kAccountingNumeric value should not have a kOtherNumeric or kPrimaryNumeric value as well.

Property	**kAlternateTotalStrokes**
Status	Provisional
Category	Dictionary-like Data
Introduced	15.0
Delimiter	space
Syntax	(\d+:[BGHJKMPSTUV]+)|-
Description	The total number of strokes in the ideograph (including the radical). Each value consists either of a decimal value followed by an IRG source specifier as defined in Section 3.10, or of the special value "-" (U+002D - HYPHEN-MINUS). The IRG source specifier indicates the IRG sources for which a particular value is preferred. ~~The source identifiers "G" and "T" are not used in this property, as these IRG sources are fully covered by the kTotalStrokes property.~~ The stroke count value is the one for the representative glyph as shown in the code charts. Multiple stroke counts are listed in increasing numeric order. Stroke counts may not be repeated. ~~If there is a single kTotalStrokes value for a ideograph, t~~The IRG sources sharing ~~this stroke count~~the kTotalStrokes value should not be explicitly listed. If all IRG sources share ~~this stroke count~~the kTotalStrokes value, then the value of "-" is used. The kAlternateTotalStrokes value for U+4E95 井 is therefore "-" instead of "4:GHJKPTV." ~~The kAlternateTotalStrokes "-" value may not be used where there are two kTotalStrokes values for an ideograph. Thus, the kAlternateTotalStrokes value for U+9AA8 骨 is "10:HJKPV."~~ For IRG sources which do not include a source reference, the kAlternateTotalStrokes property should not have a corresponding value. Unlike the kTotalStrokes property, the data in this property is not to be taken as exhaustive. Where it is defined for an ideograph, however, it includes explicit or implicit values for all IRG sources containing the deograph.

Review note: The changes to the Syntax and Description of the kAlternateTotalStrokes property are for aligning it with the change to the kTotalStrokes property that now allows only a single property value.

Property	**kBigFive**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9A-F]{4}\'?
Description	The Big Five mapping for this ideograph in hexadecimal; note that this does not cover any of the Big Five extensions in common use, including the ETEN extensions. An apostrophe (U+0027 ' APOSTROPHE) at the end of the property value indicates an alternate Big Five mapping for two ideographs that map differently in CNS 11643, specifically U+5284 劄 (Big

	Five) versus U+7B9A 箚 (CNS 11643) and U+5F5D 彝 (Big Five) versus U+5F5E 彞 (CNS 11643).

Property	**kCangjie**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1.1
Delimiter	N/A
Syntax	[A-Z]+
Description	The cangjie input code for the ideograph. This incorporates data from the file cangjie-table.b5 by Christian Wittern.

Property	**kCantonese**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	[a-z]{1,6}[1-6]
Description	The most customary jyutping (Cantonese) reading for this ideograph. This property is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep jyutping-based Han collation (and its tailorings) and transliteration reasonably stable. The values may not in all cases track the preferred reading in some dictionaries. Among the sources used for Cantonese data are the following: Casey, G. Hugh, S.J. *Ten Thousand Characters: An Analytic Dictionary*. Hong Kong: Kelley and Walsh, 1980. (kPhonetic) Cheung Kwan-hin, and Robert S. Bauer, *The Representation of Cantonese with Chinese Characters*, Journal of Chinese Linguistics Monograph Series Number 18, 2002. ISSN 0091-3723 (kCheungBauer, kCheungBauerIndex) Cowles, Roy T. *A Pocket Dictionary of Cantonese*. Hong Kong: University Press, 1999. ISBN 962-209-122-9 (kCowles) Jiu Bingcoi 饒秉才, ed. *Guangzhou Yin Zidian* / *Gwongzau Jam Zidin* 廣州音字典 (*Guangzhou Pronouncing Character Dictionary*). Hong Kong: Joint Publishing (H.K.) Co., Ltd, 1989. ISBN 962-04-0389-4 *Langwen Chuji Zhongwen Cidian* / *Longman Cokap Zungman Cidin* 朗文初級中文詞典 (*Longman's Elementary Chinese Dictionary*). Hong Kong: Longman, 2001. ISBN 962-00-5148-3 Lau, Sidney. *A Practical Cantonese-English Dictionary*. Hong Kong: Government Printer, 1977 (kLau). Meyer, Bernard F., and Theodore F. Wempe. *Student's Cantonese-English Dictionary*. Maryknoll, New York: Catholic Foreign Mission Society of America, 1947 (kMeyerWempe). Wong Gongsang 黃港生, ed. ~~*Xin*~~ *Shangwu* *Xin* *Cidian* / ~~*San*~~ *Soengmou* *San* *Cidin* 商務新詞典 (*New Commercial Press Dictionary*). Hong Kong: 商務印書館(香港)有限公司 (Commercial Press [Hong Kong], Ltd.), 1991. ISBN 962-07-0133-X

	Wong Gongsang 黃港生, ed. ~~*Xin*~~ *Shangwu* *Xin* *Zidian* / ~~*San*~~ *Soengmou* *San* *Zidin* 新商務新字典 (*New Commercial Press Character Dictionary*). Hong Kong: 商務印書館(香港)有限公司 (Commercial Press [Hong Kong], Ltd.), 2003. ISBN 962-07-0140-2 (kSMSZD2003Index and kSMSZD2003Readings) *Zhonghua Xin Zidian* / *Zungwaa San Zidin* 中華新字典 (*New Chung Hwa Character Dictionary*). Hong Kong: 中華書局 (Chung Hwa Book Co.), 2003. ISBN 962-231-001-X

Review note: The changes are to correct the titles and transliterations of two dictionaries and to reference the kSMSZD2003Readings property to the second dictionary.

Property	**kCCCII**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9A-F]{6}
Description	The mapping for this ideograph in the *Chinese Character Code for Information Interchange* 中文資訊交換碼 (CCCII) in hexadecimal, published by the Chinese Character Analysis Group 國字整理小組 (CCAG) in 1987. Earlier versions of CCCII served as the basis for ANSI/NISO Z39.64-1989 (see kEACC), so many values are common to the two properties.

Property	**kCheungBauer**
Status	Provisional
Category	Dictionary-like Data
Introduced	5.0
Delimiter	space
Syntax	~~[0-9]~~\d{3}\/~~[0-9]~~\d{2};[A-Z]*;[a-z1-6\[\]\/,]+
Description	Data regarding the ideograph in Cheung Kwan-hin and Robert S. Bauer, *The Representation of Cantonese with Chinese Characters*, Journal of Chinese Linguistics, Monograph Series Number 18, 2002. Each data value consists of three pieces, separated by semicolons: (1) the ideograph's radical-stroke index as a three-digit radical, slash, two-digit stroke count; (2) the ideograph's cangjie input code (if any); and (3) a comma-separated list of Cantonese readings using the jyutping romanization in alphabetical order.

Property	**kCheungBauerIndex**
Status	Provisional
Category	Dictionary Indices
Introduced	5.0
Delimiter	space
Syntax	~~[0-9]~~\d{3}\.[01]~~[0-9]~~\d
Description	The position of the ideograph in Cheung Kwan-hin and Robert S. Bauer, *The Representation of Cantonese with Chinese Characters*, Journal of Chinese Linguistics, Monograph Series Number 18, 2002. The format is a three-digit page number followed by a two-digit position number, separated by a period.

Property	**kCihaiT**
Status	Provisional

Category	Dictionary Indices
Introduced	3.2
Delimiter	space
Syntax	[1-9][0-9]\d{0,3}\.[0-9]\d{3}
Description	The position of this ideograph in the *Cihai* (辭海) dictionary, single volume edition, published in Hong Kong by the Zhonghua Bookstore, 1983 (reprint of the 1947 edition), ISBN 962-231-005-2. The position is indicated by a decimal number. The digits to the left of the decimal are the page number. The first digit after the decimal is the row on the page, and the remaining two digits after the decimal are the position on the row.

Property	**kCNS1986**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[12E]-[0-9A-F]{4}
Description	The CNS 11643-1986 mapping for this ideograph in hexadecimal.

Property	**kCNS1992**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[1-9]-[0-9A-F]{4}
Description	The CNS 11643-1992 mapping for this ideograph in hexadecimal.

Property	**kCompatibilityVariant**
Status	Normative
Category	IRG Sources
Introduced	3.2
Delimiter	N/A
Syntax	U\+[23]?[0-9A-F]{4}
Description	The canonical Decomposition_Mapping value for the ideograph, derived from `UnicodeData.txt` in the [UCD]. This property is derived by taking the non-null Decomposition_Mapping values from Field 5 of `UnicodeData.txt`, for ideographs contained within the CJK Compatibility Ideographs block and the CJK Compatibility Ideographs Supplement block.

Property	**kCowles**
Status	Provisional
Category	Dictionary Indices
Introduced	3.1.1
Delimiter	space
Syntax	[0-9]\d{1,4}(\.[0-9]\d{1,2})?
Description	The index or indices of this ideograph in Roy T. Cowles, *A Pocket Dictionary of Cantonese*, Hong Kong: University Press, 1999. The Cowles indices are numerical, usually integers but occasionally fractional where an ideograph was added after the original indices were determined. Cowles is missing

	indices 1222 and 4949, and four ideographs in Cowles are part of Unicode's "Hangzhou" numeral set: 2964 (U+3025), 3197 (U+3028), 3574 (U+3023), and 4720 (U+3027).

Property	**kDaeJaweon**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	N/A
Syntax	~~[0-9]~~\d{4}\.~~[0-9]~~\d{2}[01]
Description	The position of this ideograph in the *Dae Jaweon* (Korean) dictionary used in the four-dictionary sorting algorithm. The full name of this dictionary in Korean is 漢韓大辭典 大字源 (한한대사전 대자원). The position is in the form "page.position" with the final digit in the position being "0" for ideographs actually in the dictionary and "1" for ideographs not found in the dictionary and assigned a "virtual" position in the dictionary. Thus, "1187.060" indicates the sixth ideograph on page 1187. An ideograph not in this dictionary but assigned a position between the 6th and 7th ideographs on page 1187 for sorting purposes would have the code "1187.061" The edition used is the first edition, published in Seoul by Samseong Publishing Co., Ltd. (三星省出版社 삼성출판사), 1988.

Review note: The title of the dictionary was corrected.

Property	**kDefinition**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	N/A
Syntax	[^\t"]+
Description	An English definition for this ideograph. Definitions are for modern written Chinese and are usually (but not always) the same as the definition in other Chinese dialects or non-Chinese languages. In some cases, synonyms are indicated. Fuller variant information can be found using the various variant properties. Definitions specific to non-Chinese languages or Chinese dialects other than modern Mandarin are marked, for example, (Cant.) or (J). Major definitions are separated by semicolons, and minor definitions by commas. However, semicolons and commas may also occur anywhere within major definitions and minor definitions, meaning that the text cannot be parsed into separate definitions using those punctuation characters. Any valid Unicode character (except for tab, double-quote, and any line break character) may be used within the `kDefinition` property.

Property	**kEACC**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9A-F]{6}
Description	The hexadecimal code point of this ideograph in the *East Asian Character Code for Bibliographic Use* (ANSI/NISO Z39.64-1989, withdrawn in 2012). EACC is used by the

	Library of Congress for the CJK portions of MARC-8; MARC-8 itself is one of the character sets used by the Library of Congress for encoding bibliographic information. EACC's original repertoire was derived from pre-1987 versions of CCCII (see kCCCII) and is therefore identical with CCCII for many characters. The kEACC property was originally derived from data supplied and proofed by the Research Libraries Group. It has since been extended and corrected with mapping data supplied by the Library of Congress.

Property	**kFanqie**
Status	Provisional
Category	Readings
Introduced	16.0
Delimiter	space
Syntax	[\x{3400}-\x{4DBF}\x{4E00}-\x{9FFF}\x{20000}-\x{2A6DF}]{2}
Description	Fanqie (反切) is a method commonly found in ancient Chinese dictionaries and rhyming books to specify the reading of an ideograph. The method uses two ideographs to specify a reading: the first one shares the same initial part, and the second one shares the same final part. For example, 德 (***tək***) and 紅 (***ɣuŋ***) are used to indicate the Middle Chinese reading of 東 (*tuŋ*), and therefore the kFanqie property value of U+6771 東 is the ideograph pair 德紅. The method uses a third and final ideograph, which can be either 反 or 切 (the two ideographs that correspond to Fanqie), and is therefore not included as part of the kFanqie property value. Much of the property data is based on the dictionary that serves as the basis of the kSBGY property, but the scope of this property is not limited to that particular dictionary.

Property	**kFenn**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1.1
Delimiter	space
Syntax	~~[0-9]~~\d+a?[A-KP*]
Description	Data on the ideograph from *The Five Thousand Dictionary* by Courtenay H. Fenn, Cambridge, Mass.: Harvard University Press, 1979. The data here consists of a decimal number followed by a letter A through K, the letter P, or an asterisk. The decimal number gives the Soothill number for the ideograph's phonetic, and the letter is a rough frequency indication, with A indicating the 500 most common ideographs, B the next five hundred, and so on. P is used by Fenn to indicate a rare ideograph included in the dictionary only because it is the phonetic element in other ideographs. An asterisk is used instead of a letter in the final position to indicate an ideograph which belongs to one of Soothill's phonetic groups but is not found in Fenn's dictionary. Ideographs which have a frequency letter but no Soothill phonetic group are assigned group 0.

