Technical Reports

L2/09-133

Proposed Update Unicode Standard Annex #38

UNICODE HAN DATABASE (UNIHAN)

Version	5.2.0
Authors	John H. Jenkins 井作恆 Richard Cook 曲理查
Date	2009-01-29
This Version	http://www.unicode.org/reports/tr38/tr38-6.html
Previous Version	http://www.unicode.org/reports/tr38/tr38-5.html
Latest Version	http://www.unicode.org/reports/tr38/
Revision	6

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*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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 radical, 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.

The Unihan database is available to the public in two forms: one, as a text file, Unihan.txt, which is distributed as part of the Unicode Standard; two, via the World Wide Web on the Unicode Web site. The text file is the best indication of the overall size of the database, consisting as it does of twenty-nine megabytes of data with over one million lines, covering over 71,000 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. The Unihan database is truly a work in progress, with new data (and even new fields) being added on a regular basis. If the Unihan database has a weakness, however, it is that it is maintained by volunteers. Nobody is paid to work on it. There is a great deal of useful information which would be added if only someone would provide it. Despite this, the Unihan database provides solid, useful data for everyday implementation needs and beyond.

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 reflections 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 puncutation except for underscore. For historical reasons, they all start with a lower-case “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 purposes or are fields which are in the process of being filled and which will be made public in a future release. The remaining two private fields are convenience fields only; since 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+4DBF are mapped to 0x05200 through 0x06BBF; that is, 0x1E00 is added to the Unicode Scalar Value.
  - U+20000 through U+2A6DF are mapped to 0x06C00 through 0x112DF; 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+2FA1F are mapped to 0x1F800 through 0x1FA1F; 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, Compatibility Ideographs, Compatibility Extension), and to leave a gap of over 58,000 code points between the end of Extension B and the first Compatibility Ideographs block, and over 12,000 after 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.

2.2 Web Access

The URI for accessing the Unihan database via the Web is http://www.unicode.org/charts/unihan.html.

Chinese and Japanese compound data are presented in the online database and come from the online CEDICT and Jim Breen’s EDICT projects. These additional data are not available in the other versions.

There are also two indices for the database, 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 “UTF-8” check-box allows you to control whether UTF-8 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.

2.3 Unihan.txt

The final form of the Unihan database is the `Unihan.txt` file. This is the only version included in beta releases of the Unicode Standard.

The released version of `Unihan.txt` consists of a header followed by data. Unix line breaks are used. Each line in the data section 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 the `Unihan.txt` file.

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 more 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.

The header contains detailed information on the database and its contents, including the specific syntax for and contents of individual fields.

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 eight 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_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.

At the moment, the U-source consists only of the Unicode Standard itself, and the field value is always equal to the character's Unicode Scalar Value. This will change when Extension C1 is formally encoded, because Extension C1 contains a small number of characters submitted by the Unicode Technical Committee which use a different indexing system.

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 indicates that a 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 other encoded character sets or character collections. Some of the character sets covered mirror official IRG sources. For example, we have data for mapping GB 12345, which 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 extension 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 (e.g., 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 range, is spoken throughout Guangdong Province and in much of neighboring Guangxi, and covers two large urban centers (Guangzhou 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 social context. For example, in Cantonese, the -ing and -eng finals are fairly interchangeable, with the former preferred in more formal settings, and the latter having a distinct colloquial feel.

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 *kun* pronunciation for a single kanji can easily be polysyllablic (e.g., *ichi* for 一). In essence, the *on* pronunciation is the Japanese way of pronouncing the Chinese word, whereas the *kun* pronunciation is the Japanese translation of the word.

Moreover, some characters have rare pronunciations known only to a minority of even native speakers, or are so rare themselves that few, if any, native speakers know how to pronounce them (e.g., U+40DF 䃟, 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 characters 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 characters' 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.

