L2/13-094

Proposed Update Unicode Standard Annex #38

UNICODE HAN DATABASE (UNIHAN)

Version	Unicode 6.3.0 (draft 6)
Editors	John H. Jenkins 井作恆 Richard Cook 曲理查 Ken Lunde 小林劍
Date	2013-03-29
This Version	http://www.unicode.org/reports/tr38/tr38-14.html
Previous Version	http://www.unicode.org/reports/tr38/tr38-13.html
Latest Version	http://www.unicode.org/reports/tr38/
Latest Proposed Update	http://www.unicode.org/reports/tr38/proposed.html
Revision	14

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*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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 CJK Unified 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 ideographic 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 is; rather, it defines it as being the equivalent of, say, 0x523B in GB 2312, 0x14421 in CNS 11643, 0x306C in JIS X 0208, and so on.

In practice, implementation of ideographs requires large amounts of ancillary data. Input methods require information such as pronunciations, 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 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 fields.

2 Mechanics

2.1 Database design

The working copy of the Unihan database is maintained privately by the Unicode Consortium. The two public versions are snapshots of this data at a particular point of time.

The database consists of a number of fields containing data for each Han ideograph in the Unicode Standard. The fields are all named, and the names 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."

Most of these are made available in the public releases. The fields not part of the public releases are, with two exceptions, either needed only for internal accounting or similar purposes. The remaining two private fields are convenience fields only; because their values can be determined algorithmically from other data in the database, there is no need to actually include them in the public releases. They are:

- ***kDefaultSortKey***

  This is a 32-bit integer which provides a default radical-stroke ordering for the characters in the database. 31 of the 32 bits are used as a bitfield as follows:

  Bits 0-16 are a representation of the character's code point:

  - U+4E00 through U+9FFF are mapped to 0x00000 through 0x051FF; that is, 0x4E00 is subtracted from the Unicode Scalar Value.
  - U+3400 through U+4DFF are mapped to 0x05200 through 0x06BFF; that is, 0x1E00 is added to the Unicode Scalar Value.
  - U+20000 through U+2F7FF are mapped to 0x06C00 through 0x1F4FF; that is, 0x19400 is subtracted from the Unicode Scalar Value.
  - U+F900 through U+FAFF are mapped to 0x1F600 through 0x1F7FF; that is,

0xFD00 is added to the Unicode Scalar Value.
    - U+2F800 through U+2FFFF are mapped to 0x1F800 through 0x1FFFF; that is, 0x10000 is subtracted from the Unicode Scalar Value.

  The net result of these remappings is to reorder the blocks (main CJK Unified Ideographs, Extension A, Extension B, Extension C, Extension D, Compatibility Ideographs, Compatibility Extension), and to leave unused blocks of several thousand code points between the mapped ranges for Extension D and the first Compatibility Ideographs block and several thousand more after the mapped range for the Compatibility Ideographs Extension.

  Bits 17-22 are the character's residual stroke count (0 through 63). The residual stroke count taken is from the first value in the character's `kRSUnicode` field.

  Bits 23-30 are the character's KangXi radical number used (1 through 214). The radical number used is that of the first value in the character's `kRSUnicode` field. The difference between simplified and traditional radical is ignored.

  Note that bit 31 is unused, so it makes no difference whether the sort key is treated as signed or unsigned.

  The `kDefaultSortKey` field thus defines a consistent way of ordering all the characters in Unihan, first by radical-stroke, then by Unicode block (with the compatibility blocks coming last), and finally by code point. It is not the most efficient sorting key possible, but it has the advantage of being easily generated and does not require existing keys to be regenerated when new ideographs or compatibility ideographs are added to the standard.

- ***`UTF8`***

  This is (as one might expect) the character's UTF-8 encoding. It is also the only field name not starting with "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 field values, operate on Normalization Form D (NFD), primarily because that makes the regular expressions simpler.

2.2 Unihan.zip

Included with the Unicode Character Database 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 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 fields in the Unihan database.

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

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 database field name, and the value for the database field for the given Unicode Scalar Value. For most of the fields, if multiple values are possible, the values are separated by spaces. No character may have more than one instance of a given field associated with it, and no empty fields are included in any of the files archived inside `Unihan.zip`.

There is no formal limit on the lengths of any of the field values. Any Unicode characters may be used in the field values except for double quotes and control characters (especially tab, newline, and carriage return). Most fields have a more restricted syntax, such as the `kKangXi` field 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 field-type as primary and secondary keys, respectively.

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

2.3 Web Access

The URI for interactive access to the contents of the Unihan database is http://www.unicode.org/charts/unihan.html. For production reasons, the version available for interactive 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 CEDICT and Jim Breen's EDICT projects. These additional data are not available in the other versions.

There are also two indices: a grid index grouping the characters in blocks of 256 and a radical-stroke index. A search page is also available. Individual characters 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 character, and click "Lookup". You will be taken to an information page for the character. 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 Field Types

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

Again, it is important to remember that all data in the Unihan database has been donated to the Unicode Consortium. Unicode currently has no staff with the responsibility to maintain or update the Unihan database. This means that, for example, the data is more complete for Chinese than for other languages simply because more

data has been donated for Chinese than for other languages.

3.1 IRG Sources

Among the few normative parts of the Unihan database, and the most exhaustively checked fields, are the nine IRG source fields: `kIRG_GSource` (PRC and Singapore), `kIRG_HSource` (Hong Kong SAR), `kIRG_JSource` (Japan), `kIRG_KPSource` (North Korea), `kIRG_KSource` (South Korea), `kIRG_MSource` (Macao), `kIRG_TSource` (Taiwan), `kIRG_USource` (Unicode/USA), 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 characters to their standards on an ad hoc basis in order to make sure they were included. The various procedures involved in submitting characters 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 character's Unicode Scalar Value. This has change with the inclusion of Extension C in version 5.2.0 of the Unicode Standard. The values are now indices as described in [UTR45].

The syntax for the values used in the various IRG source fields matches that found in ISO/IEC 10646:2011.

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

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

The `kIICore` field is also defined by the IRG and normative. It should be taken as a boolean; if a character has a valid for the kIICore field, it indicates that the character is in IICore, the IRG-produced minimal set of required ideographs for East Asian use.

Each individual value in this field is either P (for preliminary, meaning it has been approved by the IRG but not by WG2), or the ISO/IEC 10646 subset identifier for the subset(s) containing this character.

3.2 Other Mappings

There are twenty-four fields in this category. They consist of mapping tables between the ideographic portions of Unicode and those of encoded character sets or character collections *not* used by the IRG in its work, although some of the character sets covered do mirror official IRG sources. For example, data for mapping GB 12345 is included, even though GB 12345 is a part of the IRG's G-source. The difference between the two is that the `kGB1` field maps all of GB 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 12345.

3.3 Dictionary Indices

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

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

Two, standard dictionaries provide a reference for scholars or students who wish more information about a character.

Third, standard dictionaries are a source for unencoded characters. 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 HK SCS.