Property	**kFennIndex**
Status	Provisional
Category	Dictionary Indices
Introduced	4.1

Delimiter	space
Syntax	[0-9][0-9][0-2]\d{1,3}\.[01][0-9]\d
Description	The position of this ideograph in *The Five Thousand Dictionary* by Courtenay H. Fenn, Cambridge, Mass.: Harvard University Press, 1979. The position is indicated by a three-digit page number followed by a period and a two-digit position on the page.

Property	**kFourCornerCode**
Status	Provisional
Category	Dictionary-like Data
Introduced	5.0
Delimiter	space
Syntax	[0-9]\d{4}(\.[0-9]\d)?
Description	The four-corner code(s) for the ideograph. This data is derived from data provided in the public domain by Hartmut Bohn, Urs App, and Christian Wittern. Additional property values were provided by Jaemin Chung. The four-corner system assigns each ideograph a four-digit code from 0 through 9. The digit is derived from the “shape” of the four corners of the ideograph (upper-left, upper-right, lower-left, lower-right). An optional fifth digit can be used to further distinguish ideographs; the fifth digit is derived from the shape in the region immediately above the fourth corner. The four-corner system is now used only rarely for IMEs. It continues to be used, however, primarily for indexing and lookup in, for example, academic studies and reference material, especially in some Chinese dictionaries. Values in this property consist of four decimal digits, optionally followed by a period and fifth digit for a five-digit form.

Property	**kGB0**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]\d{4}
Description	The GB/T 2312-1980 mapping for this ideograph in row-cell form.

Property	**kGB1**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]\d{4}
Description	The GB/T 12345-1990 mapping for this ideograph in row-cell form.

Property	**kGB3**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]\d{4}

Description	The GB/T 13131 (unpublished GB/T 7589-1987 unsimplified form) mapping for this ideograph in row-cell form.

Property	**kGB5**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	~~[0-9]~~\d{4}
Description	The GB/T 13132 (unpublished GB/T 7590-1987 unsimplified form) mapping for this ideograph in row-cell form.

~~Property~~	~~**kGB7**~~
~~Status~~	~~Provisional~~
~~Category~~	~~Other Mappings~~
~~Introduced~~	~~2.0~~
~~Delimiter~~	~~N/A~~
~~Syntax~~	~~[0-9]{4}~~
~~Description~~	~~The *General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi* mapping for this ideograph in row-cell form.~~

Property	**kGB8**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	~~[0-9]~~\d{4}
Description	The GB/T 8565.2-1988 mapping for this ideograph in row-cell form.

Property	**kGradeLevel**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.2
Delimiter	N/A
Syntax	[1-6]
Description	The primary grade in the Hong Kong school system by which a student is expected to know the ideograph; this data is derived from 朗文初級中文詞典, Hong Kong: Longman, 2001.

Property	**kGSR**
Status	Provisional
Category	Dictionary Indices
Introduced	4.0.1
Delimiter	space
Syntax	~~[0-9]~~\d{4}[a-vx-z]\'?
Description	The position of this ideograph in Bernhard Karlgren's *Grammata Serica Recensa* (1957). This dataset contains a total of 7,405 records. References are given in the form `DDDDa(')`, where "DDDD" is a set number in the range `[0001..1260]` zero-padded to four digits, "a" is a letter in the range `[a..z]` (excluding "w"), optionally followed by apostrophe (`U+0027 '`

	APOSTROPHE). The data from which this mapping table is extracted contains a total of 10,023 references. References to inscriptional forms have been omitted. • Release notes: Changes since the initial release: Added: [U+25053] : 0995m (2009-01-01); Added: [U+65d6] : 0001l' (2008-11-17). 22-Dec-2003: Initial release. The following 32 references are to unencoded forms: 0059k, 0069y, 0079d, 0275b, 0286a, 0289a, 0289f, 0293a, 0325a, 0389o, 0391h, 0392s, 0468h, 0480a, 0516a, 0526o, 0566g', 0642y, 0661a, 0739i, 0775b, 0837h, 0893r, 0969a, 0969e, 1019e, 1062b, 1112d, 1124l, 1129c', 1144a, 1144b. In some cases, a variant mapping has been substituted in the mapping table, in other cases the reference is omitted. • Bibliographic information: Karlgren, Klas Bernhard Johannes 高本漢 (1889–1978): 2000. *Grammata Serica Recensa Electronica*. Electronic version of GSR, including indices, syllable canon, and images of the original Karlgren (1957) text. Prepared for the STEDT Project by Richard Cook; based in part on work by Tor Ulving and Ferenc Tafferner (see below), used by permission. Berkeley: University of California. Karlgren 1957. *Grammata Serica Recensa*. First published in the *Bulletin of the Museum of Far Eastern Antiquities* (BMFEA) No. 29, Stockholm, Sweden. Reprinted by Elanders Boktrycker Aktiebolag, Kungsbacka, [1972]. Reprinted also by SMC Publishing Inc., Taipei, Taiwan, ROC, [1996]. ISBN: 957-638-269-6. Karlgren 1940. *Grammata Serica: Script and Phonetics in Chinese and Sino-Japanese* 《中日漢字形聲論》Zhong-Ri Hanzi Xingsheng Lun [A study of Sino-Japanese semantic-phonetic compound characters:] BMFEA No. 12. Reprinted, Taipei: Ch'eng-Wen Publishing Company, [1966]. Ulving, Tor: 1997. *Dictionary of Old and Middle Chinese: Bernhard Karlgren's Grammata Serica Recensa Alphabetically Arranged*. With Ferenc Tafferner. Göteborg, Sweden: Acta Universitatis Gothoburgensis. Orientalia Gothoburgensia, 11. ISBN: 91-7346-294-2.

Property	**kHangul**
Status	Provisional
Category	Readings
Introduced	5.0
Delimiter	space
Syntax	[\x{1100}-\x{1112}][\x{1161}-\x{1175}][\x{11A8}-\x{11C2}]?:[01ENX]{1,3}
Description	The modern Korean pronunciation(s) for this ideograph in Hangul, with its source(s) following a colon. A value of 0 corresponds to KS X 1001, a value of 1 corresponds to KS X 1002, a value of E corresponds to 한문 교육용 기초 한자 (漢文教育用基礎漢字), and a value of N corresponds to 인명용 한자 (人名用漢字). A value of X indicates that a K-source was formerly at that code point but was later removed.

Property	**kHanYu**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space

Syntax	[1-8][0-9]\d{4}\.[0-3][0-9]\d[0-3]
Description	The position of this ideograph in the *Hànyǔ Dà Zìdiǎn* (HDZ) Chinese character dictionary (bibliographic information below). The ideograph references are given in the form “ABCDE.XYZ”, in which: “A” is the volume number [1..8]; “BCDE” is the zero-padded page number [0001..4809]; “XY” is the zero-padded number of the ideograph on the page [01..32]; “Z” is “0” for an ideograph actually in the dictionary, and greater than 0 for an ideograph assigned a “virtual” position in the dictionary. For example, 53024.060 indicates an actual HDZ ideograph, the 6th ideograph on Page 3,024 of Volume 5 (i.e. 籱 [U+7C49]). Note that the Volume 8 “BCDE” references are in the range [0008..0044] inclusive, referring to the pagination of the “Appendix of Addendum” at the end of that volume (beginning after p. 5746). The first ideograph assigned a given virtual position has an index ending in 1; the second assigned the same virtual position has an index ending in 2; and so on. -- Release information -- This data set contains a total of 56098 HDZ references, 54729 of which are actual HDZ ideograph references (positions are given for all HDZ head entries, including source-internal unifications), and 1369 of which are virtual ideograph positions (see note below). A total of 55,818 distinct Han ideographs are assigned mappings in this data. Because of IRG source-internal unifications, a given ideograph may have more than one HDZ reference. Source-internal unifications are of two types: (1) unifications of graphical variants; (2) unifications of duplicate head entries. The proofing of all references was done primarily on the basis of cross-checks of three versions of the reference data: (1) the original print source; (2) the `kIRGHanyuDaZidian` property of the Unihan database (release 3.1.1d1); (3) “HDZ.txt”, originally produced and proofed for Academia Sinica’s Institute of Information Technology (Document Processing Laboratory). In addition, the data was checked against the `kHanYu` and `kAlternateHanYu` properties of the Unihan database (release 3.1.1d1), which the present data set supersedes. String value, string length, compound key, field count, and page total validations were all performed. Altogether, 578 omissions/ errors in source (2) were identified/corrected. Any remaining errors will likely relate to virtual positions, or to the ordering of actual ideographs within a given page. It is unlikely that errors across page breaks remain. Possible future disunifications of source-internal unifications will necessitate update of the Unicode Scalar Value for some references. Under no circumstances should the source-internal unification (duplicate Unicode Scalar Value) mappings be removed from this data set. Note: Source (3) contributed only actual HDZ ideograph references to the proofing process, while source (2) contributed all virtual positions. It seems that the compilers of source (2) usually assigned virtual positions based on stroke count, though occasionally the virtual position brings the virtual ideograph together with the actual HDZ ideograph of which it is a variant, without regard to actual stroke count. -- Bibliographic information for the print source -- <Hanyu Da Zidian> [‘Great Chinese Character Dictionary’ (in 8 Volumes)]. XU Zhongshu (Editor in Chief). Wuhan, Hubei Province (PRC): Hubei and Sichuan Dictionary Publishing Collectives, 1986-1990. ISBN: 7-5403-0030-2/H.16. 《漢語大字典》。許力以主任，徐中舒主編，（漢語大字典工作委員會）。武漢：四川辭書出版社，湖北辭書出版社,1986-1990. ISBN: 7-5403-0030-2/H.16.

	Note that the property name is kHanYu instead of kHanyu to maintain compatibility with earlier versions of this file, where it was inappropriately spelled with an uppercase Y.

Property	**kHanyuPinlu**
Status	Provisional
Category	Readings
Introduced	4.0.1
Delimiter	space
Syntax	[a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+\([0-9]\d+\)
Description	The Pronunciations and Frequencies of this ideograph, based in part on those appearing in 《現代漢語頻率詞典》 <Xiandai Hanyu Pinlu Cidian> (XDHYPLCD) [Modern Standard Beijing Chinese Frequency Dictionary] (complete bibliographic information below). Data Format This dataset contains a total of 3799 records. (The original data provided to Unihan on 2003-02-04 contained a total of 3800 records, including U+3007 〇 IDEOGRAPHIC NUMBER ZERO, not included in the Unihan database since it is not a CJK Unified Ideograph.) Each entry is comprised of two pieces of data. The Hanyu Pinyin (HYPY) pronunciation(s) of the ideograph. Immediately following the pronunciation, a numeric string appears in parentheses: for example, in “ā(392)” the numeric string “392” indicates the sum total of the frequencies of the pronunciations of the ideograph as given in HYPLCD. Where more than one pronunciation exists, these are sorted by descending frequency, and the list elements are “space” delimited. Release Information The XDHYPLCD data here for Modern Standard Chinese (Putonghua) cuts across 4 genres (“News,” “Scientific,” “Colloquial,” and “Literature”), and was derived from a 1,807,389 ideograph corpus. See that text for additional information. The 8548 entries (8586 with variant writings) from p. 491–656 of XDHYPLCD were input by hand and proof-read from 1994-08-04 to 1995-03-22 by Richard Cook. Current Release Date above reflects date of last proofing. HYPY transcription for the data in this release was semiautomated and hand-corrected in 1995, based in part on data provided by Ross Paterson (Department of Computing, Imperial College, London). Tom Bishop is also due thanks for early assistance in proof-reading this data. The character set used for this digitization of HYPLCD (a “simplified” mainland PRC text) was (Mac OS 7-9) GB/T 2312-1980 (plus 嗐). These data were converted to Big5 (plus 腈), and both GB and Big5 versions were separately converted to Unicode 4.0, and then merged, resulting in the 3800 records in the original release. Frequency data for simplified polysyllabic words has been employed to generate both simplified and traditional ideograph frequencies. Bibliographic information for the primary print source

	《現代漢語頻率詞典》，北京語言學院語言教學研究所編著。 <Xiandai Hanyu Pinlu Cidian> = XDHYPLCD First edition 1986/6, 2nd printing 1990/4. ISBN 7-5619-0094-5/H.67.