The fields included in this category are `kCangjie`, `kCheungBauer`, `kFenn`, `kFourCornerCode`, `kFrequency`, `kGradeLevel`, `kHanyuPinlu`, `kHKGlyph`, `kPhonetic`, and `kXHC1983`. Note that in the case of `kFenn`, `kCheungBauer`, and `kHanyuPinlu`, the data is named for the dictionary from which the data is derived, not for the type of data it is.

3.6 Radical-Stroke Counts

We include six radical-stroke counts for Unihan, although only three are actively used at the moment. 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 (e.g., 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 (e.g., 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 form of the character as drawn in the Unicode Standard. 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 173 candidates in Unicode 4.0.1, 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, since 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 Unihan database also includes the `kCompatibilityVariant` field, which marks compatibility variants as defined by the Unicode Standard.

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

The `kSimplifiedVariant` and `kTraditionalVariant` fields are used to aid in the process of going between simplified and traditional Chinese. The People’s Republic of China, beginning in the 1950’s, undertook a series of language reforms aimed at boosting literacy by making Chinese easier to read and write, largely by reducing the number of strokes needed to write a number of characters. These reforms have also been adopted in Singapore. The traditional forms, however, are predominant in Taiwan, overseas Chinese communities, and even in China’s two Special Administrative Regions, Hong Kong and Macao.

The mapping between simplified and traditional Chinese can be quite complex. In many cases, the official simplification is an acceptable alternative even within traditional Chinese, as with our two cats above: both 猫 and 貓 are acceptable in traditional Chinese, but only 猫 is used in simplified Chinese. In a few cases, a single simplified form corresponds to multiple traditional forms, such as 台, which is not only a traditional character in its own right, but also the simplification for 檯,

臺, and 颱. And a character-by-character conversion isn't sufficient to convert between simplified and traditional Chinese because of lexical differences. A hard disk, for example, is called 硬磁盤 in the PRC, and 硬碟 in Taiwan.

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 礟 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 *tag*, its *Unicode status*, its *category* as defined above, its *level of completion*, its *separator*, its *syntax*, and its *description*.

The *tag* is the tag used in the `Unihan.txt` file and MySQL 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 also include *deprecated* as a Unicode status if the field is no longer to be used.

We use three values for the *level of completion*.

A *complete* field is one which could not cannot contain more data. The `kMatthews` field is complete, for example, because there are only so many characters in Mathews' dictionary, and all of them which are encoded have their indices in the Unihan database.

We also include as complete fields such as `kKangXi` which are Unicode counterparts to IRG data fields which are truly complete.

An *extendable* field is one which is complete as far as it goes. Fields such as the readings and definition fields are extendable, because they could theoretically be extended to cover all of Unihan, but they are sufficiently complete for most needs and we are unlikely to need to go back and revise existing data in the field.

An *incomplete* field is one which has known gaps and needs more data before it can be truly useable.

Fields which allow multiple values have a *separator* defined, usually a space. Fields which do not need or cannot have a separator do not have this defined, such as the IRG source fields.

The *syntax* is a Perl-like regular expression describing the formal structure of an individual entry in the field. The syntax for the `kKangXi` field is `[0-9]{4}\.[0-9]{2}[01]`, which means four decimal digits followed by a period, followed by two more decimal digits, followed by a zero or a one. The syntax can be used to validate the contents of a field. (Note: We may have to do more than just regexps for some fields. We'll see when we get to them.) Of course, just complying with the formal syntax is no guarantee that the data is correct: a `kKangXi` value of 9999.990 is syntactically correct but wrong anyway, since there is no page 9999 in the KangXi dictionary.

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.

Tag:	`kAccountingNumeric`
Status:	Informative
Category:	Numeric Values
Separator:	`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.

Tag: `kBigFive`
Status: Provisional
Category: Other Mappings
Separator: `space`
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.

Tag: `kCCCII`
Status: Provisional
Category: Other Mappings
Separator: `space`
Syntax: `[0-9A-F]{6}`
Description: The CCCII mapping for this character in hex.

