Han Ideograph (CJK) for Internationalized Domain Names

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

During the development of Internationalized Domain Name (IDN), it is discovered that there is a substantial lack of information and misunderstanding on Han ideographs and its folding mechanism. This document attempts to address some of the issues on doing han folding with respect to IDN. Hopefully, this will dispel some of the common misunderstanding of this problem and to discuss some of the issues with han ideograph and its folding mechanism.

This document addresses very specific problem to IDN and thus is not meant as a reference for generic Han folding. Generic Han folding are much more complicated and certainly beyond this document. However, the use of this document may be applicable to other areas that are related with names, e.g. Common Name Resolution Protocol [CNRP].

1. Definition and convention

Characters mentioned in this document are identified by their position or code point in the Unicode character set [UCS]. The notation U+12AB, for example, indicates the character at the position 12AB (hexadecimal) in the [UCS]. It is strongly recommended that a [UCS] table is available for reference for the ideograph described.

Han ideographs are defined as the Chinese ideographs starting from U+3400 to U+9FFF or commonly known as CJK Unification Ideographs. This covers Chinese ‘hanzi’ (U+6F22 U+5B57/U+6C49 U+5B57), Japanese ‘kanji’ (U+6F22 U+5B57) and Korean ‘hanja’ (U+6F22 U+5B57/U+D55C U+C790). Additional Han ideographs will appear in other location (not necessary in plane 0) in the future.
Conversion between ideographs can be done using four different approaches: Code-base substitution, character-based substitution, lexicon-based substitution and context-based substitution. Han folding refers only to code-base substitution, similar to case mapping of alphabetic characters.

2. Introduction

Traditionally, domain names have been case insensitive (as defined in [RFC1035] Section 2.3.3). While this is not a problem when domain names are restricted to English alphanumeric letters and digits, it becomes a serious problem for IDN. An important criterion for having a robust IDN is to have good normalization and canonicalization forms. This is to ensure domain name duplications are kept to the minimal.

Fortunately, Unicode Consortium is developing technical reports on canonicalization [UTR21] and normalization [UTR15]. Hence, it becomes simple for IDN to ride upon the work of Unicode and use these references.

Unfortunately, both [UTR15] and [UTR21] are limited in scope and do not address many other scripts. In particular, Han ideographs are not discussed in detail in these documents and most experts are quick to point out that this problem is technically impossible.

2.1 Han ideographs

While there are many forms or writing style for Chinese characters, the most common used ‘zhengti’ (U+6B63 U+4F53/U+6B63 U+9AD4) represent Chinese ideographs by radicals (U+2E80-U+2FDF) that is composed of simple strokes.

When the Unicode Consortium started work on Universal Character Set, it was suggested that Hanzi, Kanji and Hanja ideographs should be unified into a single code space. This resulted in the CJK Unification, whereby 27,786 Han ideographs are allocated in U+3400-U+9FFF and U+F900-U+FAFF range. Another 41,000 Han ideographs will be added to Plane 2.

Ideographs are common in China, Korea and Japan but as ideographs spread and evolve, the form of the ideographs sometimes differs slightly from country to country. For example, the word ‘villa’ (U+838A) ‘zhuang’ in Chinese, in Japanese is ‘sou’ (U+8358). These are given different code points in Unicode.

3. Chinese (Hanzi)

Chinese ideographs or hanzi (U+6F22 U+5B57/U+6C49 U+5B57) originated from pictograph. They are ‘pictures’ which evolved into ideographs during several thousand years. For instance, the ideograph for "hill" (U+5C71) still bears some resembles to 3 peaks of a hill.

Not all ideographs are pictograph. There are other classifications such as compound ideographs, phonetic ideographs etc. For example, ‘endurance’ (U+5FC3) is a pierced ‘knife’ (U+5200) above the ‘heart’ (U+5FC3), or as a Chinese saying goes, ‘endurance is like having a pierced knife in your heart’.

Hence, almost all Han ideographs are associated with some meaning by itself which is very different from most other scripts. This causes some confusion that Han folding is a form of lexicon-substitution.
Chinese ideographs underwent a major change in the 1950s after the establishment of People’s Republic of China. A committee on Language Reform was established in China whose activities include simplification of Chinese ideographs. The Simplified Chinese (SC) are used in China and Singapore and Traditional Chinese (TC) in Taiwan, Hong Kong PRC, Macau PRC, and most other overseas Chinese.