Property	**kHanyuPinyin**
Status	Provisional
Category	Readings
Introduced	5.2
Delimiter	space
Syntax	(\d{5}\.\d{2}0,)*\d{5}\.\d{2}0:([a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+,)*[a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+
Description	The 漢語拼音 Hànyǔ Pīnyīn reading(s) appearing in the edition of 《漢語大字典》 Hànyǔ Dà Zìdiǎn (HDZ) specified in the kHanYu property description (q.v.). Each location has the form “ABCDE.XYZ” (as in kHanYu); multiple locations for a given pīnyīn reading are separated by commas. The list of locations is followed by a colon, followed by a comma-separated list of one or more pīnyīn readings. Where multiple pīnyīn readings are associated with a given mapping, these are ordered as in HDZ (for the most part reflecting relative commonality). The following are representative records. `| U+34CE | 浸 | 10297.260: qīn,qìn,qǐn | | U+34D8 | 風 | 10278.080,10278.090: sù | | U+5364 | 卤 | 10093.130: xī,lǔ 74609.020: lǔ,xī | | U+5EFE | 廾 | 10513.110,10514.010,10514.020: gǒng |` For example, the kHanyuPinyin value for U+5364 卤 is “10093.130: xī,lǔ 74609.020: lǔ,xī.” This means that U+5364 卤 is found in kHanYu at entries 10093.130 and 74609.020. The former entry has the two pīnyīn readings xī and lǔ (in that order), whereas the latter entry has the readings lǔ and xī (reversing the order). This data was originally input by 井作恆 Jǐng Zuòhéng, proofed by 聃媽歌 Dān Māgē (Magda Danish, using software donated by 文林 Wénlín Institute, Inc. and tables prepared by 曲理查 Qū Lǐchá), and proofed again and prepared for the Unicode Consortium by 曲理查 Qū Lǐchá (2008-01-14). -- Release Notes -- This data set includes readings for 34,130 distinct HDZ Hànzì, 34,302 HDZ references, and 1,457 distinct pīnyīn syllables.

Property	**kHDZRadBreak**
Status	Provisional
Category	Dictionary-like Data
Introduced	4.1
Delimiter	N/A
Syntax	[\x{2F00}-\x{2FD5}]\[U\+2F[0-9A-D][0-9A-F]\]:[1-8][0-9]\d{4}\.[0-3][0-9]\d0
Description	Indicates that 《漢語大字典》 Hanyu Da Zidian has a radical break beginning at this ideograph’s position. The property value consists of the radical (with its Unicode code point), a colon, and then the Hanyu Da Zidian position as in the kHanyu property.

Property	**kHKGlyph**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1.1
Delimiter	space
Syntax	[0-9]\d{4}

Description	The index of the ideograph in 常用字字形表 (二零零零年修訂本),香港: 香港教育學院, 2000, ISBN 962-949-040-4. This publication gives the “proper” shapes for 4759 ideographs as used in the Hong Kong school system. The index is an integer, zero-padded to four digits.

Property	**kIBMJapan**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	F[ABC][0-9A-F]{2}
Description	The IBM Japanese mapping for this ideograph in hexadecimal.

Property	**kIICore**
Status	Normative
Category	IRG Sources
Introduced	4.1
Delimiter	space
Syntax	[ABC][GHJKMPT]{1,7}
Description	Used for ideographs which are in IICore, the IRG-produced minimal set of required ideographs for East Asian use. An ideograph is in IICore if and only if it has a value for the `kIICore` property. Each value consists of a letter (A, B, or C), indicating priority value, followed by an IRG source specifier as defined in Section 3.10 above.

Property	**kIRGDaeJaweon**
Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	~~[0-9]~~\d{4}\.~~[0-9]~~\d{2}[01]
Description	The position of this ideograph in the *Dae Jaweon* (Korean) dictionary used in the four-dictionary sorting algorithm. The full name of this dictionary in Korean is 漢韓大辭典 大字源 (한한대사전 대자원). The position is in the form “page.position” with the final digit in the position being “0” for ideographs actually in the dictionary and “1” for ideographs not found in the dictionary and assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth ideograph on page 1187. An ideograph not in this dictionary but assigned a position between the 6th and 7th ideographs on page 1187 for sorting purposes would have the code “1187.061” This property represents the official position of the ideograph within the *Dae Jaweon* dictionary as used by the IRG in the four-dictionary sorting algorithm. The edition used is the first edition, published in Seoul by Samseong Publishing Co., Ltd. (三星省出版社 삼성출판사), 1988.

Review note: The title of the dictionary was corrected.

Property	**kIRGHanyuDaZidian**

Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	[1-8][0-9]\d{4}\.[0-3][0-9]\d[01]
Description	The position of this ideograph in the *Hànyǔ Dà Zìdiǎn* (PRC) dictionary used in the four-dictionary sorting algorithm. The position is in the form “volume page.position” with the final digit in the position being “0” for ideographs actually in the dictionary and “1” for ideographs not found in the dictionary and assigned a “virtual” position in the dictionary. Thus, “32264.080” indicates the eighth ideograph on page 2264 in volume 3. An ideograph not in this dictionary but assigned a position between the 8th and 9th ideographs on this page for sorting purposes would have the code “32264.081” This property represents the official position of the ideograph within the *Hànyǔ Dà Zìdiǎn* dictionary as used by the IRG in the four-dictionary sorting algorithm. The edition of the Hanyu Da Zidian used is the first edition, published in Chengdu by Sichuan Cishu Publishing, 1986.

Property	**kIRGKangXi**
Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	[01][0-9]\d{3}\.[0-7][0-9]\d[01]
Description	The official IRG position of this ideograph in the 《康熙字典》 *Kangxi Dictionary* used in the four-dictionary sorting algorithm. The position is in the form “page.position” with the final digit in the position being “0” for ideographs actually in the dictionary and “1” for ideographs not found in the dictionary but assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth ideograph on page 1187. An ideograph not in this dictionary but assigned a position between the 6th and 7th ideographs on page 1187 for sorting purposes would have the code “1187.061”. The edition of the *Kangxi Dictionary* used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989. The values in the `kIRGKangXi` property are a strict subset of those in the `kKangXi` property.

Property	**kIRG_GSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	G[013578EKS]-[0-9A-F]{4} | G4K(-\d{5})? | G(DZ|GH|RM|WZ|XC|XH|ZH)-\d{4}\.\d{2} | G(BK|CH|CY|HC)(-\d{4}\.\d{2})? | GKX-\d{4}\.\d{2,3} | G(HZ|HZR)-\d{5}\.\d{2} | G(CE|FC|IDC23|OCD|XHZ)-\d{3} | G(H|HF|LGYJ|PGLG|T|ZHSJ)-\d{4} | G(4K|CESI|CYY|DM|GT|JZ|KJ|XM|WY|ZFY|ZJW|ZYS)-\d{5} | G(FZ|IDC)-[0-9A-F]{4}

	| GCA-[A-Z]\d{4} | GGFZ-\d{6} | G(BK|LK|Z)-\d{7} | G(CH|CY|HC|U)-[023][0-9A-F]{4} | GZA-[123467]\d{5}
Description	The IRG “G” source mapping for this ideograph in hexadecimal or decimal. The IRG “G” source consists of data from the following national standards, publications, and lists from the People’s Republic of China and Singapore. The versions of the standards used are those provided by the PRC to the IRG and may not always reflect published versions of the standards generally available. G0 GB/T 2312-1980 (formerly GB 2312-80) G1 GB/T 12345-1990 (formerly GB/T 12345-90) G3 GB/T 13131 (unpublished GB/T 7589-1987 unsimplified forms) G5 GB/T 13132 (unpublished GB/T 7590-1987 unsimplified forms) G7 General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi G8 GB/T 8565.2-1988 (formerly GB 8565.2-88) GE GB/T 16500-1998 GK GB/T 12052-1989 (formerly GB 12052-89) GS Singapore Characters G4K Siku Quanshu (四庫全書) GDZ Geographic Publishing House Ideographs (地质出版社用字) GGH *Gudai Hanyu Cidian* (古代汉语词典) GGT Characters collected by the National Library of China (中国国家图书馆) GRM People’s Daily Ideographs (人民日报用字) GWY Cultural Heritage Ideographs (文化遗产用字) GWZ Hanyu Dacidian Publishing House Ideographs (漢語大詞典出版社用字) GXC *Xiandai Hanyu Cidian* (现代汉语词典) GXH *Xinhua Zidian* (新华字典) GZH *ZhongHua ZiHai* (中华字海) GZHSJ Characters collected by the Zhonghua Book Company (中华书局) GBK Chinese Encyclopedia (中國大百科全書) GCH *Ci Hai* (辞海) GCY *Ci Yuan* (辭源) GHC *Hanyu Dacidian* (漢語大詞典) GKX *Kangxi Dictionary* ideographs (康熙字典) 9th edition (1958) including the addendum (康熙字典)補遺 GHZ *Hanyu Dazidian* ideographs (漢語大字典) GHZR 汉语大字典编辑委员会:《汉语大字典（第二版）》, 武汉: 湖北长江出版集团崇文书局 & 成都 : 四川出版集团四川辞书出版社 , 2010, ISBN 978-7-5403-1744-7 GCA Culture and Art Publishing House Ideographs (文化艺术出版社用字) GCE Names of newly-discovered chemical elements as assigned by the China National Committee for Terms in Sciences and Technologies and the China National Language and Character Working Committee" (全国科学技术名词审定委员会，国家语言文字工作委员会); the value is the atomic number of the element GCESI Characters collected by China Electronics Standardization Institute (中国电子技术标准化研究院) GFC *Modern Chinese Standard Dictionary* (现代汉语规范词典第二版。主编:李行健。北京:外语 教学与研究出版社) 2010, ISBN:978-7-5600-9518-9 GIDC Supplementary Characters of the Public Security Population Information Special Font (公安人口信息专用字库补充汉字) of PRC; the value is the original PUA code point in hexadecimal GIDC23 ID system of the Ministry of Public Security of China, 2023 GOCD *Oxford English-Chinese Chinese-English Dictionary* (牛津英汉汉英词典。主

	编:Julie Kleeman, 于海江。牛津:牛津大学出版社。2010年。ISBN:978-0-19-920761-9) GXHZ *Xinhua Da Zidian* (新华大字典) GH GB/T 15564-1995 GHF 鄭賢章:《漢文佛典疑難俗字彙釋與研究》(Hànwén Fódiǎn Yínánsúzì Huìshì Yǔ Yánjiū; *Explanation and Research on Difficult Vulgar Characters in Chinese Buddhist Classics*), 成都: 巴蜀書社, 2016, ISBN 978-7-5531-0700-4 GLGYJ ZhuangLiaoSongsResearch,《壮族嘹歌研究》2008年广西民族出版社,ISBN78-7-5363-5069-4 GPGLG *Zhuang Folk Song Culture Series - Pingguo County Liao Songs* (壮族民歌文化丛书•平果嘹歌) 2004-2006, ISBN 7-5363-[4820-7 | 5012-0 | 5013-9 |5014-7 | 5015-5] GT 标准电码本（修订本） (Standard Telegraph Codebook (revised)), 1983 GCYY Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字) GJZ Commercial Press Ideographs (商务印书馆用字) GKJ Terms in Sciences and Technologies (科技用字) approved by the China National Committee for Terms in Sciences and Technologies (CNCTST) GZFY *Hanyu Fangyan Dacidian* (汉语方言大词典) GZJW Yinzhou Jinwen Jicheng Yinde (殷周金文集成引得) GZYS Chinese Ancient Ethnic Characters Research (中国民族古文字研究), 1984 GFZ Founder Press System (方正排版系统) GGFZ *Tongyong Guifan Hanzi Zidian* (通用规范汉字字典) GLK 《龍龕手鑑》 (續古逸叢書) GZ *Ancient Zhuang Character Dictionary*, (古壮字字典) 1989, ISBN 7-5363-0614-8 GXM *Characters for use in personal names in China.* Source from Public Order Administration, The Ministry of Public Security of the People's Republic of China. GZA-1 *A Vibrant and Unbroken Transmission—Filial Piety and Zhuang Funeral Songs* 生生不息的传承•孝与壮族行孝歌之研究, 2010年北京民族出版社, ISBN 978-7-1051-0648-6 GZA-2 *Annotated Long Zhuang Morality Songs* 壮族伦理道德长诗传扬歌译注, 2005 年广西民族出版社, ISBN 7-5363-4922-X GZA-3 *Compendium of Old Zhuang Folksong Texts—Wooing Songs vol. 1—Liao Songs* 壮族民歌古籍集成•情歌（一）嘹歌, 1993广西民族出版社, ISBN 7-5363-2714-5 GZA-4 *Compendium of Old Zhuang Folksong Texts—Wooing Songs vol. 1|2—Fwen Nganx* 壮族民歌古籍集成•情歌（二）欢𥍉, 1997年南宁广西民族出版社, ISBN 7-5363-3298-X GZA-6 *Zhuang Proverbs from China* 中国壮族谚语 2015 广州世界图书出版广东有限公司, ISBN 978-7-192-0492-1 GZA-7 *Ancient Remembrance—Zhuang Creation Myth Songs* 远古的追忆•壮族创世神话古歌研究, 2012 年北京民族出版社, ISBN 978-7-1051-2242-4 GDM Place name characters from the Public Order Administration, Ministry of Public Security, People's Republic of China GU The source reference for this ideograph has been moved; the value is its code point.