Tag: `kCNS1986`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[12E]-[0-9A-F]{4}`
Description: The CNS 11643-1986 mapping for this character in hex.

Tag: `kCNS1992`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[123]-[0-9A-F]{4}`
Description: The CNS 11643-1992 mapping for this character in hex.

Tag: `kCangjie`
Status: Provisional
Category: Dictionary-like Data
Separator: space
Syntax: `[A-Z]+`

Description: The *cangjie* input code for the character. This incorporates data from the file `cangjie-table.b5` by Christian Wittern.

Tag: `kCantonese`

Status: Provisional

Category: Readings

Separator: space

Syntax: `[a-z]+[1-6]`

Description: The Cantonese pronunciation(s) for this character using the *jyutping* romanization.

A full description of *jyutping* can be found at http://cpct92.cityu.edu.hk/lshk/Jyutping/Jyutping.htm. 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.

The *jyutping* phrase box from the Linguistic Society of Hong Kong, http://cpct92.cityu.edu.hk/lshk/Jyutping/ . The copyright of the *jyutping* phrase box belongs to the Linguistic Society of Hong Kong. We would like to thank the Jyutping Group of the Linguistic Society of Hong Kong for permission to use the electronic file in our research and/or product development. Note that the inclusion of the phrase box in the Unihan database requires that any products developed using the `kCantonese` field needs to include this acknowledgment.

Tag:	`kCheungBauer`
Status:	Provisional
Category:	Dictionary-like Data
Separator:	NA
Introduced:	5.0
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. The data consist 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.

Tag:	`kCheungBauerIndex`
Status:	Provisional
Category:	Dictionary Indices
Separator:	space

Syntax: [0-9]{3}\.[0-9]{2}

Introduced: 5.0

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.

Tag: kCihaiT

Status: Provisional

Category: Dictionary-like Data

Separator: 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.

Tag: kCompatibilityVariant

Status: Normative

Category: Variants

Separator: space

Syntax: U\+2?[0-9A-F]{4}

Description: The compatibility decomposition for this ideograph, derived from the `UnicodeData.txt` file.

Tag: kCowles

Status: Provisional

Category: Dictionary Indices

Separator: 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.

Tag: `kDaeJaweon`

Status: Provisional

Category: Dictionary Indices

Separator: space

Syntax: `[0-9]{4}\.[0-9]{2}[0158]`

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.

Tag: `kDefinition`

Status: Provisional

Category: Dictionary-like Data

Separator: space

Syntax: `See Description`

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.

Tag: `kEACC`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[0-9A-F]{6}`
Description: The EACC mapping for this character in hex.

Tag: `kFenn`
Status: Provisional
Category: Dictionary-like Data
Separator: 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.

Tag: `kFennIndex`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[1-9][0-9]{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.

Tag: `kFourCornerCode`
Status: Provisional
Category: Dictionary-like Data
Separator: 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.

Tag: `kFrequency`
Status: Provisional
Category: Dictionary-like Data
Separator: space
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.

Tag: `kGB0`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[0-9A-F]{4}`
Description: The GB 2312-80 mapping for this character in ku/ten form.

Tag:	`kGB1`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The GB 12345-90 mapping for this character in ku/ten form.

Tag:	`kGB3`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The GB 7589-87 mapping for this character in ku/ten form.

Tag:	`kGB5`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The GB 7590-87 mapping for this character in ku/ten form.

Tag:	`kGB7`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The GB 8565-89 mapping for this character in ku/ten form.

Tag:	`kGB8`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9]{4}`
Description:	The GB 8565-89 mapping for this character in ku/ten form

Tag: `kGSR`

Status: Provisional

Category: Dictionary Indices

Separator: 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 7,403 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

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, & images of the original Karlgren (1957) text. Prepared for the STEDT Project by Richard Cook; based in part on work by Tor Ulving & Ferenc Tafferner (see below), used by permission. Berkeley: University of California, http://stedt.berkeley.edu/

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.