The process is to take complex ideographs and simplify them. The main purpose is to make it easier to remember and write and thus to raise the literacy of the population.

For example, 'lightning' TC {U+96FB} becomes SC {U+6535} (They drop the 'rain' {U+96E8} part from the TC). In many cases, they bear no resemblance to any of the original traditional forms e.g. 'dragon' TC (U+9F8D) SC (U+9F99). Two different TC may also have the same SC since it means fewer ideographs to learn, e.g. SC (U+53D1) can be (U+667C) or (U+9AEE) depending on semantics. The official ‘Comprehensive List of Simplified Characters’ latest published in 1986 listed 2244 SC [ZONGBIAO].

Therefore, the process of SC-to-TC is very complicated. It is not possible to do it accurately without considering the semantics of the phrase.

On the other hand, TC-to-SC is much simpler although different TCs may map to one single SC. While Unicode does not handle TC & SC, in the informal [UNIHAN] document, it listed 2145 TC and its equivalent mapping of SC. However, because that document is informal and not part of the Unicode standard, it is incomplete and has mistakes in the code points. Hence, precise tables for TC-to-SC conversion have not been fully laid out.

In domain names, we are particularly interested in is to equivalences comparison of the names, and not converting SC-to-TC. Therefore, for this purpose, it is possible that equivalency matching be done in the TC-to-SC folding prior to comparison, similar to lower-case English strings before comparing them, e.g. 'taiwan' SC (U+53F0 U+6E7E) will match with TC (U+81FA U+5F4E) or TC (U+53F0 U+5F4E).

The side effect of this method is that comparing SC (U+53D1) to TC (U+667C) or TC (U+9AEE) will both be positive. This implies that SC ‘hair' SC .borderWidth [U+5934 U+53D1] will match TC (U+982D U+9AEE). It will also match TC (U+982D U+9AEE) that does not have any meaning in Chinese.

It should also be noted that SC are not used together with TC. Hence, 'hair' is either written as SC (U+5934 U+53D1) or TC (U+982D U+9AEE) but (almost) never (U+5934 U+9AEE) or (U+982D U+53D1). So the problem of SC and TC may not too serious for IDN.

Unfortunately, when it comes to names in Chinese, places where SC are used (i.e. Singapore and China), traditional and simplified ideographs are sometimes mixed within a single name for artistic reasons. Some of them even ‘create’ ideographs for their names.

[Need to add a section on Bopomofo U+3118 to U+312A in future draft]

4. Korean (Hanja and Hangeul)

Korean is one of the first cultures to imported Chinese ideographs into Korean language as a written form. These Korean ideographs are known as
‘hanja’ (U+6F22 U+5B57/U+D55C U+C790) and they are widely used until recently where ‘hangeul’ (U+D55C U+AE00) become more popular.

Hangeul (U+D55C U+AE00) is a systemic script designed by a 15th century ruler and linguistic expert, King Sejong (U+4E16 U+5B97). It is based on the pronunciation of the Korean language, hanmal. A Korean syllable is composed of ‘jamo’ (U+5B57 U+6BCD/U+C790 U+BAA8) elements that represent different sound. Hence, unlike Han ideographs, each hangeul syllable does not have any meaning.

Each hanja ideographs can be represented by hangeul syllable. For example, ‘samsung’ hanja (U+4E09 U+661F) hangeul (U+C0BC U+C131). Note that (U+4E09) is pronounced as ‘sa-ah-am’ or in jamo (U+3145) (U+314F) (U+3141), which gives hangeul (U+C0BC). While Jamo decompositions are described in [UTR15] in Form D decomposition, this document also suggested another hanguel canonical decomposition in Appendix A to accommodates both modern and old hangeul.

[Need to fill up Appendix A when information is more complete]

Most hanja characters have only one pronunciation. However, some hanja pronunciation differs as according to orthography (same for Chinese & Japanese) or the position in a word, which make this more complex. And of course, conversation of Hangeul back to hanja is impossible by code substitution without consideration for semantics.

Korean also invented their own ideographs that are called ‘gugja’ (U+56FD U+5B57/U+AD6D U+C790).

5. Japanese (Kanji, Hiragana, Katakana)

Japanese adopted Chinese ideograph from the Korean and the Chinese since the 5th century. Chinese ideographs in Japanese are known as ‘kanji’ (U+6F22 U+5B57). They also developed their own syllabary hiragana (U+5E73 U+4EEE U+540D) (U+3040-U+309F) and katakana (U+7247 U+4EEE U+540D) (U+30A0-U+30FF), both are derivative of kanji that has same pronunciation. Hiragana is a simplified cursive form, for example, ‘a’ (U+3042) was derived from ‘an’ (U+5B89). Katakana is a simplified part form, for example, ‘a’ (U+30A2) was derived from ‘a’ (U+963F). However, kanji all remain very integrated within the Japanese language.