Property	**kIRG_HSource**
Status	Normative
Category	IRG Sources
Introduced	3.1
Delimiter	N/A
Syntax	H-[0-9A-F]{4} | H(B[012])-[0-9A-F]{4} | HD-[23]?[0-9A-F]{4} | HU-[023][0-9A-F]{4}
Description	The IRG "H" source mapping for this ideograph in hexadecimal. The IRG "H" source consists of data from the following sources: H Hong Kong Supplementary Character Set – 2008

	HB0 Big-5: Computer Chinese Glyph and Character Code Mapping Table, Technical Report C-26, 電腦用中文字型與字碼對照表, 技術通報C-26, 1984, Symbols HB1 Big-5, Level 1 HB2 Big-5, Level 2 HD Hong Kong Supplementary Character Set – 2016 HU The source reference for this ideograph has been moved; the value is its code point.

Property	**kIRG_JSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	J[014]-[0-9A-F]{4} | J3A?-[0-9A-F]{4} | J13A?-[0-9A-F]{4} | J14-[0-9A-F]{4} | JA[34]?-[0-9A-F]{4} | JARIB-[0-9A-F]{4} | JH-(JT[ABC][0-9A-F]{3}S?|IB\d{4}|\d{6}) | JK-\d{5} | JMJ-\d{6}
Description	The IRG “J” source mapping for this ideograph in hexadecimal or decimal. The IRG “J” source consists of data from the following national standards and lists from Japan. J0 JIS X 0208-1990 J1 JIS X 0212-1990 J4 JIS X 0213:2004 level-4 J3 JIS X 0213:2004 level-3 J3A JIS X 0213:2004 level-3 addendum from JIS X 0213:2000 level-3 J13 JIS X 0213:2004 level-3 ideographs replacing J1 ideographs J13A JIS X 0213:2004 level-3 ideograph addendum from JIS X 0213:2000 level-3 replacing J1 ideographs J14 JIS X 0213:2004 level-4 ideographs replacing J1 ideographs JA Unified Japanese IT Vendors Contemporary Ideographs, 1993 JA3 JIS X 0213:2004 level-3 ideographs replacing JA ideographs JA4 JIS X 0213:2004 level-4 ideographs replacing JA ideographs JARIB Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007 JH Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009 JK Japanese KOKUJI Collection JMJ Moji Joho Kiban Project (文字情報基盤整備事業)

Property	**kIRG_KPSource**
Status	Normative
Category	IRG Sources
Introduced	3.1.1
Delimiter	N/A
Syntax	KP([01]-[0-9A-F]{4}|U-[023][0-9A-F]{4})
Description	The IRG “KP” source mapping for this ideograph in hexadecimal. The IRG “KP” source consists of data from the following national standards and lists from the Democratic People’s Republic of Korea (North Korea). KP0 KPS 9566-97 KP1 KPS 10721-2000 KPU The source reference for this ideograph has been moved; the value is its code point.

	It is currently not possible to communicate with standards bodies within the DPRK. There may, therefore, be erroneous data in the values for this property. Please refer to Unicode Technical Note #50, “KP-Source Property Value History” [UTN50], for the complete history of changes that have been made to this property.

Property	**kIRG_KSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	K[0-6]-[0-9A-F]{4} | KC-\d{5} | KU-[023][0-9A-F]{4}
Description	The IRG “K” source mapping for this ideograph in hexadecimal or decimal. The IRG “K” source consists of data from the following national standards and lists from the Republic of Korea (South Korea). K0 KS X 1001:2004 (formerly KS C 5601-1987) K1 KS X 1002:2001 (formerly KS C 5657-1991) K2 KS X 1027-1:2011 (formerly PKS C 5700-1 1994) K3 KS X 1027-2:2011 (formerly PKS C 5700-2 1994) K4 KS X 1027-3:2011 (formerly PKS 5700-3:1998) K5 KS X 1027-4:2011 (formerly Korean IRG Hanja Character Set 5th Edition: 2001) K6 KS X 1027-5:2021 KC Korean History On-Line (한국 역사 정보 통합 시스템) KU The source reference for this ideograph has been moved; the value is its code point. Note that the K4 and K5 sources are expressed in hexadecimal, but unlike the K0 through K3 sources, they are not organized in row/column format. Also note that the KC source is expressed as a zero-padded five-digit decimal value.

Property	**kIRG_MSource**
Status	Normative
Category	IRG Sources
Introduced	5.2
Delimiter	N/A
Syntax	MA-[0-9A-F]{4} | MB[12]-[0-9A-F]{4} | MC-\d{5} | MDH?-[23]?[0-9A-F]{4}
Description	The IRG “M” source mapping for this ideograph in hexadecimal or decimal. The IRG “M” source corresponds to MSCS (*Macao Supplementary Character Set*). MA HKSCS-2008 code point in hexadecimal MB Big Five code point in hexadecimal MC MSCS reference MD MSCS horizontal extensions MDH HKSCS-2016 horizontal extension

Review note: The Description was updated to restore the MDH source prefix and to add *hexadecimal*.

Property	**kIRG_SSource**
Status	Normative
Category	IRG Sources
Introduced	13.0
Delimiter	N/A
Syntax	SATM?-\d{5}
Description	The IRG “S” source mapping for this ideograph in decimal ~~that corresponds to *Taishō Shinshū Daizōkyō* (大正新脩大藏經), 1924–1934, which is accessible in the SAT Daizōkyō Text Database. The source references consist of “SAT” followed by a hyphen and five decimal digits, zero padded~~. SAT Taishō Shinshū Daizōkyō (大正新脩大藏經), 1924-1934, which is accessible in the SAT Daizōkyō Text Database SATM SAT manuscript collection for Buddhist studies (IRG N2485)

Property	**kIRG_TSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	T([1-79A-F]|1[1-3])-[0-9A-F]{4} | TU-[023][0-9A-F]{4}
Description	The IRG “T” source mapping for this ideograph in hexadecimal. The IRG “T” source consists of data from the following standards and lists. “TCA” stands for “Taipei Computer Association,” and “CNS” stands for “Chinese National Standard.” T1 TCA-CNS 11643-~~1992~~1986 1st plane T2 TCA-CNS 11643-~~1992~~1986 2nd plane with one additional ideograph from TCA-CNS 11643:2007/Amd.1:2023 T3 TCA-CNS 11643-1992 3rd plane with some additional ideographs T4 TCA-CNS 11643-1992 4th plane T5 TCA-CNS 11643-1992 5th plane T6 TCA-CNS 11643-1992 6th plane T7 TCA-CNS 11643-1992 7th plane T9 TCA-CNS 11643 9th plane (pending new version) TA TCA-CNS 11643-2007 10th plane TB TCA-CNS 11643-2007 11th plane TC TCA-CNS 11643-2007 12th plane TD TCA-CNS 11643-2007 13th plane TE TCA-CNS 11643-2007 14th plane TF TCA-CNS 11643-2007 15th plane T11 TCA-CNS 11643 17th plane (pending new version) T12 TCA-CNS 11643 18th plane (pending new version) T13 TCA-CNS 11643 19th plane (pending new version) TU The source reference for this ideograph has been moved; the value is its code point. CNS 11643-1992 (p. 319) lists the following reference works: 參考文件: (1) “教育部常用國字標準字體表”， 正中書局， 民國 71 年 9 月。[‘ROC Ministry of Education: Table Standardizing Common Characters’. Sept., 1982.] (2) “教育部次常用國字標準字體表”， 教育部， 民國 71 年 12 月。[‘ROC Ministry of Education: Table Standardizing Less-Common Characters’. Dec., 1982.] (3) “教育部罕用字體表”， 正中書局， 民國 72 年 10 月。[‘ROC Ministry of Education: Table Standardizing Rare Characters’. Oct., 1983.] (4) “教育部異體國字字表”， 教育部， 民國 73 年 3 月。[‘ROC Ministry of Education: Table

	of Character Variants'. Mar., 1984.] (5) "通用漢字標準交換碼 — 使用者加字區交換碼，行政院主計處理資料中心，民國 77 年 6 月。['Standard Interchange Encoding of Common Characters — Private-Use Area Codes (Executive Office, Central Accounting Data Processing Center, ROC)'. June, 1988.] (6) 《中文大辭典》，中國文化大學出版部，民國 71 年 8 月。['Zhōng Wén Dà Cídiǎn: Encyclopedic Dictionary of Written Chinese'. Aug., 1982. 中文大辭典 (Wikipedia)] (7) 《康熙字典》，第六版，中華書局，民國 78 年 2 月。 ['Kāng Xī Dictionary'. Feb., 1989] (8) 國字標準字體研習會資料，民國 80 年 7 月。['National Script Standardization Conference Data Resources'. July, 1991.] (9) 警政署常用字頻率分析。['High-frequency characters in police reports'.] (10) 國中教科書用字整理分析報告，資訊工業策進會。['Statistical analysis of common characters in junior highschool (grades 7-9) textbooks'.] (11) "Information Technology — Universal Multi-Octet Coded Character Set (UCS), Part 1: Architecture and Basic Multi-Lingual Plane", Working Document, ISO/IEC DIS 10646 - 1.2, Dec. 26, 1991.

Review note: The description of the T1 and T2 source prefixes were updated.

Property	**kIRG_UKSource**
Status	Normative
Category	IRG Sources
Introduced	13.0
Delimiter	N/A
Syntax	UK-\d{5}
Description	The IRG "UK" source mapping for this ideograph in decimal. The source references consist of "UK" followed by a hyphen and five decimal digits, zero padded. The IRG "UK" source currently consists of data from the documents IRG N2107R2 and IRG N2232R that are available in the UK-source Ideographs repository.

Property	**kIRG_USource**
Status	Normative
Category	IRG Sources
Introduced	4.0.1
Delimiter	N/A
Syntax	UTC-\d{5}
Description	The IRG "U" source mapping for this ideograph in decimal. The IRG "U" source corresponds to the U-source ideograph database; see Unicode Standard Annex #45, "U-Source Ideographs" [UAX45]. They consist of "UTC" followed by a hyphen and a five-digit, zero-padded index into the database.

Property	**kIRG_VSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	V[0-4]-[0-9A-F]{4} | VN-[023F][0-9A-F]{4}
Description	The IRG "V" source mapping for this ideograph in hexadecimal. The IRG "V" source consists of data from the following national standards and lists from Vietnam.

	V0 TCVN 5773:1993 V1 TCVN 6056:1995 V2 VHN 01:1998 V3 VHN 02:1998 V4 *Kho Chữ Hán Nôm Mã Hoá* (Hán Nôm Coded Character Repertoire), Hà Nội, 2007 VN Vietnamese horizontal and vertical extensions; the value is its code point or its original PUA code point in the open source Nom Na Tong font

Review note: The VN source prefix is used for both horizontal and vertical extensions, and sometimes the value is not its code point, but rather the original PUA code point in the V-source font.

~~Property~~	~~**kJa**~~
~~Status~~	~~Provisional~~
~~Category~~	~~Other Mappings~~
~~Introduced~~	~~8.0.0~~
~~Delimiter~~	~~space~~
~~Syntax~~	~~[0-9A-F]{4}S?~~
~~Description~~	~~The source identifier for this ideograph in *Unified Japanese IT Vendors Contemporary Ideographs*, 1993 (JA). This property is used for ideographs whose original `kIRG_JSource` was "JA" and later changed to a different source standard.~~

Property	**kJapanese**
Status	Provisional
Category	Readings
Introduced	15.1
Delimiter	space
Syntax	[\x{3041}-\x{3096}\x{3099}\x{309A}\x{30A1}-\x{30FA}\x{30FC}]+
Description	The Japanese readings(s) for this ideograph expressed in Kana. Readings expressed in Hiragana are generally considered Kun-yomi (訓読み), and readings expressed in Katakana are generally considered On-yomi (音読み). The Moji Jōhō Kiban database and its Japanese readings are owned by CITPC (*Character Information Technology Promotion Council* 文字情報技術促進協議会), and are used under license.

Property	**kJapaneseKun**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	[A-Z]+
Description	The Japanese pronunciation(s) of this ideograph in the Hepburn romanization.

Property	**kJapaneseOn**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space

Syntax	[A-Z]+
Description	The Sino-Japanese pronunciation(s) of this ideograph.

Property	**kJinmeiyoKanji**
Status	Provisional
Category	Other Mappings
Introduced	11.0
Delimiter	space
Syntax	(20~~[0-9]~~\d{2})(:U\+[23]?[0-9A-F]{4})?
Description	The year that corresponds to the Jinmei-yō Kanji (人名用漢字) table in which the ideograph appears, and followed by a colon and the code point of its standard form if it is considered a variant. Published by Japan's Ministry of Justice (法務省) in 2010 and amended in 2015 and 2017 with one additional ideograph during each year, the Jinmei-yō Kanji table (人名用漢字表) includes 863 ideographs for use in personal names in Japan. The version year is either 2010 (861 ideographs), 2015 (one ideograph), or 2017 (one ideograph), and 230 ideographs are variants for which the code point of the standard Japanese form is specified.

Property	**kJis0**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	~~[0-9]~~\d{4}
Description	The JIS X 0208-1990 mapping for this ideograph in row-cell form.

Property	**kJIS0213**
Status	Provisional
Category	Other Mappings
Introduced	3.1.1
Delimiter	space
Syntax	[12],~~[0-9]~~\d{2},~~[0-9]~~\d{1,2}
Description	The JIS X 0213:2004 mapping for this ideograph in plane-row-cell form.