Tag: `kGradeLevel`
Status: Provisional
Category: Dictionary-like Data
Separator: space
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.

Tag: `kHDZRadBreak`
Status: Provisional
Category: Dictionary-like Data
Separator: NA
Syntax: `[\x{2F00}-\x{2FD5}][U+2?[0-9A-F]{4}]:[1-8][0-9]{4}\.[0-9]{2}[012]`
Introduced: 4.1
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.

Tag: `kHKGlyph`
Status: Provisional
Category: Dictionary-like Data
Separator: 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.

Tag: `kHKSCS`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[0-9A-F]{4}`

Description: Mappings to the Big Five extended code points used for the Hong Kong Supplementary Character Set.

Tag: `kHangul`
Status: Provisional
Category: Readings
Separator: space
Syntax: `[\x{AC00}-\x{D7AF}]+`
Description: The modern Korean pronunciation(s) for this character in Hangul.

Tag: `kHanYu`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[1-8][0-9]{4}\.[0-9]{2}[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 3,044 of Volume 5 (i.e. 籉). 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 56097 records, 54728 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).

All 55817 HDZ references in this data set are unique. Because of IRG source-internal unifications, a given UCS-4 Scalar Value (USV) 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 Unihan.txt (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 Unihan.txt (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.

Tag:	`kHanyuPinlu`
Status:	Provisional
Category:	Dictionary Indices
Separator:	space
Syntax:	`[a-zü]+[1-5]\([0-9]+\)`
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 3800 records. 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 "comma + 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 440799 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" PRC text) was (Mac OS 7-9) GB 2312-80 (plus 啫).

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 current 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.

Tag: `kIBMJapan`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `F[ABC][0-9A-F]{2}`
Description: The IBM Japanese mapping for this character in hexadecimal.

Tag: `kIICore`
Status: Normative
Category: Dictionary-like Data
Separator: space
Syntax: `[1-9]\.[1-9]`
Description: Indicates that a 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.

Tag: `kIRGDaeJaweon`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[0-9]{4}\.[0-9]{2}[01]|0000\.555`
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.

Tag: `kIRGDaiKanwaZiten`

Status: Provisional

Category: Dictionary Indices

Separator: 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.

Tag: `kIRGHanyuDaZidian`

Status: Provisional

Category: Dictionary Indices

Separator: 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.

Tag: `kIRGKangXi`

Status: Provisional

Category: Dictionary Indices

Separator:	space
Syntax:	`[01][0-9]{3}\.[0-7][0-9][01]`
Description:	The position of this character in the *KangXi* dictionary used in the four-dictionary sorting algorithm. The position is in the form "page.position" with the final digit in the position being "0" for 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 *KangXi* dictionary as used by the IRG in the four-dictionary sorting algorithm. The edition of the *KangXi* dictionary used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989.

Tag: `kIRG_GSource`

Status: Normative

Category: IRG Sources

Separator: space

Syntax: `(4K|BK|CH|CY|FZ(_BK)?|HC|HZ|KX|[0135789ES]-[0-9A-F]{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.

- 4K Siku Quanshu
- BK Chinese Encyclopedia
- CH The Ci Hai (PRC edition)
- CY The Ci Yuan
- FZ and FZ_BK Founder Press System
- 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 GB8685-88
- GE GB16500-95

- HC The Hanyu Da Cidian
- HZ The Hanyu Da Zidian
- KX The KangXi dictionary

Tag: kIRG_HSource
Status: Normative
Category: IRG Sources
Separator: N/A
Syntax: [0-9A-F]{4}
Description: The IRG "H" source mapping for this character in hex. The IRG "H" source consists of data from the Hong Kong Supplementary Characer Set.