As elsewhere, the set of dictionaries covered represent data that has been volunteered. There are important dictionaries (for example, the *Hanyu Da Cidian*, the *Shuowen*) for which formal indices should be provided. And as elsewhere, the data which has been volunteered is weighted heavily in favor of Chinese.

Four of the dictionary fields 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 two (kIRGDaeJaweon and kIRGDaiKanwaZiten) are not. We have, nonetheless, retained their data for reference purposes.

For all four, there are clone fields to hold Unicode indices into the same four dictionaries. By and large, the data in the IRG fields and their Unicode counterparts is the same—but not always.

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

The general-purpose Chinese dictionary fields are: kCihaiT, kFennIndex, kGSR, kKarlgren, kMatthews, and kSBGY. These represent large, standard Chinese-Chinese, Chinese-English dictionaries, or definitive sinological studies.

The Cantonese dictionary fields are kCheungBauerIndex, kCowles, kLau, and kMeyerWempe. All but Cheung-Bauer are large character-based Cantonese-English dictionaries.

At present, the only other dictionary field is kNelson, the character’s index in the first edition of Andrew N. Nelson’s excellent and popular *Modern Reader’s Japanese-English Character Dictionary*.

In selecting dictionaries for inclusion—outside of the general consideration of who is willing to volunteer what data—we aim for including large dictionaries rather than small ones, and standard dictionaries such as serious students might have on their shelves.

3.4 Readings

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

Any attempt at providing a reading or set of readings for a character 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, and covers four large urban centers (Guangzhou, Shenzhen, Macao, and Hong Kong), with Guangzhou Cantonese somewhat infected by Mandarin and Hong Kong Cantonese more than a little infected by English.

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 a character 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 characters 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 礒, used in a Hong Kong place name). 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 fields including information one might find in a dictionary (such as a character'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 character'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 six radical-stroke counts for Unihan, although only three (`kRSAdobe_Japan1_6`, `kRSKangXi`, and `kRSUnicode`) can be considered complete; the others (`kRSJapanese`, `kRSKanWa`, and `kRSKorean`) are placeholders to be filled in later. Three are based on IRG standard dictionaries: the *Hanyu Da Zidian*, which uses a slightly different radical system from the others, is not included, although *Hanyu Da Zidian* radical-stroke data can be calculated using the `kHDZRadBreak` field.

All the radical-stroke fields are based on the radical-system introduced by the 18th-century *KangXi* dictionary. Each ideograph is assigned one of 214 radicals. In most cases, the radical assigned is the natural radical, giving a clue as to the character's meaning; in the rest, the radical is arbitrary, based on the character's structure. One also counts the character's residual strokes, that is, the number of brush strokes required to write everything in the character except the radical.

To find a character using the radical-stroke system, one determines its radical and the number of residual strokes, then looks through the list of characters 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 a character does not have a natural radical, it can sometimes be hard to tell what the radical ought to be (for example, 井 being assigned arbitrarily the radical 二). Even if the character naturally falls into radical-like pieces, it can be hard to tell which is the radical and which the phonetic (for example, 和, which looks like it belongs to the radical 禾, actually belongs to the radical 口). Moreover, since Unicode encodes characters, not glyphs, two different glyphs for the same character may have different residual strokes (such as 者, which can be written either with or without a dot, altering its stroke count between nine and eight, respectively).

We include multiple radical-stroke systems to allow for this. Three of the radical-stroke

fields represent the character's radical-stroke count as determined by its position within a standard IRG dictionary. Two more (`kRSJapanese` and `kRSUnicode`) are intended to cover a "typical" Japanese radical-stroke count, and everything else, respectively. Finally, there is the `kRSAdobe_Japan1_6` field which contains more detailed information on the glyph used for the character in the Adobe Japan 1-6 character set.

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

The `kRSUnicode` field also uses an apostrophe after the radical number to indicate that the character uses a standard simplification. In simplified Chinese, many radicals have standard, simplified forms, such as 讠, which is the simplified form of the radical 言

There is, by the way, no standard way of ordering characters within a given radical-stroke group. Unicode's radical-stroke charts order characters with the same radical-stroke count by the Unicode block in which they occur. If looking for a character with radical 64 (手) and ten residual strokes, one knows that of the 175 candidates in Unicode 5.2.0, 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 to one of five categories, and sorting by those categories. When this "first stroke" data is available for all of Unihan, it will be added to the Unihan database and simplify the process of finding a character 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. Both are described in some detail in the Unicode Standard. Briefly, however, the three-dimensional model uses the x-axis to represent meaning, and the y-axis to represent abstract shape. The z-axis is used for stylistic variations.

To illustrate, 說 and 貓 have different positions along the x-axis, because they mean two entirely different things (*to speak* and *cat*, respectively). 貓 and 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, 說 and 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` field.

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

3.7.1 Simplified and Traditional Chinese Variants

The kTraditionalVariant and kSimplifiedVariant fields are used in character-by-character conversions between simplified and traditional Chinese (SC and TC, respectively). For any character 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 并 U+4E95. This is the most common case, and is indicated by both the kSimplifiedVariant and kTraditionalVariant fields 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 field lists the character(s) to which it is mapped and the kTraditionalVariant field is empty. An example would be 書 U+66F8 whose kSimplifiedVariant field is 书 U+4E66.
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 field lists the character(s) to which it is mapped and the kSimplifiedVariant field is empty. An example would be 学 U+5B66 whose kTraditionalVariant field is 學 U+5B78.
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 character when converting between SC and TC. In this case, the character is its own simplification as well as the simplification for other characters. An example would be 后 U+540E, which is the simplification for itself and for 後 U+5F8C. 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 fields 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 苧 U+82E7. 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 苎 U+82CE, and when converting SC to TC it is mapped to 薴 U+85B4. In this case, both kTraditionalVariant and kSimplifiedVariant fields are defined but X is not included in the values for either.

In practice, conversion between simplified and traditional Chinese is complicated by three 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, 脏 U+810F is mapped to 臟 U+81DF when it means "viscera" and to 髒 U+9AD2 when it means "dirty."
2. An SC character 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 characters is identified as the preferred or correct character to use in SC. For example, both 猫 U+732B and 貓 U+8C93 are acceptable TC characters meaning "cat," but only 猫 U+732B should be used in SC.
3. 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 硬磁盘 in the PRC, and 硬碟 in Taiwan.

3.7.2 Semantic Variants

The remaining two variation fields, `kSemanticVariant` and `kSpecializedSemanticVariant`, are used to mark cases where two characters 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. 井 means *a well*, and although 丼 can also mean *a well* and be used for 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.

3.8 Numeric Values

Finally, we have three fields, `kAccountingNumeric`, `kOtherNumeric`, and `kPrimaryNumeric` to 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 three fields.

4 The Fields

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

4.1 Alphabetical Listing

For each field 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 field.

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 field is no longer to be used.

Fields which allow multiple values have a *Delimiter* defined as “space”. Fields which do not have multiple values (such as the IRG source fields) have this defined as “N/A”. Some fields do not currently have multiple values in the data but may do so in the future.