Property	**kJis1**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	~~[0-9]~~\d{4}
Description	The JIS X 0212-1990 mapping for this ideograph in row-cell form.

Property	**kJoyoKanji**
Status	Provisional
Category	Other Mappings
Introduced	11.0
Delimiter	space

Syntax	(20[0-9]\d{2})|(U\+[23]?[0-9A-F]{4})
Description	The year that corresponds to the Jōyō Kanji (常用漢字) table in which the ideograph appears, or the code point of the JIS X 0208 variant for ideographs that are specific to the JIS X 0213 standard and allowed for compatibility with implementations that support only JIS X 0208. Published by Japan's Agency for Cultural Affairs (文化庁) in 2010, the Jōyō Kanji table (常用漢字表) includes 2,136 ideographs for common use in Japan. The current version year is 2010, and there are only four ideographs that are considered JIS X 0208 variants of JIS X 0213 ideographs.

Property	**kKangXi**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[0-9]\d{4}\.[0-9]\d{2}[01]
Description	The position of this ideograph in the《康熙字典》*Kangxi Dictionary* used in the four-dictionary sorting algorithm. The position is in the form "page.position" with the final digit in the position being "0" for ideographs actually in the dictionary and "1" for ideographs not found in the dictionary but assigned a "virtual" position in the dictionary. Thus, "1187.060" indicates the sixth ideograph on page 1187. An ideograph not in this dictionary but assigned a position between the 6th and 7th ideographs on page 1187 for sorting purposes would have the code "1187.061". The edition of the *Kangxi Dictionary* used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989. The values in the `kKangXi` property are a strict superset of those in the `kIRGKangXi` property.

Property	**kKarlgren**
Status	Provisional
Category	Dictionary Indices
Introduced	3.1.1
Delimiter	space
Syntax	[1-9][0-9]\d{0,3}[A*]?
Description	The index of this ideograph in *Analytic Dictionary of Chinese and Sino-Japanese* by Bernhard Karlgren, New York: Dover Publications, Inc., 1974. If the index is followed by an asterisk (*), then the index is an interpolated one, indicating where the ideograph would be found if it were to have been included in the dictionary. Note that while the index itself is usually an integer, there are some cases where it is an integer followed by an "A."

Property	**kKorean**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	[A-Z]+

Description	The Korean pronunciation(s) of this ideograph, using the Yale romanization system. See Romanization of Korean (Wikipedia) for a discussion of the various Korean romanization systems. Use of the `kKorean` property is not recommended. The `kHangul` property, which is aligned to the KS X 1001 and KS X 1002 standards, 한문 교육용 기초 한자 (漢文教育用基礎漢字), and 인명용 한자 (人名用漢字), is recommended to be used instead.

Property	**kKoreanEducationHanja**
Status	Provisional
Category	Other Mappings
Introduced	11.0
Delimiter	space
Syntax	20[0-9]\d{2}
Description	The year that corresponds to the 한문 교육용 기초 한자 (漢文教育用基礎漢字) list of 1,800 ideographs for general use in which the ideograph appears. ~~The Supreme Court of Korea published this table of ideographs for use in personal names, and this property corresponds to an 1,800-ideograph subset that is separate from those intended only for use in personal names and covered by the `kKoreanName` property.~~This property corresponds to an 1,800-ideograph subset for educational purposes. The current version year is 2007.

Review note: The Description of this property was adjusted based on private feedback from ROK.

Property	**kKoreanName**
Status	Provisional
Category	Other Mappings
Introduced	11.0
Delimiter	space
Syntax	20[0-9]\d{2}
Description	The year that corresponds to the 인명용 한자 (人名用漢字) list in which the ideograph first appears, regardless of its readings. The Supreme Court of Korea publishe~~d~~s this table of ideographs~~, and this property excludes 1,800 ideographs that represent a subset that the `kKoreanEducationHanja` property covers~~.Since all 1,800 `kKoreanEducationHanja` ideographs are automatically included in the `kKoreanName` ideographs, the `kKoreanName` property is not shown for the 1,800 `kKoreanEducationHanja` ideographs. The current version years are 2015 and 2018. Note that 40 ideographs were added to this table in 2022, and that 1,070 were added in 2024, but we do not have the data for updating this property accordingly.

Review note: The Description of this property was adjusted based on private feedback from ROK.

Property	**kLau**
Status	Provisional

Category	Dictionary Indices
Introduced	3.1.1
Delimiter	space
Syntax	[1-9][0-9]\d{0,3}
Description	The index of this ideograph in *A Practical Cantonese-English Dictionary* by Sidney Lau, Hong Kong: The Government Printer, 1977. The index consists of an integer. Missing indices indicate ideographs to be found in Unicode Standard Annex #45, “U-Source Ideographs” [UAX45].

Property	**kMainlandTelegraph**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]\d{4}
Description	The PRC telegraph code for this ideograph, derived from 漢字電報コード変換表 (*Chinese character telegraph code conversion table*), Lin Jinyi, KDD Engineering and Consulting, Tokyo, 1984.

Property	**kMandarin**
Status	Informative
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	[a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+
Description	The most customary pīnyīn reading for this ideograph. When there are two values, then the first is preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both. This property is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep pīnyīn-based Han collation (and its tailorings) and transliteration reasonably stable. The values may not in all cases track the preferred use in some dictionaries.

Property	**kMatthews**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]\d{0,3}(a|\.5)?
Description	The index of this ideograph in *Chinese-English Dictionary* by Robert H. Mathews, Cambridge: Harvard University Press, 1975. Note that the property name is `kMatthews` instead of `kMathews` to maintain compatibility with earlier versions of this file, where it was inadvertently misspelled.

Property	**kMeyerWempe**
Status	Provisional
Category	Dictionary Indices
Introduced	3.1

Delimiter	space
Syntax	[1-9][0-9]\d{0,3}[a-t*]?
Description	The index of this ideograph in the *Student's Cantonese-English Dictionary* by Bernard F. Meyer and Theodore F. Wempe (3rd edition, 1947). The index is an integer, optionally followed by a lowercase Latin letter if the listing is in a subsidiary entry and not a main one. In some cases, where the ideograph is found in the radical-stroke index, but not in the main body of the dictionary, the integer is followed by an asterisk: for example, U+50E5 僥, which is listed as 736* as well as 1185a.

Property	**kMojiJoho**
Status	Provisional
Category	Dictionary-like Data
Introduced	15.1
Delimiter	space
Syntax	MJ\d{6}(:(FE0[01]|E01[01][0-9A-F]))?
Description	This property provides mappings from CJK Unified Ideographs, along with SVSes (*Standardized Variation Sequences*) and registered Moji_Joho IVSes (*Ideographic Variation Sequences*) that use the CJK Unified Ideograph as a BC (*Base Character*), to Moji Jōhō Kiban database (文字情報基盤データベース) serial numbers. The property is based on Version 006.01 of the Moji Jōhō Kiban database. See MJ文字情報一覧表. If a colon (:) and VS (*Variation Selector*) follow a Moji Jōhō Kiban database serial number, the sequence of the CJK Unified Ideograph, serving as a BC, followed by the VS, corresponds to the Moji Jōhō Kiban database serial number. Such sequences are SVSes or Moji_Joho IVSes. If a Moji Jōhō Kiban database serial number appears both by itself and followed by a colon and VS, the registered Moji_Joho IVS that corresponds to the latter is considered the default (that is, encoded) form. The Moji Jōhō Kiban database and its mappings are owned by CITPC (*Character Information Technology Promotion Council* 文字情報技術促進協議会), and are used under license.

Property	**kMorohashi**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	(\d{5}\'{0,2}|H\d{3})(:(FE0[01]|E010[0-9A-F]))?
Description	The index of the ideograph in the *Dai Kanwa Jiten* (大漢和辞典) Japanese kanji dictionary (1984–1986, 大修館書店)—often referred to as Morohashi (諸橋), the family name of its chief editor—or in the Dai Kanwa Jiten Hokan (大漢和辞典 補巻) supplemental volume (2000, 大修館書店). Index numbers are five zero-padded integer values with an optional single apostrophe (U+0027 ' APOSTROPHE) or double apostrophe ('') suffix that corresponds in appearance to a prime or double prime. Index numbers that appear in the supplemental volume (補巻) are prefixed with "H" and consist of three zero-padded integer values. If a colon (:) and VS (*Variation Selector*) follow an index number, the sequence of the CJK Unified Ideograph, serving as a BC (*Base Character*), followed by the VS, corresponds to the index number. Such sequences are SVSes or Moji_Joho IVSes.

	If an index number appears both by itself and followed by a colon and VS, the registered Moji_Joho IVS that corresponds to the latter is considered the default (that is, encoded) form of the CJK Unified Ideograph. The Moji Jōhō Kiban database and its mappings are owned by CITPC (*Character Information Technology Promotion Council* 文字情報技術促進協議会), and are used under license.

Property	**kNelson**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[0-9]\d{4}
Description	The index of this ideograph in *The Modern Reader's Japanese-English Character Dictionary* by Andrew Nathaniel Nelson, Rutland, Vermont: Charles E. Tuttle Company, 1974.

Property	**kOtherNumeric**
Status	Informative
Category	Numeric Values
Introduced	3.2
Delimiter	space
Syntax	[0-9]\d+
Description	One or more values of the ideograph when used as a numeral in Chinese and derivative numeric systems. Ideographs with this property are rarely used, obsolete, domain-specific, non-standard, or non-compositional as a numeral. For example, 王 is a rare ideograph whose meaning, "five," would not be recognized by most native readers; and 幺 "tiny," normally not a numeral, can be used as the phonetic code for "one" in some regions. An English-language equivalent is "gross," whose numeric value, "one hundred forty-four," is not universally understood by native readers. The three Chinese numeric-value properties should have no overlap; that is, ideographs with a `kOtherNumeric` value should not have a `kAccountingNumeric` or `kPrimaryNumeric` value as well.

Property	**kPhonetic**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1
Delimiter	space
Syntax	[1-9][0-9]\d{0,3}[A-D]?*?
Description	The phonetic class for the ideograph, as adopted from *Ten Thousand Characters: An Analytic Dictionary*, by G. Hugh Casey, S.J. Hong Kong: Kelly and Walsh, 1980. Ideographs in the same phonetic class have a common phonetic element, such as `U+8015` 耕 and `U+9631` 阱, both assigned to the phonetic class 103. Most classes have a prototype ideograph, which serves as the common phonetic element for the remaining members of the class. For example, `U+4E4D` 乍 is the prototype for ideographs of class 10. Some classes are associated with one to four subsidiary classes, indicated by the letters A through D.

	Some ideographs are assigned multiple classes. This can happen, for example, when an ideograph belongs to one class but is also the prototype for a different class. For example, U+570B 國 is the prototype for class 748, but is also a member of class 1416, which has U+6216 或 as its prototype. Its kPhonetic value is therefore “748 1416.” Multiple values are always in ascending numerical order. An asterisk is appended when an ideograph has the given phonetic class but is not explicitly included in the ideograph list for that class. For example, U+8753 蝓 belongs to the class 1611 but is not explicitly listed in that class. Its kPhonetic value is therefore “1611*.” The Chinese Phonetic Groups page is a useful resource for browsing the kPhonetic property data.

Property	**kPrimaryNumeric**
Status	Informative
Category	Numeric Values
Introduced	3.2
Delimiter	space
Syntax	[0-9]\d+
Description	One or more values of the ideograph when used as a numeral in Chinese and derivative numeric systems. Ideographs which have this property have numeric values that are common, or are standardized to convey a fixed numeric value. For example, 千 always means “thousand”. A native reader is expected to understand the numeric value for these ideographs. If an ideograph has more than one numeric value, the first one is to be considered the most common one, and that first value is used for the Numeric_Value property of the ideograph. The three Chinese numeric-value properties should have no overlap; that is, ideographs with a kPrimaryNumeric value should not have a kAccountingNumeric or kOtherNumeric value as well.

Property	**kPseudoGB1**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]\d{4}
Description	A GB/T 12345-1990 code point assigned to this ideograph for the purposes of including it within Unihan. Pseudo-GB1 codes were used to provide official code points for ideographs not already in national standards, such as ideographs used to write Cantonese, and so on.

Property	**kRSAdobe_Japan1_6**
Status	Provisional
Category	Radical-Stroke Counts
Introduced	4.1
Delimiter	space
Syntax	[CV]\+[0-9]\d{1,5}\+[1-9][0-9]\d{0,2}\.[1-9][0-9]\d?\ [0-9]\d{1,2}
Description	Information on the glyphs in Adobe-Japan1-6 as contributed by Adobe. The value consists of a number of space-separated entries. Each entry consists of three pieces of information separated by a plus sign:

	1) C or V. “C” indicates that the Unicode code point maps directly to the Adobe-Japan1-6 CID that appears after it, and “V” indicates that it is considered a variant form, and thus not directly encoded. 2) The Adobe-Japan1-6 CID. 3) Radical-stroke data for the indicated Adobe-Japan1-6 CID. The radical-stroke data consists of three pieces separated by periods: the Kangxi radical (1–214), the number of strokes in the form the radical takes in the glyph, and the number of strokes in the residue. The standard Unicode radical-stroke form can be obtained by omitting the second value, and the total strokes in the glyph from adding the second and third values.