Tag: kIRG_JSource
Status: Normative
Category: IRG Sources
Separator: space
Syntax: ([0134A]|3A)-[0-9A-F]{4}
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
- J4 JIS X 0213:2000
- JA Unified Japanese IT Vendors Contemporary Ideographs, 1993
- J3A JIS X 0213:2004 level-3

Tag: kIRG_KPSource
Status: Normative
Category: IRG Sources
Separator: 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

Tag: `kIRG_KSource`

Status: Normative

Category: IRG Sources

Separator: N/A

Syntax: `[01234]-[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 C 5601-1987
- K1 KS C 5657-1991
- K2 PKS C 5700-1 1994
- K3 PKS C 5700-2 1994
- K4 PKS 5700-3:1998

Note that the K4 source is expressed in hexadecimal, but unlike the other sources, it is not organized in row/column.

Tag: `kIRG_TSource`

Status: Normative

Category: IRG Sources

Separator: N/A

Syntax: `[1-7F]-[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 CNS 11643-1992, plane 1
- T2 CNS 11643-1992, plane 2
- T3 CNS 11643-1992, plane 3 (with some additional characters)
- T4 CNS 11643-1992, plane 4
- T5 CNS 11643-1992, plane 5
- T6 CNS 11643-1992, plane 6
- T7 CNS 11643-1992, plane 7
- TF CNS 11643-1992, plane 15

Tag: `kIRG_USource`

Status:	Normative
Category:	IRG Sources
Separator:	space
Syntax:	`U\+2?[0-9A-F]{4}`
Description:	The IRG "U" source mapping for this character. Currently, the IRG U source is limited to a small number of characters in the CJK Compatibility Ideographs block, where the value is the Unicode code point.

Tag:	`kIRG_VSource`
Status:	Normative
Category:	IRG Sources
Separator:	space
Syntax:	`[0123]-[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 VHN 01:1998 • V2 VHN 02:1998 • V3 TCVN 6056:1995

Tag:	`kJIS0213`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[12],[0-9]{2},[0-9]{1,2}`
Description:	The JIS X 0213-2000 mapping for this character in min,ku,ten form.

Tag:	`kJapaneseKun`
Status:	Provisional
Category:	Readings
Separator:	space
Syntax:	`[A-Z]+`
Description:	The Japanese pronunciation(s) of this character.

Tag:	`kJapaneseOn`
Status:	Provisional
Category:	Readings
Separator:	space
Syntax:	`[A-Z]+`
Description:	The Sino-Japanese pronunciation(s) of this character.

Tag:	`kJis0`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9]{4}`
Description:	The JIS X 0208-1990 mapping for this character in ku/ten form.

Tag:	`kJis1`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9]{4}`
Description:	The JIS X 0212-1990 mapping for this character in ku/ten form.

Tag:	`kKPS0`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The KPS 9566-97 mapping for this character in hexadecimal form.

Tag:	`kKPS1`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9A-F]{4}`
Description:	The KPS 10721-2000 mapping for this character in hexadecimal form.

Tag:	`kKSC0`
Status:	Provisional
Category:	Other Mappings
Separator:	space
Syntax:	`[0-9]{4}`
Description:	The KS X 1001:1992 (KS C 5601-1989) mapping for this character in ku/ten form.

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

Tag:	`kKangXi`
Status:	Provisional
Category:	Dictionary Indices
Separator:	space
Syntax:	`[0-9]{4}\.[0-9]{2}[01]`
Description:	The position of this character in the *KangXi* dictionary used in the four-dictionary sorting algorithm. The position is in the form "page.position" with the final digit in the position being "0" for 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 of the *KangXi* dictionary used is the 7th edition published by Zhonghua Bookstore in Beijing, 1989.

Tag:	`kKarlgren`
Status:	Provisional
Category:	Dictionary Indices
Separator:	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".

Tag:	`kKorean`
Status:	Provisional
Category:	Readings
Separator:	space
Syntax:	`[A-Z]+`
Description:	The Korean pronunciation(s) of this character, using the Yale romanization system. (See http://www.coffeesigns.com/Resources/romanization/korean.asp for a comparison of the various Korean romanization systems.)

Tag:	`kLau`
Status:	Provisional
Category:	Dictionary Indices
Separator:	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.