For most fields 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 field.

However, for certain fields the ordering of values may be significant; in such cases, the significance is specified in the Description for the field~~, with the heading “Multiple Value Order:”~~. For example, see the kMandarin field. In later versions of the Unicode Character Database, a field 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 field’s *Syntax* regular expression. Note that the value for any given field'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 field contains, but also source information, known limitations, methodology used in deriving the data, and so on.

The fields covered in the table are: kAccountingNumeric, kBigFive, kCangjie, kCantonese, kCCCII, kCheungBauer, kCheungBauerIndex, kCihaiT, kCNS1986, kCNS1992, kCompatibilityVariant, kCowles, kDaeJaweon, kDefinition, kEACC, kFenn, kFennIndex, kFourCornerCode, kFrequency, kGB0, kGB1, kGB3, kGB5, kGB7, kGB8, kGradeLevel, kGSR, kHangul, kHanYu, kHanyuPinlu, kHanyuPinyin, kHDZRadBreak, kHKGlyph, kHKSCS, kIBMJapan, kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_MSource, kIRG_TSource, kIRG_USource, kIRG_VSource, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kJapaneseKun, kJapaneseOn, kJis0, kJis1, kJIS0213, kKangXi, kKarlgren, kKorean, kKPS0, kKPS1, kKSC0, kKSC1, kLau, kMainlandTelegraph, kMandarin, kMatthews, kMeyerWempe, kMorohashi, kNelson, kOtherNumeric, kPhonetic, kPrimaryNumeric, kPseudoGB1, kRSAdobe_Japan1_6, kRSJapanese, kRSKangXi, kRSKanWa, kRSKorean, kRSUnicode, kSBGY,

kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTaiwanTelegraph, kTang, kTotalStrokes, kTraditionalVariant, kVietnamese, kXerox, kXHC1983, and kZVariant.

Property	**kAccountingNumeric**
Status	Informative
Category	Numeric Values
Introduced	3.2
Delimiter	space
Syntax	[0-9]+
Description	The value of the character when used in the writing of accounting numerals. Accounting numerals are used in East Asia to prevent fraud. Because a number like ten (十) is easily turned into one thousand (千) with a stroke of a brush, monetary documents will often use an accounting form of the numeral ten (such as 拾) in their place. The three numeric-value fields should have no overlap; that is, characters with a kAccountingNumeric value should not have a kPrimaryNumeric or kOtherNumeric value as well.

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 character in hex; note that this does not cover any of the Big Five extensions in common use, including the ETEN extensions.

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 character. 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 Cantonese pronunciation(s) for this character using the jyutping romanization. A full description of jyutping can be found at <http://www.lshk.org/cantonese.php>. The main differences between jyutping and the Yale romanization previously used are: 1) Jyutping always uses tone numbers and does not distinguish the high falling and high level tones. 2) Jyutping always writes a long a as “aa”. 3) Jyutping uses “oe” and “eo” for the Yale “eu” vowel. 4) Jyutping uses “c” instead of “ch”, “z” instead of “j”, and “j” instead of “y” as initials. 5) A non-null initial is always explicitly written (thus “jyut” in jyutping instead of Yale’s “yut”). Cantonese pronunciations are sorted alphabetically, not in order of frequency. N.B., the Hong Kong dialect of Cantonese is in the process of dropping initial NG- before non-null finals. Any word with an initial NG- may actually be pronounced without it, depending on the speaker and circumstances. Many words with a null initial may

	similarly be pronounced with an initial NG-. Similarly, many speakers use an initial L- for words previously pronounced with an initial N-. Cantonese data are derived from the following sources: 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. Roy T. Cowles, A Pocket Dictionary of Cantonese, Hong Kong: University Press, 1999 (kCowles). Sidney Lau, A Practical Cantonese-English Dictionary, Hong Kong: Government Printer, 1977 (kLau). Bernard F. Meyer and Theodore F. Wempe, Student's Cantonese-English Dictionary, Maryknoll, New York: Catholic Foreign Mission Society of America, 1947 (kMeyerWempe). 饒秉才, ed. 廣州音字典, Hong Kong: Joint Publishing (H.K.) Co., Ltd., 1989. 中華新字典, Hong Kong:中華書局, 1987. 黃港生, ed. 商務新詞典, Hong Kong: The Commercial Press, 1991. 朗文初級中文詞典, Hong Kong: Longman, 2001.

Property	kCCCII
Status	Provisional
Category	Other Mappings
Introduced	2.0

Delimiter	space
Syntax	[0-9A-F]{6}
Description	The CCCII mapping for this character in hex.

Property	**kCheungBauer**
Status	Provisional
Category	Dictionary-like Data
Introduced	5.0
Delimiter	space
Syntax	[0-9]{3}\/[0-9]{2};[A-Z]*;[a-z1-6\[\]\/,]+
Description	Data regarding the character 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 character's radical-stroke index as a three-digit radical, slash, two-digit stroke count; (2) the character'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]{3}\.[01][0-9]
Description	The position of the character 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-like Data
Introduced	3.2
Delimiter	space
Syntax	[1-9][0-9]{0,3}\.[0-9]{3}
Description	The position of this character 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 character in hex.

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 character in hex.

Property	**kCompatibilityVariant**
Status	Normative
Category	Variants

Introduced	3.2
Delimiter	N/A
Syntax	U\+2?[0-9A-F]{4}
Description	The canonical Decomposition_Mapping value for the ideograph, derived from UnicodeData.txt. This field is derived by taking the non-null Decomposition_Mapping values from Field 5 of UnicodeData.txt, for characters 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]{1,4}(\.[0-9]{1,2})?
Description	The index or indices of this character 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 a character was added after the original indices were determined. Cowles is missing indices 1222 and 4949, and four characters in Cowles are part of Unicode's "Hangzhou" numeral set: 2964 (U+3025), 3197 (U+3028), 3574 (U+3023), and 4720 (U+3027). ~~Approximately 100 characters from Cowles which are not currently encoded are being submitted to the IRG by Unicode for inclusion in future versions of the standard.~~

Property	**kDaeJaweon**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	N/A

Syntax	[0-9]{4}\.[0-9]{2}[01]
Description	The position of this character in the Dae Jaweon (Korean) 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 characters actually in the dictionary and “1” for characters not found in the dictionary and assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth character on page 1187. A character not in this dictionary but assigned a position between the 6th and 7th characters 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., 1988.

Property	**kDefinition**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	N/A
Syntax	[^\t"]+
Description	An English definition for this character. 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 fields. Definitions specific to non-Chinese languages or Chinese dialects other than modern Mandarin are marked, e.g., (Cant.) or (J). Major definitions are separated by semicolons, and minor definitions by commas. Any valid Unicode character (except for tab, double-quote, and any line break character) may be used within the definition field.