Japanese also invented ideographs known as ‘kokuji’ (U+56FD U+5B57). For example, ‘iwashi’ (U+9C2F) is a Japanese kokuji ideograph. Kokuji are invented according to Han ligature rules. For example, ‘touge” “mountain pass” (U+5CE0) is a conjunction of meaning with ‘yama” “mountain” (U+5C71) + ‘ue” “up” (U+4E0A) + ‘shita” “down” (U+4E0B).

Japanese is also a vocal language, i.e. the script itself is based on pronunciation. Each hiragana corresponding to one pronunciation and 48 hiragana forms the basic of the Japanese language, including the less commonly used ‘we’ (U+3091). Furthermore, hiragana has more 35 forms to represent voiced sound, P-sound, double consonant. For example, ‘ga’ (U+304C) is a voiced sound of ‘ka’ (U+304B). Katakana is a mirror of hiragana with few more forms and they are used to integrate foreign words or phrases into Japanese, or to emphasize words or phrases even in Japanese, or to represent onomatopoeia. For example, ‘hamburger’ pronounced as ‘han-baa-gaa’ in Japanese is written as (U+30CF U+30F3 U+30DD U+30FC U+30AC U+30FC) instead of (U+306F U+3093 U+3070 U+3041 U+304C U+3041) because it is a foreign word.

If Japanese uses hiragana and katakana only, then it is fairly obvious that written Japanese is going to be very long. Hence, kanji are used
when referring to nouns or verbs. Each kanji corresponds to one or more hiragana characters. For example, ‘japan’ pronounced as ‘nippon’ (U+306B U+3063 U+307D U+3093) are written as (U+65E5 U+672C) instead.

Hiragana, like Korean jamo, has no meaning itself. And also, Kanji can take on different pronunciation (which means different hiragana) depending where and how it is use in the sentence. For example, ‘sky’ (U+7A7A) can be pronounced as (U+305D U+3089) or (U+30BD U+30E9). Hence, a code substitution between hiragana and kanji is impractical.

On the other hand, there are Kanji that has the same meaning with the same pronunciation and equivalent. For example, ‘river’ "kawa" can be either {U+5DDD} or {U+6CB3}. The only differential between the two ideographs is that it signifies the ‘size of the river’ (the latter is bigger river).

Japanese also reduce complex Chinese ideographs to a simplified form. For example, ‘both’ {U+5169} was simplified {U+4E21}. Note that Chinese simplified it to {U+4E24} instead. However, traditional Japanese kanji are seldom used nowadays beyond documenting old historical text that they are treated different from the more commonly used simplified form, or used to express proper noun such as person’s name or trademarks. Hence, Han folding here is not recommended.

4. Vietnamese

While Vietnamese also adopted Chinese ideographs (‘chu han’) and created their own ideographs (‘chu nom’), they were now replaced by romanized ‘quoc ngu’ today. Hence, this document does not attempt to address any issues with ‘chu han’ or ‘chu nom’.

5. zVariant

Unicode has a three dimension conceptual model to Ideograph Unification. The three dimensions are semantic (X axis - meaning, function), abstract shape (Y-axis - general form) and actual shape (Z-axis \(\text{Å}Â\) instantiated, type-faced).

When two ideographs have similar etymology but are given two different code points in Unicode, they are known as zVariant ideograph i.e. they belong to the same ‘Z’ axis. For example, ‘villa’ {U+838A} and {U+8358}.

6. Ideographic Description

In Unicode v3.0, an ideographic description (U+2FF0-U+2FFB) was introduced allowing Han ideograph to be constructed using radical (U+2EB0-U+2FD5) and Han ideograph (U+3400-U+9FFF).

The intention of this description method is to allow ideograph that is not defined by Unicode to be described. Hence, it is not necessary that these ideograph can be display properly. In addition, this method are not deterministic and allowing same ideograph to be represented in different sequence.