Tag:	`kMainlandTelegraph`
Status:	Provisional
Category:	Other Mappings
Separator:	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.

Tag: `kMandarin`
Status: Provisional
Category: Readings
Separator: space
Syntax: `[A-ZÜ]+[1-5]`
Description: The Mandarin pronunciation(s) for this character in pinyin; Mandarin pronunciations are sorted in order of frequency, not alphabetically.

Tag: `kMatthews`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[0-9]{1,4}(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.

Tag: `kMeyerWempe`
Status: Provisional
Category: Dictionary Indices
Separator: 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).

Tag: `kMorohashi`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[0-9]{5}'?`

Description: The index of this character in the *Dae Kanwa Ziten*, aka Morohashi dictionary (Japanese) used in the four-dictionary sorting algorithm.

The edition used is the revised edition, published in Tokyo by Taishuukan Shoten, 1986.

Tag: `kNelson`
Status: Provisional
Category: Dictionary Indices
Separator: 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.

Tag: `kOtherNumeric`
Status: Informative
Category: Numeric Values
Separator: 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.

Tag: `kPhonetic`
Status: Provisional
Category: Dictionary-like Data
Separator: 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.

Tag: `kPrimaryNumeric`
Status: Informative
Category: Numeric Values

Separator: 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.

Tag: `kPseudoGB1`
Status: Provisional
Category: Other Mappings
Separator: space
Syntax: `[0-9]{4}`
Description: A "GB 12345-90" code point assigned 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.

Tag: `kRSAdobe_Japan1_6`
Status: Provisional
Category: Radical-Stroke Counts
Separator: space
Syntax: `[CV]\+[0-9]{1,5}\+[1-9][0-9]{0,2}\.[1-9][0-9]?\.[0-9]{1,2}`
Introduced: 4.1
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.

Tag:	`kRSJapanese`
Status:	Provisional
Category:	Radical-Stroke Counts
Separator:	space
Syntax:	`[0-9]{1,3}\.[0-9]{1,2}`
Description:	A Japanese radical/stroke count for this character in the form "radical.additional strokes". A ' after the radical indicates the simplified version of the given radical.

Tag:	`kRSKanWa`
Status:	Provisional
Category:	Radical-Stroke Counts
Separator:	space
Syntax:	`[0-9]{1,3}\.[0-9]{1,2}`
Description:	A Morohashi radical/stroke count for this character in the form "radical.additional strokes". A ' after the radical indicates the simplified version of the given radical.

Tag:	`kRSKangXi`
Status:	Provisional
Category:	Radical-Stroke Counts
Separator:	space
Syntax:	`[0-9]{1,3}\.[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". A ' after the radical indicates the simplified version of the given radical.

Tag:	`kRSKorean`
Status:	Provisional
Category:	Radical-Stroke Counts
Separator:	space
Syntax:	`[0-9]{1,3}\.[0-9]{1,2}`
Description:	A Korean radical/stroke count for this character in the form "radical.additional strokes". A ' after the radical indicates the simplified version of the given radical

Tag: `kRSUnicode`
Status: Informative
Category: Radical-Stroke Counts
Separator: space
Syntax: `[0-9]{1,3}\'?\.[0-9]{1,2}`
Description: A standard radical/stroke count for this character in the form "radical.additional strokes". A ' after the radical indicates the simplified version of the given radical

This field is 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 intended to reflect the same radical as the kRSKangXi field and the stroke count of the glyph used to print the character within the Unicode Standard.

Tag: `kSBGY`
Status: Provisional
Category: Dictionary Indices
Separator: space
Syntax: `[0-9]{3}\.[0-9]{2}`
Description: The position of this character in the *Song Ben Guang Yun* (SBGY) Medieval Chinese character dictionary (bibliographic and general information below).

The 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.

The current 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.

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 Unihan.txt 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.