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 character 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 earlier versions of CCCII (see kCCCII) and is therefore identical with CCCII for many characters. The kEACC field 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	kFenn
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1.1
Delimiter	space
Syntax	[0-9]+a?[A-KP*]
Description	Data on the character from _The Five Thousand Dictionary_ (aka _Fenn's Chinese-English Pocket 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 character'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 character included in the dictionary only because it is the phonetic element in other characters. An asterisk is used instead of a letter in the final position to indicate a character which belongs to one of Soothill's phonetic groups but is not found in Fenn's dictionary. Characters 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	[1-9][0-9]{0,2}\.[01][0-9]
Description	The position of this character in _Fenn's Chinese-English Pocket Dictionary_ by Courtenay H. Fenn, Cambridge, Mass.: Harvard University Press, 1942. 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]{4}(\.[0-9])?
Description	The four-corner code(s) for the character. This data is derived from data provided in the public domain by Hartmut Bohn, Urs App, and Christian Wittern. The four-corner system assigns each character a four-digit code

	from 0 through 9. The digit is derived from the “shape” of the four corners of the character (upper-left, upper-right, lower-left, lower-right). An optional fifth digit can be used to further distinguish characters; the fifth digit is derived from the shape in the character’s center or region immediately to the left of the fourth corner. The four-corner system is now used only rarely. Full descriptions are available online, e.g., at <http://en.wikipedia.org /wiki/Four_corner_input>. Values in this field consist of four decimal digits, optionally followed by a period and fifth digit for a five-digit form.

Property	**kFrequency**
Status	Provisional
Category	Dictionary-like Data
Introduced	3.2
Delimiter	N/A
Syntax	[1-5]
Description	A rough frequency measurement for the character based on analysis of traditional Chinese USENET postings; characters with a kFrequency of 1 are the most common, those with a kFrequency of 2 are less common, and so on, through a kFrequency of 5.

Property	**kGB0**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 2312-80 mapping for this character in ku/ten form.

Property	**kGB1**
Status	Provisional

Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 12345-90 mapping for this character in ku/ten form.

Property	**kGB3**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 7589-87 mapping for this character in ku/ten form.

Property	**kGB5**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 7590-87 mapping for this character in ku/ten form.

Property	**kGB7**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 8565-89 mapping for this character in ku/ten form.

Property	**kGB8**
Status	Provisional

Category	Other Mappings
Introduced	2.0
Delimiter	N/A
Syntax	[0-9]{4}
Description	The GB 8565-89 mapping for this character in ku/ten 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 character; 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]{4}[a-vx-z]\'?
Description	The position of this character 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 4-digits, "a" is a letter in the range [a..z] (excluding "w"), optionally followed by 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 <http://stedt.berkeley.edu/> 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{11FF}]+
Description	The modern Korean pronunciation(s) for this character in Hangul.

Property	kHanYu
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[1-8][0-9]{4}\.[0-3][0-9][0-3]
Description	The position of this character in the Hanyu Da Zidian (HDZ) Chinese character dictionary (bibliographic information below). The character 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 character on the page [01..32]; "Z" is "0" for a character actually in the dictionary, and greater than 0 for a character assigned a "virtual" position in the dictionary. For example, 53024.060 indicates an actual HDZ character, the 6th character on Page 3,024 of Volume 5 (i.e. 鱻 [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 character 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 character references (positions are given for all HDZ head entries, including source-internal unifications), and 1369 of which are virtual character positions (see note below).

A total of 55818 distinct Unihan characters are assigned mappings in this data. Because of IRG source-internal unifications, a given character 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” field 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” fields 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 characters within a given page. It is unlikely that errors across page breaks remain. Possible future deunifications of source-internal unifications will necessitate update of USV for some references. Under no circumstances should the source-internal unification (duplicate USV) mappings be removed from this data set.

	Note: Source (3) contributed only actual HDZ character 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 character together with the actual HDZ character 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.

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}]+\([0-9]+\)
Description	The Pronunciations and Frequencies of this character, 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 2003/02/04 contained a total of 3800 records, including ○ [U+3007] líng 'IDEOGRAPHIC NUMBER ZERO', not included in Unihan 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 character, with numeric tone marks (1-5, where 5 indicates the “neutral tone”) immediately following each alphabetic string.

Immediately following the numeric tone mark, a numeric string appears in parentheses: e.g. in “a1(392)” the numeric string “392” indicates the sum total of the frequencies of the pronunciations of the character 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 character 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 <http://www.wenlin.com> 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 2312–80 (plus 嗇). 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 character 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 “,” (comma). The list of locations is followed by “:” (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 | 㛚 | 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 卤 U+5364 is “10093.130: xī,lǔ 74609.020: lǔ,xī”. This means that 卤 U+5364 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). ~~Multiple Value Order: Individual entries are in same order as they are found in the Hanyu Da Zidian. This is true both for the locations and the individual readings. While this is generally in the order of utility for modern Chinese, such is not invariably the case, as the example above illustrates.~~ 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 34,131 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]{4} \.[0-3][0-9]0
Description	Indicates that 《漢語大字典》 Hanyu Da Zidian has a radical break beginning at this character’s position. The field consists of the

	radical (with its Unicode code point), a colon, and then the Hanyu Da Zidian position as in the kHanyu field.

Property	kHKGlyph
Status	Provisional
Category	Dictionary-like Data
Introduced	3.1.1
Delimiter	space
Syntax	[0-9]{4}
Description	The index of the character 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 characters as used in the Hong Kong school system. The index is an integer, zero-padded to four digits.

Property	kHKSCS
Status	Provisional
Category	Other Mappings
Introduced	3.1.1
Delimiter	N/A
Syntax	[0-9A-F]{4}
Description	Mappings to the Big Five extended code points used for the Hong Kong Supplementary Character Set-2008 (HKSCS-2008).

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 character in hexadecimal.

Property	kIICore
Status	Normative

Category	IRG Sources
Introduced	4.1
Delimiter	space
Syntax	2\.1
Description	A boolean indicating that a character is in IICore, the IRG-produced minimal set of required ideographs for East Asian use. A character is in IICore if and only if it has a value for the kIICore field. The only value currently in this field is “2.1”, which is the identifier of the version of IICore used to populate this field.

Property	**kIRGDaeJaweon**
Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	[0-9]{4}\.[0-9]{2}[01]
Description	The position of this character in the Dae Jaweon (Korean) 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 characters actually in the dictionary and “1” for characters not found in the dictionary and assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth character on page 1187. A character not in this dictionary but assigned a position between the 6th and 7th characters on page 1187 for sorting purposes would have the code “1187.061” This field represents the official position of the character 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., 1988.

Property	**kIRGDaiKanwaZiten**
Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	[0-9]{5}\'?
Description	The index of this character in the Dai Kanwa Ziten, aka Morohashi dictionary (Japanese) used in the four-dictionary sorting algorithm. This field represents the official position of the character within the DaiKanwa dictionary as used by the IRG in the four-dictionary sorting algorithm. The edition used is the revised edition, published in Tokyo by Taishuukan Shoten, 1986.