Property	**kRSUnicode**
Status	Informative
Category	IRG Sources
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]\d{0,2}\'{0,3}\.-?[0-9]\d{1,2}
Description	The standard radical-stroke count for this ideograph in the form “radical.additional strokes.” The radical is indicated by a number in the range 1–214, followed by an optional single apostrophe (U+0027 ' APOSTROPHE), double apostrophe (''), or triple apostrophe (''') suffix. A single apostrophe after the radical indicates a Chinese simplified version of the given radical. Two apostrophes after the radical indicates a non-Chinese simplified version of the given radical. Three apostrophes after the radical indicates a second non-Chinese simplified version of the given radical. The “additional strokes” value is the residual stroke-count, the count of all strokes remaining after eliminating all strokes associated with the radical. This property is also used for additional radical-stroke indices where either an ideograph may be reasonably classified under more than one radical, or alternate stroke count algorithms may provide different stroke counts. This property is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep pīnyīn-based Han collation (and its tailorings) and transliteration reasonably stable. The residual stroke count may be negative. This is because some ideographs (for example, U+225A9 ⺗ and U+29C0A ⻖) are constructed by removing strokes from a standard radical.

Property	**kSBGY**
Status	Provisional
Category	Dictionary Indices
Introduced	3.2
Delimiter	space
Syntax	[0-9]\d{3}\.[0-7][0-9]\d
Description	The position of this ideograph in the *Song Ben Guang Yun* (SBGY) Medieval Chinese character dictionary (bibliographic and general information below). The 25,334 ideograph references are given in the form “ABC.XY”, in which: “ABC” is the zero-padded page number [004..546]; “XY” is the zero-padded number of the ideograph on the page [01..73]. For example, 364.38 indicates the 38th ideograph on Page 364 (i.e. 澍). Where a given Unicode Scalar Value has more than one reference, these are space-delimited.

-- Release information (20080814) --

This release corrects several mappings. This data set now contains a total of 25,334 references, for 19,583 different hanzi.

-- Release information (2003-10-05) --

This release corrects several mappings.

-- Release information (2002-03-10) --

This data set contains a total of 25,334 references, for 19,572 different hanzi (up from 25,330 and 19,511 in the previous release).

This release of the kSBGY data fixes a number of mappings, based on extensive work done since the initial release (compare the initial release counts given below). See the end of this header for additional information.

-- Initial release information (2002-03-10) --

The original data was input under the direction of Professor LUO Fengzhu at Taiwan Taoyuanxian Yuan Zhi University (see below) using an early version of the Big5-based CDP encoding scheme developed at Academia Sinica. During 2000–2002 this raw data was processed and revised by Richard Cook as follows: the data was converted to Unicode encoding using his revised kHanYu mapping tables (first provided to the Unicode Consortium for the Unihan database release 3.1.1d1) and also using several other mapping tables developed specifically for this project; the kSBGY indices were generated based on hand-counts of all page totals; numerous indexing errors were corrected; and the data underwent final proofing.

-- About the print sources --

The SBGY text, which dates to the beginning of the Song Dynasty (c. 1008, edited by 陳彭年 CHEN Pengnian et al.) is an enlargement of an earlier text known as 《切韻》 Qie Yun (dated to c. 601, edited by 陸法言 LU Fayan). With 25,330 head entries, this large early lexicon is important in part for the information which it provides for historical Chinese phonology. The GY dictionary employs a Chinese transcription method (known as 反切) to give pronunciations for each of its head entries. In addition, each syllable is also given a brief gloss.

It must be emphasized that the mapping of a particular SBGY glyph to a single Unicode Scalar Value may in some cases be merely an approximation or may have required the choice of a “best possible glyph” (out of those available in the Unicode repertoire). This indexing data in conjunction with the print sources will be useful for evaluating the degree of distinctive variation in the ideograph forms appearing in this text, and future proofing of this data may reveal additional Chinese glyphs for IRG encoding.

-- Bibliographic information on the print sources --

《宋本廣韻》 <<Song Ben Guang Yun>> [‘Song Dynasty edition of the Guang Yun Rhyming Dictionary’], edited by 陳彭年 CHEN Pengnian et al. (c. 1008).

Two modern editions of this work were consulted in building the kSBGY indices:

《新校正切宋本廣韻》。台灣黎明文化事業公司 出版，林尹校訂1976 年出版。[This was the edition used by Prof. LUO (台灣桃園縣元智大學中語系羅鳳珠), and in the subsequent revision, conversion, indexing and proofing.]

《新校互註·宋本廣韻》。香港中文大學,余迺永 1993, 2000 年出版。ISBN: 962-201-413-

	5; 7-5326-0685-6. [Textual problems were resolved on the basis of this extensively annotated modern edition of the text.] -- Additional Information -- For further information on this index data and the databases from which it is excerpted, see: Cook, Richard S. 2003. 《說文解字·電子版》 Shuo Wen Jie Zi - Dianzi Ban: Digital Recension of the Eastern Han Chinese Grammaticon. PhD Dissertation. Department of Linguistics. Berkeley: University of California.

Property	**kSemanticVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}(<[ks][A-Za-z0-9_]+(:[TBZFJ]+)?(,[ks][A-Za-z0-9_]+(:[TBZFJ]+)?)*)?
Description	The Unicode value for a semantic variant for this ideograph. A semantic variant is an x- or y-variant with similar or identical meaning which can generally be used in place of the indicated ideograph. Also see Section 3.7.2. The basic syntax is a Unicode Scalar Value. It may optionally be followed by additional data. The additional data is separated from the Unicode Scalar Value by a less-than sign (<), and may be subdivided itself into substrings by commas, each of which may be divided into two pieces by a colon. The additional data consists of a series of property tags for another property in the Unihan database indicating the source of the information. If these property tags are themselves subdivided by a colon, the final piece is a string consisting of the letters T (for *tòng*, U+540C 同) B (for *bù*, U+4E0D 不), Z (for *zhèng*, U+6B63 正), F (for *fán*, U+7E41 繁), or J (for *jiǎn* U+7C21 簡/U+7B80 简). T is used if the indicated source explicitly indicates the two are the same (for example, by saying that the one ideograph is “the same as” the other). B is used if the source explicitly indicates that the two are used improperly one for the other. Z is used if the source explicitly indicates that the given ideograph is the preferred form. Thus, kHanYu indicates that U+5231 刱 and U+5275 創 are semantic variants and that U+5275 創 is the preferred form. F is used if the source explicitly indicates that the given ideograph is the traditional form. J is used if the source explicitly indicates that the given ideograph is the simplified form. Data on simplified and traditional variations can be included in this property to document cases where different sources disagree on the nature of the relationship between two ideographs. The kSemanticVariant and kSpecializedSemanticVariant properties need not be consulted when interconverting between traditional and simplified Chinese. As an example, U+3A17 揆 has the kSemanticVariant value "U+6377<kHanYu:TZ". This means that, according to the *Hanyu Da Zidian*, U+3A17 揆 and U+6377 捷 have identical meaning and that U+6377 捷 is the preferred form.

Property	**kSimplifiedVariant**
Status	Provisional

Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}
Description	The Unicode value(s) for the simplified Chinese variant(s) for this ideograph. A full discussion of the `kSimplifiedVariant` and `kTraditionalVariant` properties is found in Section 3.7.1 above. Much of the data on simplified and traditional variants was graciously supplied by Wenlin Institute, Inc.

Property	**kSMSZD2003Index**
Status	Provisional
Category	Dictionary Indices
Introduced	15.1
Delimiter	space
Syntax	\d{1,3}\.\d{2}
Description	This represents the position(s) of the ideograph in the *Soengmou San Zidin* (商務新字典, *New Commercial Press Character Dictionary*). The format is the page within the dictionary followed by the position on the page. If multiple values are present, the first is the primary entry for the ideograph. Other entries are simply cross-references to the primary entry and are in numeric order. The complete bibliographic information for the *Soengmou San Zidin* is: Wong Gongsang 黃港生, ed. *Shangwu Xin Zidian* / *Soengmou San Zidin* 商務新字典 (*New Commercial Press Character Dictionary*). Hong Kong: 商務印書館(香港)有限公司 (Commercial Press [Hong Kong], Ltd.), 2003. ISBN 962-07-0140-2.

Property	**kSMSZD2003Readings**
Status	Provisional
Category	Readings
Introduced	15.1
Delimiter	space
Syntax	[a-z\x{300}\x{301}\x{302}\x{304}\x{308}\x{30C}]+(,[a-z\x{300}\x{301}\x{302}\x{304}\x{308}\x{30C}]+)*\x{7CB5}[a-z]+[1-6]([a-z]+[1-6])?(,[a-z]+[1-6]([a-z]+[1-6])?)*
Description	This represents the Mandarin and Cantonese readings(s) of the ideograph in the *Soengmou San Zidin* (商務新字典, *New Commercial Press Character Dictionary*). The full bibliographic information for this dictionary is found in the description of the `kSMSZD2003Index` property. Mandarin readings are in *hànyǔ pīnyīn*. Cantonese readings are in *jyutping*. Note that some ideographs have readings which would ordinarily be considered invalid, such as polysyllabic readings. If an ideograph has multiple entries, it means that the ideograph has multiple definitions and the readings are grouped in order of those definitions.

Property	**kSpecializedSemanticVariant**
Status	Provisional
Category	Variants

Introduced	2.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}(<[ks][A-Za-z0-9_]+(:[TBZFJ]+)?(,[ks][A-Za-z0-9_]+(:[TBZFJ]+)?)*)?
Description	The Unicode value for a specialized semantic variant for this ideograph. The syntax is the same as for the `kSemanticVariant` property. A specialized semantic variant is an x- or y-variant with similar or identical meaning only in certain contexts. See Section 3.7.2 for a full description.

Property	**kSpoofingVariant**
Status	Provisional
Category	Variants
Introduced	13.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}
Description	The spoofing variants for the ideograph, if any. Spoofing variants include ideograph pairs which look similar, particularly at small point sizes, which are not already z-variants or compatibility variants. See Section 3.7.3 for a full description of spoofing variants. The syntax consists of the ideograph's code point.

Property	**kStrange**
Status	Provisional
Category	Dictionary-like Data
Introduced	14.0
Delimiter	space
Syntax	[ACU] | B:U\+31[0-2AB][0-9A-F] | [FMORY](:U\+[23]?[0-9A-F]{4})? | H(:U\+31[3-8][0-9A-F])+ | I(:U\+[23]?[0-9A-F]{4})* | K(:U\+30[A-F][0-9A-F])+ | S:[4-9][0-9]\d
Description	This property identifies CJK Unified Ideographs that are considered "strange" in one or more ways per the following 12 categories: Category A = [A]symmetric (exhibits a structure that is asymmetric) Category B = [B]opomofo (visually resembles a bopomofo character) Category C = [C]ursive (is cursive or includes one or more cursive components that do not adhere to Han ideograph stroke conventions) ~~Category F = [F]ully-reflective (is fully-reflective or includes components that are fully-reflective, meaning that the mirrored and unmirrored components are arranged side-by-side or stacked top-and-bottom)~~ Category H = [H]angul Component (includes ~~a~~one or more hangul component~~s~~) Category I = [I]ncomplete (appears to be an incomplete version of an existing or possible ideograph, meaning that one or more components appear to be incomplete, without regard to semantics) Category K = [K]atakana Component (includes one or more components that visually resemble a katakana syllable) Category M = [M]irrored (is either mirrored or includes one or more components that are mirrored) Category O = [O]dd Component (includes one or more components that are symbol-like or are otherwise considered odd) Category R = [R]otated (is either rotated or includes one or more components that are rotated) Category S = [S]troke-heavy (has 40 or more strokes)

	Category U = [U]nusual Arrangment/Structure (has an unusual structure or component arrangement) Category Y = S[Y]mmetric (is symmetric, or includes components that are symmetric, meaning that the mirrored and unmirrored components are arranged side-by-side or stacked top-and-bottom) This property is fully documented in Unicode Technical Note #43, “Unihan Database Property ‘kStrange’” [UTN43].

Review note: The Syntax and Description of this property were adjusted to change Category F to Category Y.

Property	**kTaiwanTelegraph**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	~~[0-9]~~\d{4}
Description	The Taiwanese telegraph code for this ideograph, derived from 漢字電報コード変換表 (*Chinese character telegraph code conversion table*), Lin Jinyi, KDD Engineering and Consulting, Tokyo, 1984.

Property	**kTang**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	*?[A-Za-z()\x{E6}\x{251}\x{259}\x{25B}\x{300}\x{30C}]+
Description	The Tang dynasty pronunciation(s) of this ideograph, derived from or consistent with *T’ang Poetic Vocabulary* by Hugh M. Stimson, Far Eastern Publications, Yale University 1976. An asterisk indicates that the word or morpheme represented *in toto* or in part by the given ideograph with the given reading occurs more than four times in the seven hundred poems covered.