Tag: `kSemanticVariant`

Status: Provisional

Category: Variants

Separator: space

Syntax: `U+2?[0-9A-F]{4}(<k[A-Za-z:]+(,k[A-Za-z]+)*)?`

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), or Z (for *zhèng*, U+6B63 正).

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, the Hanyu Da Zidian indicates that U+5231 刱 and U+5275 創 are semantic variants and that U+5275 創 is the preferred form.

Tag: `kSimplifiedVariant`

Status: Provisional

Category: Variants

Separator: space

Syntax: `U\+2?[0-9A-F]{4}`

Description: The Unicode value for the simplified Chinese variant for this character (if any).

Note that a character can be *both* a traditional Chinese character in its own right *and* the simplified variant for other characters (e.g., U+53F0).

In such case, the character is listed as its own simplified variant and one of its own traditional variants. This distinguishes this from the case where the character is not the simplified form for any character (e.g., U+4E95).

Much of the of the data on simplified and traditional variants was supplied by Wenlin http:/www.wenlin.com.

Tag: `kSpecializedSemanticVariant`
Status: Provisional
Category: Variants
Separator: space
Syntax: `U+2?[0-9A-F]{4}(<k[A-Za-z]+(,k[A-Za-z]+)*)?`
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).

Tag: `kTaiwanTelegraph`
Status: Provisional
Category: Other Mappings
Separator: 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.

Tag: `kTang`
Status: Provisional
Category: Readings
Separator: 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.

Tag: `kTotalStrokes`

Status: Provisional

Category: Dictionary-like Data

Separator: space

Syntax: `[1-9][0-9]{0,2}`

Description: The total number of strokes in the character (including the radical). This value is for the character as drawn in the Unicode charts.

Tag: `kTraditionalVariant`

Status: Provisional

Category: Variants

Separator: space

Syntax: `U\+2?[0-9A-F]{4}`

Description: The Unicode value(s) for the traditional Chinese variant(s) for this character.

Note that a character can be *both* a traditional Chinese character in its own right *and* the simplified variant for other characters (e.g., 台 U+53F0).

In such case, the character is listed as its own simplified variant and one of its own traditional variants. This distinguishes this from the case where the character is not the simplified form for any character (e.g., 井 U+4E95).

Much of the of the data on simplified and traditional variants was supplied by Wenlin http:/www.wenlin.com.

Tag: `kVietnamese`

Status: Provisional

Category: Readings

Separator: space

Syntax: `[A-Za-z\x{E0}-\x{1B0}\x{1EA1}-\x{1EF9}]+`

Description: The character's pronunciation(s) in Quốc ngữ.

Tag: `kXerox`

Status: Provisional

Category: Other Mappings

Separator: space

Syntax: `[0-9]{3}:[0-9]{3}`

Description: The Xerox code for this character.

Tag: `kXHC1983`

Status: Provisional

Category: Dictionary-like Data

Separator: space

Syntax: `[0-9,.*]+:[a-z\x{FC}\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 reading is preceded by the character's location(s) in the dictionary, separated from the reading by a colon (:). Multiple locations for a given reading are separated by commas (,). Each location reference is of the form [0-9]{4}\.[0-9]{3}*?. The number preceding the period is the page number, zero justified to four digits. The first two digits of the number following the period are the entry's position on the page, zero-justified. The third digit is 0 for a main entry and greater than 0 for a parenthesized variant. A trailing askterisk (*) 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].

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).

As of the present writing (Unicode 5.1), the XHC source data contains 204 unencoded characters, for the most part simplified variants. Each unencoded character in the source is replaced by one or more encoded variants (references in these 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.

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.

Tag:	`kZVariant`
Status:	Provisional
Category:	Variants
Separator:	space
Syntax:	`U+2?[0-9A-F]{4}(:k[A-Za-z]+)?`
Description:	The Unicode value(s) for known z-variants of this character.

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
5.1	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)*

The remaining fields were added prior to Unicode 2.1.

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 6

- **Proposed Update** for Unicode 5.2.0

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