Property	**kIRGHanyuDaZidian**
Status	Provisional
Category	Dictionary Indices
Introduced	3.0
Delimiter	space
Syntax	[1-8][0-9]{4}\.[0-3][0-9][01]
Description	The position of this character in the Hanyu Da Zidian (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 characters actually in the dictionary and “1” for characters not found in the dictionary and assigned a “virtual” position in the dictionary. Thus, “32264.080” indicates the eighth character on page 2264 in volume 3. A character not in this dictionary but assigned a position between the 8th and 9th characters on this page for sorting purposes would have the code “32264.081” This field represents the official position of the character within

	the Hanyu Da Zidian 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]{3}\.[0-7][0-9][01]
Description	The official IRG position of this character in the 《康熙字典》 Kang Xi 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 characters actually in the dictionary and “1” for characters not found in the dictionary but assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth character on page 1187. A character not in this dictionary but assigned a position between the 6th and 7th characters on page 1187 for sorting purposes would have the code “1187.061”. The edition of the Kang Xi Dictionary used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989.

Property	**kIRG_GSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	G(4K|BK|CH|CY|FZ|HC|HZ|((BK|CH|GH|HC|XC|ZH)-[0-9]{4}\.[0-9]{2})|HZ-[0-9]{5}\.[0-9]{2}|(KX-[01][0-9]{3}\.[0-9]{2})|((CYY|FZ|JZ|ZFY|ZJW)-[0-9]{5})|([0135789ES]-[0-9A-F]{4})|

	(IDC-[0-9]{3})|(K-[0-9A-F]{4})|(H-\d{4})
Description	The IRG “G” source mapping for this character in hex. 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 GB2312-80 G1 GB12345-90 with 58 Hong Kong and 92 Korean “Idu” characters G3 GB7589-87 unsimplified forms G5 GB7590-87 unsimplified forms G7 General Purpose Hanzi List for Modern Chinese Language, and General List of Simplified Hanzi GS Singapore Characters G8 GB8565-88 G9 GB18030-2000 GE GB16500-95 G4K Siku Quanshu (四庫全書) GBK Chinese Encyclopedia (中國大百科全書) GCH Ci Hai (辞海) GCY Ci Yuan (辭源) GCYY Chinese Academy of Surveying and Mapping Ideographs (中国测绘科学院用字) GFZ Founder Press System (方正排版系统) GGH Gudai Hanyu Cidian (古代汉语词典) GHC Hanyu Dacidian (漢語大詞典) GHZ Hanyu Dazidian ideographs (漢語大字典) GIDC ID system of the Ministry of Public Security of China, 2009 GJZ Commercial Press Ideographs (商务印书馆用字) GKX Kangxi Dictionary ideographs(康熙字典)9th edition (1958) including the addendum (康熙字典)補遺 GXC Xiandai Hanyu Cidian (现代汉语词典) GZFY Hanyu Fangyan Dacidian (汉语方言大辞典) GZH ZhongHua ZiHai (中华字海) GZJW Yinzhou Jinwen Jicheng Yinde (殷周金文集成引得)

Property	**kIRG_HSource**
Status	Normative
Category	IRG Sources
Introduced	3.1
Delimiter	N/A
Syntax	H((3)|(B[012]))?-[0-9A-F]{4}
Description	The IRG “H” source mapping for this character in hex. The IRG “H” source consists of data from the following sources: H, H3 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

Property	**kIRG_JSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	J((([0134AK]|3A|ARIB)-[0-9A-F]{4,5})|(H-(((IB|JT|[0-9]{2})[0-9A-F]{4}S?))))
Description	The IRG “J” source mapping for this character in hex. 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 J3 JIS X 0213:2000 level-3 J3A JIS X 0213:2004 level-3 J4 JIS X 0213:2000 level-4 JA Unified Japanese IT Vendors Contemporary Ideographs, 1993 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 JARIB Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

Property	**kIRG_KPSource**
Status	Normative
Category	IRG Sources
Introduced	3.1.1
Delimiter	N/A
Syntax	KP[01]-[0-9A-F]{4}
Description	The IRG “KP” source mapping for this character in hex. 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

Property	**kIRG_KSource**
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	K[0-57]-[0-9A-F]{4}
Description	The IRG “K” source mapping for this character in hex. 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 PKS C 5700-1 1994 K3 PKS C 5700-2 1994 K4 PKS 5700-3:1998 K5 Korean IRG Hanja Character Set 5th Edition: 2001

	Note that the K4 source is expressed in hexadecimal, but unlike the other sources, it is not organized in row/column. The content of the repertoire covered by the K2, K3, K4, and K5 sources is in the process of being reedited in new Korean standards.

Property	kIRG_MSource
Status	Normative
Category	IRG Sources
Introduced	5.2
Delimiter	N/A
Syntax	MAC-[0-9]{5}
Description	The IRG “M” source mapping for this character. The IRG “M” source consists of data from the Macao Information System Character Set (澳門資訊系統字集).

Property	kIRG_TSource
Status	Normative
Category	IRG Sources
Introduced	3.0
Delimiter	N/A
Syntax	T[1-7B-F]-[0-9A-F]{4}
Description	The IRG “T” source mapping for this character in hex. The IRG “T” source consists of data from the following national standards and lists from the Republic of China (Taiwan). T1 TCA-CNS 11643-1992 1st plane T2 TCA-CNS 11643-1992 2nd plane T3 TCA-CNS 11643-1992 3rd plane with some additional characters 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 TB TCA-CNS Ministry of Education, Hakka dialect, May 2007 TC TCA-CNS 11643-1992 12th plane

TD TCA-CNS 11643-1992 13th plane
TE TCA-CNS 11643-1992 14th plane
TF TCA-CNS 11643-1992 15th plane

CNS 11643, X 5012 (p.3) 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. http://ap6.pccu.edu.tw/Dictionary/]
(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.

Property	kIRG_USource
Status	Normative
Category	IRG Sources
Introduced	4.0.1
Delimiter	N/A
Syntax	U(TC|CI)-[0-9]{5}
Description	The IRG “U” source mapping for this character. U-source references are a reference into the U-source ideograph database; see UTR #45. These consist of “UTC” or “UCI” 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}
Description	The IRG “V” source mapping for this character in hex. 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 Dictionary on Nom 2006, Dictionary on Nom of Tay ethnic 2006, Lookup Table for Nom in the South 1994

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

Description	The Japanese pronunciation(s) of this character.

Property	**kJapaneseOn**
Status	Provisional
Category	Readings
Introduced	2.0
Delimiter	space
Syntax	[A-Z]+
Description	The Sino-Japanese pronunciation(s) of this character.

Property	**kJis0**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The JIS X 0208-1990 mapping for this character in ku/ten form.

Property	**kJIS0213**
Status	Provisional
Category	Other Mappings
Introduced	3.1.1
Delimiter	space
Syntax	[12],[0-9]{2},[0-9]{1,2}
Description	The JIS X 0213:2004 mapping for this character in men,ku,ten form.

Property	**kJis1**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}

Description	The JIS X 0212-1990 mapping for this character in ku/ten form.