Property	**kTayNumeric**
Status	Provisional
Category	Numeric Values
Introduced	17.0
Delimiter	space
Syntax	\d+
Description	The value of the ideograph when used as a numeral in Tày languages with the Han script (*Chữ Nôm Tày*). It can be used alongside `kPrimaryNumeric` or `kAccountingNumeric` since the Chinese vocabulary of numbers is also imported in Tày; it can also be used alongside `kVietnameseNumeric` since the Vietnamese vocabulary of numbers is also imported. Nevertheless, in Tày text, this value should override `kPrimaryNumeric`, `kAccountingNumeric`, and `kVietnameseNumeric` if the ideograph has any of these properties.

Property	**kTGH**
Status	Provisional
Category	Other Mappings
Introduced	11.0

Delimiter	space
Syntax	20[0-9]\d{2}:[1-9][0-9]\d{0,3}
Description	The year that corresponds to the *Tōngyòng Guīfàn Hànzìbiǎo* (通用规范汉字表) table in which the ideograph appears, followed by a colon and its one- to four-digit index number in that list. Published by the Chinese government in 2013, this table includes 8,105 ideographs in three levels containing 3,500 (index numbers 1 through 3500), 3,000 (3501 through 6500), and 1,605 (6501 through 8105) ideographs, respectively. Ideographs for more general use are in the first two levels, with those in the first level being more frequently used. The ideographs in the third level are used for personal names, place names, and for science and technology. The current version year is 2013, and the index numbers range from 1 to 8105.

Property	**kTGHZ2013**
Status	Provisional
Category	Readings
Introduced	13.0
Delimiter	space
Syntax	[0-9]\d{3}\.[0-9]\d{3}(,[0-9]\d{3}\.[0-9]\d{3})*:[a-z\x{300}-\x{302}\x{304}\x{308}\x{30C}]+
Description	One or more Hànyǔ Pīnyīn readings as given in *Tōngyòng Guīfàn Hànzì Zìdiǎn* (full bibliographic information below). Each pīnyīn reading is preceded by the ideograph's location(s) in the dictionary, separated from the reading by a colon. Multiple locations for a given reading are separated by commas. Multiple "location: reading" values are separated by a space. Each location reference is of the form `/[0-9]\d{3}\.[0-9]\d{3}/`. The number preceding the period is the page number, zero-padded to three digits. The first two digits of the number following the period are the entry's position on the page, zero-padded. The third digit is 0 for a main entry and greater than 0 for a parenthesized or bracketed variant of the main entry. – Bibliographical information – 《通用规范汉字字典》(Tōngyòng Guīfàn Hànzì Zìdiǎn = TGHZ; 'General Purpose Normalized Hanzi Dictionary'). 商务印书馆辞书研究中心编 (Dictionary Research Center of the Commercial Press, eds.). 北京: 商务印书馆, 2013 [2013年7月第1版; 2013年9月北京第3次印刷; 印张 22⅞; ISBN 978-7-100-05961-9]. – Release Notes – This data was input and prepared by Jaemin Chung (initial release 2019-04-24). Distinct Unihan hànzì: 8,105 Distinct pīnyīn syllables: 1,296

Property	**kTotalStrokes**
Status	Informative
Category	IRG Sources
Introduced	3.1
Delimiter	~~space~~ N/A
Syntax	[1-9][0-9]\d{0,2}
Description	The total number of strokes in the ideograph (including the radical) according to the stroke-counting conventions of the IRG. ~~When there are two values, then the first is~~

	~~preferred for zh-Hans (CN) and the second is preferred for zh-Hant (TW). When there is only one value, it is appropriate for both.~~ ~~The preferred value is the one most commonly associated with the ideograph in modern text using customary fonts.~~ This property is targeted specifically for use by CLDR collation and transliteration. As such, it is subject to considerations that help keep pīnyīn-based Han collation (and its tailorings) and transliteration reasonably stable.

Property	**kTraditionalVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}
Description	The Unicode value(s) for the traditional Chinese variant(s) for this ideograph. A full discussion of the `kSimplifiedVariant` and `kTraditionalVariant` properties is found in Section 3.7.1 above. Much of the data on simplified and traditional variants was graciously supplied by Wenlin Institute, Inc.

Property	**kUnihanCore2020**
Status	Informative
Category	Dictionary-like Data
Introduced	13.0
Delimiter	N/A
Syntax	[GHJKMPT]{1,7}
Description	Used for ideographs which are in the Unihan Core 2020 set, the minimal set of required ideographs for East Asia. An ideograph is in the Unihan Core 2020 set if and only if it has a value for the `kUnihanCore2020` property. The property value consists of an IRG source specifier as defined in Section 3.10 above.

Property	**kVietnamese**
Status	Provisional
Category	Readings
Introduced	3.1.1
Delimiter	space
Syntax	[A-Za-z\x{110}\x{111}\x{300}-\x{303}\x{306}\x{309}\x{31B}\x{323}]+
Description	The ideograph’s pronunciation(s) in Quốc ngữ.

Property	**kVietnameseNumeric**
Status	Provisional
Category	Numeric Values
Introduced	15.1
Delimiter	space
Syntax	\d+
Description	The value of the character when used as a numeral in Vietnamese with Han script (Hán Nôm). It can be used alongside `kPrimaryNumeric` since the Chinese vocabulary of numbers

	is also imported in Vietnamese. Nevertheless, in Vietnamese text, this value should override kPrimaryNumeric if the character should have both properties.

Property	**kXerox**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]\d{3}:[0-9]\d{3}
Description	The Xerox code for this ideograph.

Property	**kXHC1983**
Status	Provisional
Category	Readings
Introduced	5.1
Delimiter	space
Syntax	[0-9]\d{4}\.[0-9]\d{3}*?(,[0-9]\d{4}\.[0-9]\d{3}*?)*:[a-z\x{300}\x{301}\x{304}\x{308}\x{30C}]+
Description	One or more Hànyǔ Pīnyīn readings as given in the *Xiàndài Hànyǔ Cídiǎn* (full bibliographic information below). Each pīnyīn reading is preceded by the ideograph's location(s) in the dictionary, separated from the reading by a colon; multiple locations for a given reading are separated by commas; multiple "location: reading" values are separated by a space. Each location reference is of the form /[0-9]\d{4}\.[0-9]\d{3}*?/. The number preceding the period is the page number, zero-padded to four digits. The first two digits of the number following the period are the entry's position on the page, zero-padded. The third digit is 0 for a main entry and greater than 0 for a parenthesized variant of the main entry. A trailing asterisk (*) on the location indicates a unifiable variant substituted for an unencoded ideograph. -- Bibliographical information -- 《现代汉语词典》 [Xiàndài Hànyǔ Cídiǎn = XHC; 'Modern Chinese Dictionary']. 中国社会科学院语言研究所词典编辑室编 [Chinese Academy of Social Sciences, Linguisitics Research Institute, Dictionary Editorial Office, eds.]. 北京: 商务印书馆, 1983 [1978 年 12 月第 1 版; 1983 年 1 月第 2 版; 1984 年 1 月北京第 49 次印刷印张 54; 统一书号: 17017.91]. Note that although there are later editions of this important PRC dictionary, reflecting developments and refinements in language and orthographic standardization, these editions should not be used in future revisions to this property. -- Release Notes -- The Unihan version of this data was originally prepared by Richard Cook (initial release 2007-12-12), proofing and revising a subset of data contributed by Dr. George Bell (who input it with the help of Joy Zhao Rouzer, Steve Mann, et al., as one part of their "Quick and Easy Index of Chinese Characters with Attributes"; Bell 1995-2005). Additional data and corrections were provided by Andrew West in 2022 for Unicode Version 15.1. Distinct Unihan hànzì: 10,992; Distinct hànzì: 11,190;

	Distinct pīnyīn syllable types: 1,337; As of Unicode Version 15.1, all ideographs in the dictionary which are not unifiable variants have been encoded. All are encoded as CJK Unified Ideographs, with one exception. The print source includes the entry “0719.100: líng” for U+3007 〇 IDEOGRAPHIC NUMBER ZERO. As this is not a CJK Unified Ideograph, it is not included in the Unihan database; see U+96F6 零.

Property	**kZhuang**
Status	Provisional
Category	Readings
Introduced	16.0
Delimiter	space
Syntax	[a-z]+*?
Description	The most customary Zhuang reading for this ideograph. Readings of words not part of the Standard Zhuang lexicon are suffixed by an asterisk. Among the sources used for the property data are the following: Ancient Zhuang Character Dictionary (古壮字字典), 1989, ISBN 7-5363-0614-8

Property	**kZhuangNumeric**
Status	Provisional
Category	Numeric Values
Introduced	15.1
Delimiter	space
Syntax	\d+
Description	The value of the ideograph when used as a numeral in Zhuang languages with the Han script (Sawndip). It can be used alongside kPrimaryNumeric since the Chinese vocabulary of numbers is also imported in Zhuang. Nevertheless, in Zhuang text, this value should override kPrimaryNumeric if the ideograph should have both properties.

Property	**kZVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+[23]?[0-9A-F]{4}(<[ks][A-Za-z0-9_]+(:[TBZ]+)?(,[ks][A-Za-z0-9_]+(:[TBZ]+)?)*)?
Description	The z-variants for the ideograph, if any. Z-variants are instances where the same abstract shape has been encoded multiple times, either in error or because of the now-abolished Source Separation Rule. Z-variant pairs also have identical semantics. The basic syntax is a Unicode Scalar Value. It may optionally be followed by additional data. The additional data is separated from the Unicode Scalar Value by a less-than sign (<), and may be subdivided itself into substrings by commas. The additional data consists of a series of property tags for another property in the Unihan database indicating the source of the information.

4.2 Listing by Version of Addition to the Unicode Standard

The table below lists the properties of the Unihan database by the version of the Unicode Standard in which they were first added. Also included are properties that were removed in a particular version.

Properties that were removed in Version 4.1 or later are linked to either the UTC document in which their removal was proposed or the particular consensus in UTC meeting minutes that confirms their removal.

Version	Properties Added	Properties Removed
17.0.0	kTayNumeric	kGB7, kJa
16.0.0	kFanqie, kZhuang	kFrequency
15.1.0	kJapanese, kMojiJoho, kSMSZD2003Index, kSMSZD2003Readings, kVietnameseNumeric, kZhuangNumeric	kHKSCS, kIRGDaiKanwaZiten, kKPS0, kKPS1, kKSC0, kKSC1, kRSKangXi
15.0.0	kAlternateTotalStrokes	
14.0.0	kStrange	
13.0.0	kIRG_SSource, kIRG_UKSource, kSpoofingVariant, kTGHZ2013, kUnihanCore2020	kRSJapanese, kRSKanWa, kRSKorean
12.0.0		kDefaultSortKey (private property)
11.0.0	kJinmeiyoKanji, kJoyoKanji, kKoreanEducationHanja, kKoreanName, kTGH	
8.0.0	kJa	
5.2	kHanyuPinyin, kIRG_MSource	
5.1	kXHC1983	
5.0	kCheungBauer, kCheungBauerIndex, kFourCornerCode, kHangul	
4.1	kFennIndex, kIICore, kRSAdobe_Japan1_6	kAlternateKangXi, kAlternateMorohashi
4.0.1	kGSR, kHanyuPinlu, kIRG_USource	
3.2	kAccountingNumeric, kCihaiT, kCompatibilityVariant, kFrequency, kGradeLevel, kOtherNumeric, kPrimaryNumeric, kSBGY	kAlternateHanYu
3.1.1	kCangjie, kCowles, kFenn, kHKGlyph, kHKSCS, kIRG_KPSource, kJIS0213, kKPS0, kKPS1, kKarlgren, kLau, kVietnamese	
3.1	kIRG_HSource, kMeyerWempe, kPhonetic, kTotalStrokes	kAlternateJEF, kJHJ, kRSMerged
3.0	kAlternateJEF, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kIRG_GSource, kIRG_JSource, kIRG_KSource, kIRG_TSource, kIRG_VSource, kJHJ, kRSMerged, kSemanticVariant (reintroduced), kSpecializedSemanticVariant (reintroduced)	
2.1		kSemanticVariant, kSpecializedSemanticVariant
2.0	kAlternateHanYu, kAlternateKangXi, kAlternateMorohashi, kCNS1992, kCantonese, kDaeJaweon, kDefinition, kHanYu, kJapaneseKun, kJapaneseOn, kKangXi, kKorean, kMainlandTelegraph, kMandarin, kMatthews, kMorohashi, kNelson, kRSJapanese, kRSKanWa, kRSKangXi, kRSKorean, kRSUnicode, kSemanticVariant , kSimplifiedVariant, kSpecializedSemanticVariant, kTaiwanTelegraph, kTang, kTraditionalVariant, kZVariant	

The remaining properties were added prior to Unicode 2.0.

4.3 Listing by Location within `Unihan.zip`

The table below lists the properties of the Unihan database. They are organized into groups according to the file within `Unihan.zip` where their values are found. Each property name also links to its description.

The grouping of properties into files may differ between versions of the Unicode Standard. Parsers should not rely on the data for a particular property being listed in any specific file.