For example, ‘zong’ {U+9B03} (for discussion sake, we are going to use an ideograph which is already in Unicode) can be decomposed to U+2FF1 U+9ADF U+5B97 using descriptive code points and Unified Ideograph. U+9ADF can also be decomposed as U+2FF0 U+2ED2 U+2F3A and U+5B97 as
U+2FF5 U+2F28 U+2F70. In addition, U+9ADF is equivalent to U+2FBD. Hence, if we were to use only descriptive code points and radicals only, we can get U+2FF1 U+2FBD U+2FF5 U+2F28 U+2F70 or U+2FF1 U+2FF0 U+2ED2 U+2F3A U+2FF5 U+2F28 U+2F70.

In addition, certain radical has been simplified and thus, in some context, equivalent. For example, the radical for 'bird' can be either U+2EE6 or U+2FC3.

Hence, until there is a deterministic well-defined rule for ideographic description, ideographs formed by this method are not recommended for domain names use.

It should be noted that the Unicode Consortium never intended the ideographic description to be used in protocols like IDN where exact comparison must be done. But it is certainly desirable to this feature as it is commons for Chinese to invent ideographs for names by adding or removing radical from standard ideographs.

7. Mechanism

The implicit proposal in this document is that CJKV ideographs may or may not be "folded" for the purposes of comparison of domain names.

But if folding is required, there are four different ways that this folding could be done.

a) Folding by DNS clients, or by user agents
b) Folding by DNS servers
c) Folding by Domain Name registration services for the purposes of preventing confusing allocations CJKV Domain Names which, if transcoded, be the same

Before we can give much more reaction, we need to know which use is planned.

The third use is important. It should be put in place. This problem can be reduced alternately by representing non-ASCII characters that are domain names or other URL characters using hex-escaped character references in HTML pages.

To characterize Han characters as ideographs or pictograms is inadequate, because most of the Han ideograph have both a phonetic and a semantic element. Indeed, this is enough to characterize Chinese writing as phonetic, though it is other things as well. Thus, it’s difficult to comment on whether folding is useful for Chinese or not.

The first use has the problem that lightweight devices do not have enough room to fit a Unicode X-axis mapping table.

The second use has the problem that introducing mapping will limit the performance of DNS servers. Alphabetic case mapping can be performed using a single logical AND instruction; CJKV character folding requires a lookup table.

In alphabetic scripts, there is also requirement to fold Latin, Greek, Hebrew, Cyrillic, Hebrew and Arabic together. There may be a stronger requirement for CJKV characters.

Note also that because modern OS are Unicode based and have network-downloadable IMEs, "interoperability" is becoming less equivalent to
"use BIG5 characters only" or "use GB2312 character only" or "use Shift-JIS characters only".

If conservative safety is really required, then
1) find the x-axis characters which are available in all major CJK character sets used on the internet;
2) only allow variants of those in domain names;
3) when one variant is used, no other can be allocated. So comparisons are made on x-axis characters, but the license of that domain name can pick which y or z variants they wish to use.

Acknowledgement

The editor gratefully acknowledge the contributions of:

Paul Hoffman <phoffman@imc.org>
Jiang Mingliang <jiang@i-DNS.net>
Dongman Lee <dlee@icu.ac.kr>
Karlsson Kent <keka@im.se>

Author(s)

James SENG  Å­¸Å·Á¶Å·Á«Å
i-DNS.net International Pte Ltd.
8 Temasek Boulevard
Suntec Tower 3 #24-02
Singapore 038988
Email: James@Seng.cc
Tel: +65 2468208

Yoshiro YONEYA
NTT Software Corporation
Shinagawa IntercityBldg., B-13F
2-15-2 Kohnan, Minato-ku Tokyo 108-6113 Japan
Email: yone@po.ntts.co.jp
Tel: +81-3-5782-7291

Kenny HUANG  Å©Á¶Å·Ã¥ÄÅ
Geotempo International Ltd; TWNIC
3F, No 16 Kang Hwa Street, Nei Hu
Taipei 114, Taiwan
Email: huangk@alum.sinica.edu
Tel: +886-2-2658-6510

KIM Kyongsok/GIM Gyeongseog

References

[IDN] "IETF Internationalized Domain Names Working Group",
  idn@ops.ietf.org, James Seng, Marc Blanchet
[CNRP] "Common Name Resolution Protocol",
  cnrp-ietf@lists.netsol.com, Leslie Daigle
Sanseido's Unicode Kanji Information Dictionary
ISBN 4-385-13690-4

Unicode chart http://charts.unicode.org/


Unicode Han Database, Unicode Consortium
ftp://ftp.unicode.org/Public/UNIDATA/Unihan.txt