Property	**kKangXi**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}\.[0-9]{2}[01]
Description	The position of this character in the 《康熙字典》 Kang Xi 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 characters actually in the dictionary and “1” for characters not found in the dictionary but assigned a “virtual” position in the dictionary. Thus, “1187.060” indicates the sixth character on page 1187. A character not in this dictionary but assigned a position between the 6th and 7th characters on page 1187 for sorting purposes would have the code “1187.061”. The edition of the Kang Xi Dictionary used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989.

Property	**kKarlgren**
Status	Provisional
Category	Dictionary Indices
Introduced	3.1.1
Delimiter	space
Syntax	[1-9][0-9]{0,3}[A*]?
Description	The index of this character 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 character 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 character, using the Yale romanization system. (See <http://en.wikipedia.org /wiki/Korean_romanization> for a discussion of the various Korean romanization systems.)

Property	**kKPS0**
Status	Provisional
Category	Other Mappings
Introduced	3.1.1
Delimiter	space
Syntax	[0-9A-F]{4}
Description	The KPS 9566-97 mapping for this character in hexadecimal form.

Property	**kKPS1**
Status	Provisional
Category	Other Mappings
Introduced	3.1.1
Delimiter	space
Syntax	[0-9A-F]{4}
Description	The KPS 10721-2000 mapping for this character in hexadecimal form.

Property	**kKSC0**

Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The KS X 1001:1992 (KS C 5601-1989) mapping for this character in ku/ten form.

Property	**kKSC1**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The KS X 1002:1991 (KS C 5657-1991) mapping for this character in ku/ten form.

Property	**kLau**
Status	Provisional
Category	Dictionary Indices
Introduced	3.1.1
Delimiter	space
Syntax	[1-9][0-9]{0,3}
Description	The index of this character in A Practical Cantonese-English Dictionary by Sidney Lau, Hong Kong: The Government Printer, 1977. The index consists of an integer. Missing indices indicate unencoded characters which are being submitted to the IRG for inclusion in future versions of the standard.

Property	**kMainlandTelegraph**
Status	Provisional

Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The PRC telegraph code for this character, derived from “Kanzi denpou koudo henkan-hyou” (“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 pinyin reading for this character. 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. ~~Multiple Value Order: 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.~~

Property	**kMatthews**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,3}(a|\.5)?
Description	The index of this character in Mathews’ Chinese-English Dictionary by Robert H. Mathews, Cambrige: Harvard University Press, 1975.

	Note that the field 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]{0,3}[a-t*]?
Description	The index of this character 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 lower-case Latin letter if the listing is in a subsidiary entry and not a main one. In some cases where the character is found in the radical-stroke index, but not in the main body of the dictionary, the integer is followed by an asterisk (e.g., U+50E5, which is listed as 736* as well as 1185a).

Property	**kMorohashi**
Status	Provisional
Category	Dictionary Indices
Introduced	2.0
Delimiter	space
Syntax	[0-9]{5}\'?
Description	The index of this character in the Dai Kanwa Ziten, aka Morohashi dictionary (Japanese) used in the four-dictionary sorting algorithm. The edition used is the revised edition, published in Tokyo by Taishūkan Shoten, 1986.

Property	**kNelson**
Status	Provisional
Category	Dictionary Indices

Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The index of this character 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]+
Description	The numeric value for the character in certain unusual, specialized contexts. The three numeric-value fields should have no overlap; that is, characters 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]{0,3}[A-D]?*?
Description	The phonetic index for the character from _Ten Thousand Characters: An Analytic Dictionary_, by G. Hugh Casey, S.J. Hong Kong: Kelley and Walsh, 1980.

Property	**kPrimaryNumeric**
Status	Informative
Category	Numeric Values
Introduced	3.2

Delimiter	space
Syntax	[0-9]+
Description	The value of the character when used in the writing of numbers in the standard fashion. The three numeric-value fields should have no overlap; that is, characters 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]{4}
Description	A “GB 12345-90” code point assigned to this character for the purposes of including it within Unihan. Pseudo-GB1 codes were used to provide official code points for characters not already in national standards, such as characters 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]{1,5}\+[1-9][0-9]{0,2}\.[1-9][0-9]?\.[0-9]{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	**kRSJapanese**
Status	Provisional
Category	Radical-Stroke Counts
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,2}\.[0-9]{1,2}
Description	The Japanese radical/stroke count for this character in the form “radical.additional strokes”.

Property	**kRSKangXi**
Status	Provisional
Category	Radical-Stroke Counts
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,2}\.[0-9]{1,2}
Description	The KangXi radical/stroke count for this character consistent with the value of the kKangXi field in the form “radical.additional strokes”.

Property	**kRSKanWa**
Status	Provisional
Category	Radical-Stroke Counts

Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,2}\.[0-9]{1,2}
Description	The Morohashi radical/stroke count for this character in the form “radical.additional strokes”.

Property	**kRSKorean**
Status	Provisional
Category	Radical-Stroke Counts
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,2}\.[0-9]{1,2}
Description	The Korean radical/stroke count for this character in the form “radical.additional strokes”.

Property	**kRSUnicode**
Status	Informative
Category	Radical-Stroke Counts
Introduced	2.0
Delimiter	space
Syntax	[1-9][0-9]{0,2}\'?\.[0-9]{1,2}
Description	The standard radical/stroke count for this character in the form “radical.additional strokes”. The radical is indicated by a number in the range (1..214) inclusive. An apostrophe (') after the radical indicates a 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 field is also used for additional radical-stroke indices where either a character may be reasonably classified under more than one radical, or alternate stroke count algorithms may provide different stroke counts. The first value is equal to the normative radical-stroke value

	defined in ISO/IEC 10646.

Property	**kSBGY**
Status	Provisional
Category	Dictionary Indices
Introduced	3.2
Delimiter	space
Syntax	[0-9]{3}\.[0-7][0-9]
Description	The position of this character in the Song Ben Guang Yun (SBGY) Medieval Chinese character dictionary (bibliographic and general information below). The 25334 character 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 character on the page [01..73]. For example, 364.38 indicates the 38th character on Page 364 (i.e. 澍). Where a given Unicode Scalar Value (USV) 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 25334 references, for 19583 different hanzi. -- Release information (20031005) -- This release corrects several mappings. -- Release information (20020310) -- This data set contains a total of 25334 references, for 19572 different hanzi (up from 25330 and 19511 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 (20020310) --

The original data was input under the direction of Prof. 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 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 USV 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 character 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\+2?[0-9A-F]{4}(<k[A-Za-z0-9]+(:[TBZFJ]+)?(,k[A-Za-z0-9]+(:[TBZFJ]+)?)*)?
Description	The Unicode value for a semantic variant for this character. A semantic variant is an x- or y-variant with similar or identical meaning which can generally be used in place of the indicated character. 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 field tags for another field in the Unihan database indicating the source of the information. If subdivided, 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 (e.g., by saying that the one character 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 character 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 character is the traditional form. J is used if the source explicitly indicates that the given character is the simplified form. Data on simplified and traditional variations can be included in this

	field to document cases where different sources disagree on the nature of the relationship between two characters. The kSemanticVariant and kSpecializedSemanticVariant fields need not be consulted when interconverting between traditional and simplified Chinese.