File Name	Properties Within File
Unihan_DictionaryIndices.txt	kCheungBauerIndex, kCihaiT, kCowles, kDaeJaweon, kFennIndex, kGSR, kHanYu, kIRGDaeJaweon, kIRGHanyuDaZidian, kIRGKangXi, kKangXi, kKarlgren, kLau, kMatthews, kMeyerWempe, kMorohashi, kNelson, kSBGY, kSMSZD2003Index
Unihan_DictionaryLikeData.txt	kAlternateTotalStrokes, kCangjie, kCheungBauer, kFenn, kFourCornerCode, kGradeLevel, kHDZRadBreak, kHKGlyph, kMojiJoho, kPhonetic, kStrange, kUnihanCore2020
Unihan_IRGSources.txt	kCompatibilityVariant, kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_MSource, kIRG_SSource, kIRG_TSource, kIRG_UKSource, kIRG_USource, kIRG_VSource, kRSUnicode, kTotalStrokes
Unihan_NumericValues.txt	kAccountingNumeric, kOtherNumeric, kPrimaryNumeric, kTayNumeric, kVietnameseNumeric, kZhuangNumeric
Unihan_OtherMappings.txt	kBigFive, kCCCII, kCNS1986, kCNS1992, kEACC, kGB0, kGB1, kGB3, kGB5, ~~kGB7,~~ kGB8, kIBMJapan, ~~kJa,~~ kJinmeiyoKanji, kJis0, kJis1, kJIS0213, kJoyoKanji, kKoreanEducationHanja, kKoreanName, kMainlandTelegraph, kPseudoGB1, kTaiwanTelegraph, kTGH, kXerox
Unihan_RadicalStrokeCounts.txt	kRSAdobe_Japan1_6
Unihan_Readings.txt	kCantonese, kDefinition, kFanqie, kHangul, kHanyuPinlu, kHanyuPinyin, kJapanese, kJapaneseKun, kJapaneseOn, kKorean, kMandarin, kSMSZD2003Readings, kTang, kTGHZ2013, kVietnamese, kXHC1983, kZhuang
Unihan_Variants.txt	kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kSpoofingVariant, kTraditionalVariant, kZVariant

4.4 Listing of Ideographs Covered by the Unihan Database

The following table lists the ideographs covered by the Unihan database, together with the version in which they were added to the Unicode Standard.

Code Point Range	Block Name	Version	Count
U+3400..U+4DB5	CJK Unified Ideographs Extension A	3.0	6,582
U+4DB6..U+4DBF		13.0	10
U+4E00..U+9FA5	CJK Unified Ideographs	1.1	20,902
U+9FA6..U+9FBB		4.1	22
U+9FBC..U+9FC3		5.1	8
U+9FC4..U+9FCB		5.2	8
U+9FCC		6.1	1
U+9FCD..U+9FD5		8.0	9
U+9FD6..U+9FEA		10.0	21
U+9FEB..U+9FEF		11.0	5
U+9FF0..U+9FFC		13.0	13
U+9FFD..U+9FFF		14.0	3

U+F900..U+FA2D	**CJK Compatibility Ideographs†**	1.1	302
U+FA2E..U+FA2F		6.1	2
U+FA30..U+FA6A		3.2	59
U+FA6B..U+FA6D		5.2	3
U+FA70..U+FAD9		4.1	106
U+20000..U+2A6D6	CJK Unified Ideographs Extension B	3.1	42,711
U+2A6D7..U+2A6DD		13.0	7
U+2A6DE..U+2A6DF		14.0	2
U+2A700..U+2B734	CJK Unified Ideographs Extension C	5.2	4,149
U+2B735..U+2B738		14.0	4
U+2B739		15.0	1
U+2B73A..U+2B73F		17.0	6
U+2B740..U+2B81D	CJK Unified Ideographs Extension D	6.0	222
U+2B820..U+2CEA1	CJK Unified Ideographs Extension E	8.0	5,762
U+2CEA2..U+2CEAD		17.0	12
U+2CEB0..U+2EBE0	CJK Unified Ideographs Extension F	10.0	7,473
U+2EBF0..U+2EE5D	CJK Unified Ideographs Extension I	15.1	622
U+2F800..U+2FA1D	CJK Compatibility Ideographs Supplement	3.1	542
U+30000..U+3134A	CJK Unified Ideographs Extension G	13.0	4,939
U+31350..U+323AF	CJK Unified Ideographs Extension H	15.0	4,192
U+323B0..U+33479	CJK Unified Ideographs Extension J	17.0	4,298
Total			102,998

† Note: 12 code points in the CJK Compatibility Ideographs block (U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, and U+FA27 through U+FA29) lack a canonical Decomposition_Mapping value in UnicodeData.txt, and so are not actually CJK *Compatibility* Ideographs. These twelve ideographs are CJK *Unified* Ideographs.

Note that some CJK characters *do not* explicitly have property data in the Unihan database, such as:

Code Point Range	Block Name	Version	Count
U+2E80..U+2E99	CJK Radicals Supplement	3.0	26
U+2E9B..U+2EF3		3.0	89
U+2F00..U+2FD5	Kangxi Radicals	3.0	214
U+3000..U+3037	CJK Symbols and Punctuation	1.1	56
U+3038..U+303A		3.0	3
U+303B..U+303D		3.2	3
U+303E		3.0	1
U+303F		1.1	1
U+3190..U+319F	Kanbun	1.1	16
U+31C0..U+31CF	CJK Strokes	4.1	16
U+31D0..U+31E3		5.1	20
U+31E4..U+31E5		16.0	2
U+3220..U+3243	Enclosed CJK Letters and Months	1.1	36
U+3244..U+324F		5.2	12
U+3280..U+32B0		1.1	49
U+32C0..U+32CB		1.1	12
U+32D0..U+32FE		1.1	47
U+32FF		12.1	1

U+3358..U+3370	CJK Compatibility	1.1	25
U+337B..U+337F		1.1	5
U+33E0..U+33FE		1.1	31
U+1F210..U+1F231	Enclosed Ideographic Supplement	5.2	34
U+1F232..U+1F23A		6.0	9
U+1F23B		9.0	1
U+1F240..U+1F248		5.2	9
U+1F250..U+1F251		6.0	2

Some of the above-listed blocks of CJK characters have additional property data and documentation, in particular see:

- CJKRadicals.txt [CJKRadicals] in the [UCD]
- EquivalentUnifiedIdeograph.txt [EquivalentUnifiedIdeograph] in the Unicode Character Database [UCD]
- Appendix F: Documentation of CJK Strokes, in [Unicode]

4.5 Listing of Additional Sources Used by the Unihan Database

The following table provides a formal identification of sources used by the Unihan database, together with bibliographic information. This listing only includes sources without a corresponding property in the Unihan database. It should not be taken as exhaustive.

Identifier	Bibliographic Information
sGZJZD1989	饒秉才, ed. *Guangzhou Yin Zidian* / *Gwongzau Jam Zidin* 廣州音字典 (*Cantonese Character Dictionary*). Hong Kong: 三聯(香港)有限公司 (Joint Publishing [Hong Kong], Ltd.), 1989. ISBN 962-04-0389-4
sHanyuDaCidian1986	Luo Zhufeng 羅竹風, ed. *Hanyu Da Cidian (Suoyinben)* 漢語大詞典(縮印版) (*Hanyu Da Cidian [Compact Edition]*). n.p.: 漢語大詞典出版社 (Chinese Dictionary Publishing House), 1986. ISBN 7-5432-0014-7

5 History

The Unihan database originated as a Hypercard stack using data provided by such organizations as Apple, RLG, and Xerox. Printed versions are found in *The Unicode Standard, Version 1.0,* volume 2. Electronic versions were available on floppy disk in the form of a file called `CJKXREF.TXT`.

The first general electronic release of `CJKXREF.TXT` (961 kB) was included with Unicode 1.1.5 in July 1995. This version of the file is in a multi-column format and includes the data used in printing *The Unicode Standard, Version 1.0,* volume 2 with the exception of the Fujitsu mappings, which were found to be incorrect and withdrawn.

The electronic version of the Unihan database was substantially revised for the publication of Unicode 2.0.0 in July 1996. The file was renamed `UNIHAN.TXT`; its permanent, archival link is Unihan-1.txt (7.9 MB). The format of the file is essentially the same as the current release, although consolidated into a single file. The properties were explicitly named for the first time. The data was at the time maintained using custom, MacApp-based database software. The source code for this software used an enumerated type for the numeric property tags, and the enumerator names (each beginning with a `k` indicating their use as a constant) were used in the text file as property names.

The difficulty of downloading a file 19 MB in size with the technology of the time led to the Unihan database being made available as both a single text file and compressed archives of that text file as of Unicode 3.1.0 in March 2001. The format of the Unihan database remained essentially unchanged until Unicode 5.1.0 (April 2008), when the text file was no longer included and the database became available only as a zipped archive.

Finally, the archive was changed from containing one text file to containing multiple text files as of Unicode 5.2.0 (October 2009).

Anomalies and formatting errors are to be found in various versions of the database file. Specifically:

- `Unihan-1.txt`, the version 2.0.0 Unihan database file, was at some point accidentally truncated on line 330,553 (partway through the data for `U+8BC1` 证). No corrected version of the file was made available. Instead, it was superseded by the Unihan-2.txt (10 MB) file released with Unicode 2.1.2 in May 1998. The CD that is included with the Unicode Version 2.0 book has the same truncation issue, specifically that the files at `{DOS,MAC,UNIX}/MAPPINGS/EASTASIA/UNIHAN.TXT` are truncated at the same position.
- The Version 3.1.1 Unihan database file, `Unihan-3.1.1.txt`, includes the following anomalous record at line 246,442: `U+64AC 297`.
- The Versions 2.0.0, 2.1.2, 3.0.0, and 3.1.0 Unihan database files are not encoded in UTF-8. Rather, individual property values are encoded using an encoding appropriate for their language, such as Big Five for traditional Chinese.

Please refer to Unicode Technical Note #45, “Unihan Property History” [UTN45], for more information about the history of Unihan database properties, in terms of which properties are included in each version of the Unicode Standard from Version 1.1.5 onward, along with how many ideographs are covered by each property.

References

For references for this annex, see Unicode Standard Annex #41, “Common References for Unicode Standard Annexes.”

Acknowledgements

John H. Jenkins 井作恆 (RIP) was the author of the initial version of this annex, and served as a co-editor up through and including Version 33 for Unicode Version 15.0.0. Richard Cook served as a co-editor up through and including Version 37 for Unicode Version 16.0.0.

Modifications

The following summarizes modifications from the previous revision of this annex.

Revision 38

- **Proposed Update** for Unicode 17.0.0.
- Removed Richard Cook as a co-editor.
- Removed the provisional `kGB7` (see UTC Consensus 181-C16 and UTC Action Item 181-A40) and `kJa` (see UTC Consensus 183-C43 and UTC Action Item 183-A98) properties.
- Added the provisional `kTayNumeric` property (see UTC Consensus 183-C50 and UTC Action Item 183-A126).
- Updated Section 3.3 to add references to the `kDaeJaweon`, `kHanYu`, `kIRGDaiKanwaZiten`, `kKangXi`, and `kMorohashi` properties (see UTC Action Item 183-A99).
- Updated the descriptions of the `kCantonese`, `kDaeJaweon`, `kIRGDaeJaweon`, `kIRG_MSource`, `kIRG_VSource`, `kKoreanEducationHanja`, and `kKoreanName` properties.
- Updated the syntax and descriptions of the `kAlternateTotalStrokes`, `kIRG_GSource` (see UTC Action Items 181-A70, 182-A101, 183-A139, and 183-A156), `kIRG_SSource` (see UTC Action Item 180-A17), `kIRG_TSource` (see UTC Action Item 180-A17), and `kStrange` properties.

- Updated the delimiter and description of the `kTotalStrokes` property (see UTC Consensus 183-C72 and UTC Action Item 183-A178).
- Updated the syntax of 50 properties to consistently use `\d` instead of `[0-9]`.
- Added U+2B73A..U+2B73F, U+2CEA2..U+2CEAD, and the CJK Unified Ideographs Extension J block to the first table in Section 4.4 (see UTC Action Item 180-A17)

~~**Revision 37**~~

- ~~**Reissued** for Unicode 16.0.0.~~
- ~~Clarified the relationship between the `Equivalent_Unified_Ideograph` property and the Unihan database in Section 2.1.1.~~
- ~~Updated the sorting algorithm examples in Section 2.1.2.~~
- ~~Added a reference to the `RSIndex.txt` data file at the end of Section 2.1.2.~~
- ~~Updated Section 2.1.2 and Section 3.6 to describe a second non-Chinese simplified radical.~~
- ~~Changed the delimiter of the `kAccountingNumeric` property from *space* to *N/A*.~~
- ~~Added the provisional `kFanqie` and `kZhuang` properties.~~
- ~~Removed the provisional `kFrequency` property.~~
- ~~Updated the syntax and description of the `kIRG_GSource`, `kPhonetic`, and `kRSUnicode` properties.~~
- ~~Updated the description of the `kPrimaryNumeric` property.~~
- ~~Added U+31E4 and U+31E5 to the second table in Section 4.4~~
- ~~Added a reference to `EquivalentUnifiedIdeograph.txt` at the end of Section 4.4.~~

Previous revisions can be accessed with the “Previous Version” link in the header.