Property	**kSimplifiedVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+2?[0-9A-F]{4}
Description	The Unicode value(s) for the simplified Chinese variant(s) for this character. A full discussion of the kSimplifiedVariant and kTraditionalVariant fields is found in section 3.7.1 above. Much of the of the data on simplified and traditional variants was graciously supplied by Wenlin Institute, Inc. <http://www.wenlin.com>.

Property	**kSpecializedSemanticVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+2?[0-9A-F]{4}(<k[A-Za-z0-9]+(:[TBZFJ]+)?(,k[A-Za-z0-9]+(:[TBZFJ]+)?)*)?
Description	The Unicode value for a specialized semantic variant for this character. The syntax is the same as for the kSemanticVariant field. A specialized semantic variant is an x- or y-variant with similar or identical meaning only in certain contexts (such as accountants' numerals).

Property	**kTaiwanTelegraph**
Status	Provisional
Category	Other Mappings
Introduced	2.0
Delimiter	space
Syntax	[0-9]{4}
Description	The Taiwanese telegraph code for this character, derived from “Kanzi denpou koudo henkan-hyou” (“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 character, derived from or consistent with _T’ang Poetic Vocabulary_ by Hugh M. Stimson, Far Eastern Publications, Yale Univ. 1976. An asterisk indicates that the word or morpheme represented in toto or in part by the given character with the given reading occurs more than four times in the seven hundred poems covered.

Property	**kTotalStrokes**
Status	Informative
Category	Dictionary-like Data
Introduced	3.1
Delimiter	space
Syntax	[1-9][0-9]{0,2}
Description	The total number of strokes in the character (including the radical). 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 character in modern text using customary fonts.

Property	**kTraditionalVariant**
Status	Provisional
Category	Variants
Introduced	2.0
Delimiter	space
Syntax	U\+2?[0-9A-F]{4}
Description	The Unicode value(s) for the traditional Chinese variant(s) for this character. A full discussion of the kSimplifiedVariant and kTraditionalVariant fields is found in section 3.7.1 above. Much of the of the data on simplified and traditional variants was graciously supplied by Wenlin Institute, Inc. <http://www.wenlin.com>.

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 character's pronunciation(s) in Quốc ngữ.

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

Property	kXHC1983
Status	Provisional
Category	Readings
Introduced	5.1
Delimiter	space
Syntax	[0-9]{4}\.[0-9]{3}*?(,[0-9]{4}\.[0-9]{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 character's location(s) in the dictionary, separated from the reading by ":" (colon); multiple locations for a given reading are separated by "," (comma); multiple "location: reading" values are separated by " " (space). Each location reference is of the form /[0-9]{4}\.[0-9]{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 an encoded variant substituted for an unencoded character (see below). -- 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 there are subsequent editions of this important PRC dictionary, reflecting later developments and refinements in language and orthographic standardization, and other editions

	should not be used in future revision of this field. -- 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). Distinct Unihan hànzì: 10,992; Distinct hànzì: 11,190; Distinct pīnyīn syllable types: 1,337; As of the present writing (Unicode 5.1), the XHC source data contains 204 unencoded characters (198 of which were represented by PUA or CJK Compatibility [or in one case, by non-CJK, see below] characters), for the most part simplified variants. Each of these 198 characters in the source is replaced by one or more encoded variants (references in all 204 cases are marked with a trailing “*”; see above). Many of these unencoded forms are already in the pipeline for future encoding, and future revisions of this data will eliminate trailing asterisks from mappings. The print source and data also include a lexical entry ○ U+3007 : “0719.100: líng” (IDEOGRAPHIC NUMBER ZERO) which is currently excluded from Unihan data (as not being a CJK Unified Ideograph); see 零 U+96F6.

Property	**kZVariant**
Status	Provisional
Category	Variants
Introduced	2.0

Delimiter	space
Syntax	U\+2?[0-9A-F]{4}(<k[A-Za-z0-9]+(:[TBZ]+)?(,k[A-Za-z0-9]+(:[TBZ]+)?)*)?
Description	The Unicode value(s) for known z-variants of this character. 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 field tags for another field in the Unihan database indicating the source of the information.

4.2 Listing by Date of Addition to the Unicode Standard

The table below lists the fields of the Unihan database by the release where they were first added. Also included are fields which were dropped in a particular release. These are indicated by italics.

Unicode Version	Fields Added or Dropped
5.2	kHanyuPinyin, kIRG_MSource
5.1	kXHC1983
5.0	kCheungBauer, kCheungBauerIndex, kFourCornerCode, kHangul
4.1	kAlternateKangXi (dropped), kAlternateMorohashi (dropped), kFennIndex, kIICore, kRSAdobe_Japan1_6
4.0.1	kGSR, kHanyuPinlu, kIRG_USource
3.2	kAccountingNumeric, kAlternateHanYu (dropped), kCihaiT, kCompatibilityVariant, kFrequency, kGradeLevel, kOtherNumeric, kPrimaryNumeric, kSBGY
3.1.1	kCangjie, kCowles, kFenn, kHKGlyph, kHKSCS, kIRG_KPSource, kJIS0213, kKPS0, kKPS1, kKarlgren, kLau, kVietnamese
3.1	kAlternateJEF (dropped), kIRG_HSource, kMeyerWempe, kPhonetic, kRSMerged (dropped), kTotalStrokes
3	kAlternateJEF, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kIRG_GSource, kIRG_JSource, kIRG_KSource, kIRG_TSource, kIRG_VSource, kRSMerged,

	kSemanticVariant (reintroduced), kSpecializedSemanticVariant (reintroduced)
2.1	kSemanticVariant (dropped), kSpecializedSemanticVariant (dropped)
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 fields were added prior to Unicode 2.0.

4.3 Listing by Location within Unihan.zip

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

File	Fields within file
Unihan_DictionaryIndices.txt	kCheungBauerIndex, kCowles, kDaeJaweon, kFennIndex, kGSR, kHanYu, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kKangXi, kKarlgren, kLau, kMatthews, kMeyerWempe, kMorohashi, kNelson, kSBGY
Unihan_DictionaryLikeData.txt	kCangjie, kCheungBauer, kCihaiT, kFenn, kFourCornerCode, kFrequency, kGradeLevel, kHDZRadBreak, kHKGlyph, kPhonetic, kTotalStrokes
Unihan_IRGSources.txt	kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_USource, kIRG_VSource, kIRG_MSource
Unihan_NumericValues.txt	kAccountingNumeric, kOtherNumeric, kPrimaryNumeric
Unihan_OtherMappings.txt	kBigFive, kCCCII, kCNS1986, kCNS1992, kEACC, kGB0, kGB1, kGB3, kGB5, kGB7, kGB8,

	kHKSCS, kIBMJapan, kJis0, kJis1, kJIS0213, kKPS0, kKPS1, kKSC0, kKSC1, kMainlandTelegraph, kPseudoGB1, kTaiwanTelegraph, kXerox
Unihan_RadicalStrokeCounts.txt	kRSAdobe_Japan1_6, kRSJapanese, kRSKangXi, kRSKanWa, kRSKorean, kRSUnicode
Unihan_Readings.txt	kCantonese, kDefinition, kHangul, kHanyuPinlu, kHanyuPinyin, kJapaneseKun, kJapaneseOn, kKorean, kMandarin, kTang, kVietnamese, kXHC1983
Unihan_Variants.txt	kCompatibilityVariant,kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTraditionalVariant, kZVariant

4.4 Listing of Characters Covered by the Unihan Database

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

Code Points	Block Name	Unicode Version
U+3400…U+4DB5	CJK Unified Ideographs Extension A	3.0
U+4E00…U+9FA5	CJK Unified Ideographs	1.1
U+9FA6…U+9FBB	CJK Unified Ideographs	4.1
U+9FBC…U+9FC3	CJK Unified Ideographs	5.1
U+9FC4…U+9FCB	CJK Unified Ideographs	5.2
U+9FCC	CJK Unified Ideographs	6.1
U+F900…U+FA2D	CJK Compatibility Ideographs N.B., 12 code points in this range (U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, and U+FA29) lack a canonical Decomposition_Mapping value in UnicodeData.txt and so are not	1.1

	true CJK Compatibility Ideographs. These twelve characers should be treated as proper CJK Unified Ideographs.	
U+FA2E…U+FA2F	CJK Compatibility Ideographs	6.1
U+FA30…U+FA6A	CJK Compatibility Ideographs	3.2
U+FA6B…U+FA6D	CJK Compatibility Ideographs	5.2
U+FA70…U+FAD9	CJK Compatibility Ideographs	4.1
U+20000…U+2A6D6	CJK Unified Ideographs Extension B	3.1
U+2A700…U+2B734	CJK Unified Ideographs Extension C	5.2
U+2B740…U+2B81D	CJK Unified Ideographs Extension D	6.0
U+2F800…U+2FA1D	CJK Compatibility Supplement	3.1

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

Code Points	Block Name	Unicode Version
U+2E80…U+2E99	CJK Radicals Supplement	3.0
U+2E9B…U+2EF3	CJK Radicals Supplement	3.0
U+2F00…U+2FD5	Kangxi Radicals	3.0
U+2FF0…U+2FFB	Ideographic Description Characters	3.0
U+3000…U+3037	CJK Symbols and Punctuation	1.1
U+3038…U+303A	CJK Symbols and Punctuation	3.0
U+303B…U+303D	CJK Symbols and Punctuation	3.2
U+303E	CJK Symbols and Punctuation	3.0
U+303F	CJK Symbols and Punctuation	1.1
U+3105…U+312C	Bopomofo	1.1
U+312D	Bopomofo	5.1
U+3190…U+319F	Kanbun	1.1
U+31A0…U+31B7	Bopomofo Extended	3.0
U+31C0…U+31CF	CJK Strokes	4.1
U+31D0…U+31E3	CJK Strokes	5.1
U+3220…U+3243	Enclosed CJK Letters and Months	1.1

U+3280…U+32B0	Enclosed CJK Letters and Months	1.1
U+32C0…U+32CB	Enclosed CJK Letters and Months	1.1
U+3358…U+3370	CJK Compatibility	1.1
U+337B…U+337F	CJK Compatibility	1.1
U+33E0…U+33FE	CJK Compatibility	1.1

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 fields 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 ennumerated type for the numeric field tags, and the ennumerator names (each beginning with a "k" indicating their use as a constant) were used in the text file as field names.

Unihan-1.txt was at some point accidentally truncated on line 330,553 (partway through the data for U+8BC1). 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 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).

References

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

Modifications

This section indicates the changes introduced by each revision.

Revision 14

- **Proposed update** for Unicode 6.3.0.
- Clarified the status of kCompatibilityVariant.
- Expanded the description of kEACC and kRSUnicode.
- Altered kHanuPinlu to use accents instead of numbers for tones (with concomitant changes to its regular expression).
- Changed the delimiter to "space" for the following fields: kAccountingNumeric, kIBMJapan, kIICore, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kJis0, kJIS0213, kJis1, kKangXi, kKarlgren, kKPS0, kKPS1, kKSC0, kKSC1, kMainlandTelegraph, kMatthews, kOtherNumeric, kPrimaryNumeric, kTaiwanTelegraph, and kXerox.
- Removed multiple value order notes for kHanyuPinyin, kMandarin and kTotalStrokes.
- Several minor clarifications and rewording of field descriptions.

Revision 13

- **Reissued** for Unicode 6.2.0.
- Updated the regex syntax fields for kCNS1992, kIRG_GSource, and kIRG_HSource.
- Added note re CJK unified ideographs in the CJK compatibility ideographs range in Section 4.4.

Revision 12 being a proposed update, only changes between revisions 11 and 13 are noted here.

Revision 11

- **Reissued** for Unicode 6.1.
- Updated regular expressions and removed explicit start and end markers.
- Redefined the kTotalStrokes and kMandarin fields.
- Clarified the use of the kTraditionalVariant and kSimplifiedVariant fields.
- Added section 4.4 Listing of Characters Covered by the Unihan Database
- Clarified the order used for multiple values of a single field.

Revision 10 being a proposed update, only changes between revisions 9 and 11 are noted here.

Revision 9

- **Reissued** for Unicode 6.0.0
- Clarified the nature of the contents of http://www.unicode.org/charts/unihan.html
- Added the History section.

- Altered the synax and source information for the IRG source fields to match current ISO/IEC 10646 values.
- Improved other regular expressions used to describe database field syntax and fixed some minor typographical errors in the field descriptions.
- The description of the kTang field was extended and that for the kHanyuPinlu field changed to reflect an extended source corpus. Other typographical errors in field descriptions were corrected.
- Clarified the nature of Japanese on and kun readings.

Revision 8 being a proposed update, only changes between revisions 7 and 9 are noted here.

Revision 7

- Update for Unicode 5.2.0
- Reclassified kDefinition, kHanyuPinlu, and kXHC1983 fields as Readings.
- Removed use of the Linguistic Society of Hong Kong's Jyutping Phrase Box because of viral licensing issues.
- Documented revised structure of Unihan.zip.
- Reformatted tabular listing of tags.
- Added links to tag descriptions in index tables.
- Updated regular expressions of tags.

Revision 6 being a proposed update, only changes between revisions 5 and 7 are noted here.

Revision 5

- First approved version, for Unicode 5.1.0.

Revision 4

- Upgrade from Proposed Draft to Draft.
- Correct syntax for a number of regular expressions.

Revision 3

- Changes per UTC input.

Revision 2

- Rewrite for Unicode 5.0.

Revision 1

- First working draft
