Proposal for a Chinese Script Root Zone LGR

*LGR Version 1.1*

*Date: 2018-02-04*

*Document version:10.2*

*Authors: CGP (Chinese Generation Panel)*

#  General Information/ Overview/ Abstract

The purpose of this document aims to give an overarching view of the label generation rules for the Chinese Script (Hani) including rationale behind the design decisions taken. This includes a discussion of the relevant features of the script, the communities and languages using it, as well as the process and methodology used and information of the contributors. The formal specification of the LGR can be found in the accompanying XML document:

Proposed-LGR-Hani-20171205.xml

Labels for testing can be found in the accompanying text document:

Labels-Hani-20171205.txt

[Other files accompanying this document must be listed here, including the filenames for the appendices, if they are to remain in separate files.]

#  Script for which the LGR is proposed

ISO 15924 Code: Hani

ISO 15924 Key N°: 500

ISO 15924 English Name: Han

Latin transliteration of native script name: Hanzi, Kanji, Hanja

Native name of the script: 汉字, 漢字, 한자

Maximal Starting Repertoire (MSR) version: MSR-2

#  Background on Script and Principal Languages Using It

##  Background

The Chinese Script (Hani in ISO 15924) is composed of characters, a kind of logograms used in the writing systems of Chinese and some other Asian languages. They are called Hanzi in Chinese, Kanji in Japanese and Hanja in Korean.



Figure 1: Evolution of Chinese Characters

Hanzi originated from [inscriptions](http://cn.bing.com/dict/search?q=inscriptions&FORM=BDVSP6&mkt=zh-cn) [on](http://cn.bing.com/dict/search?q=on&FORM=BDVSP6&mkt=zh-cn) [bones](http://cn.bing.com/dict/search?q=bones&FORM=BDVSP6&mkt=zh-cn) [or](http://cn.bing.com/dict/search?q=or&FORM=BDVSP6&mkt=zh-cn) [tortoise](http://cn.bing.com/dict/search?q=tortoise&FORM=BDVSP6&mkt=zh-cn) [shells](http://cn.bing.com/dict/search?q=shells&FORM=BDVSP6&mkt=zh-cn) in [the](http://cn.bing.com/dict/search?q=the&FORM=BDVSP6&mkt=zh-cn) [Shang](http://cn.bing.com/dict/search?q=Shang&FORM=BDVSP6&mkt=zh-cn) [Dynasty](http://cn.bing.com/dict/search?q=Dynasty&FORM=BDVSP6&mkt=zh-cn) ([c](http://cn.bing.com/dict/search?q=c&FORM=BDVSP6&mkt=zh-cn). 16[th](http://cn.bing.com/dict/search?q=th&FORM=BDVSP6&mkt=zh-cn)-11[th](http://cn.bing.com/dict/search?q=th&FORM=BDVSP6&mkt=zh-cn) [century](http://cn.bing.com/dict/search?q=century&FORM=BDVSP6&mkt=zh-cn) [B](http://cn.bing.com/dict/search?q=B&FORM=BDVSP6&mkt=zh-cn).[C](http://cn.bing.com/dict/search?q=C&FORM=BDVSP6&mkt=zh-cn).), [known](http://cn.bing.com/dict/search?q=known&FORM=BDVSP6&mkt=zh-cn) as the ["](http://cn.bing.com/dict/search?q=%22&FORM=BDVSP6&mkt=zh-cn)[Oracle](http://cn.bing.com/dict/search?q=Oracle&FORM=BDVSP6&mkt=zh-cn)["](http://cn.bing.com/dict/search?q=%22&FORM=BDVSP6&mkt=zh-cn) and was unified in the Qin dynasty (221-207 B.C.). In modern times, the most important changes in Chinese Hanzi occurred in the middle of the 20th century when more than two thousand simplified characters were introduced as the official forms in Mainland China. As a result, the Chinese language has two writing systems: Simplified Chinese (Hans) and Traditional Chinese (Hant). Both systems are expressed using different subsets under the common Unicode definition of the Hanzi script. The two writing systems use ISO 15924 scripts codes Hans and Hant respectively. Their repertoires are overlapping, sharing a common subset of "unchanged" Hanzi that accounts for around 60% of characters in contemporary use. The common "unchanged" Hanzi subset enables a user of simplified Chinese to understand texts written in traditional Chinese with little difficulty and vice versa. Hanzi characters in Hans and Hant share the same meaning and the same pronunciation and are typically variants.

Chinese characters have been adopted as Japanese Kanji for recording the Japanese language since the 5th century AD. Chinese words borrowed into Japanese could be written using Chinese characters, while Japanese words could be written using the characters for Chinese words of similar meaning. Finally, in Japanese, all three scripts (Kanji, and the Hiragana and Katakana syllabaries) are used as main scripts.

Chinese script spread to Korea together with Buddhism from the 2nd century BC to the 5th century AD. In times past, until the 15th century, in Korea, Literary Chinese was the dominant form of written communication, prior to the creation of Hangul, the Korean alphabet. In the modern Hangul-based Korean writing system, Chinese characters (Hanja) are no longer officially used to represent native morphemes, but still sometimes used in daily life.



Figure 2: Chinese script spread to Japan and Korea

Chinese script was also formerly used in Mongolia and Vietnam, but not anymore. Accordingly, Chinese Generation Panel does not take into account the usage of Chinese script in Mongolia and Vietnam.

##  Countries with Significant Usage for Chinese Script

Chinese script is used to write a diverse set of languages across East Asia and South East Asia. Countries and regions using Chinese script are depicted as follows:



|  |  |
| --- | --- |
|  | Traditional Chinese script used exclusively or almost exclusively (Taiwan, Macau and Hong Kong) |
|  | Simplified Chinese script used exclusively or almost exclusively(Mainland China and Singapore) |
|  | Simplified Chinese script used formally but Traditional script still used widely(Malaysia) |
|  | Chinese script used with other systems of writing in the same languageKanji (Japan) |
|  | Chinese script daily used but no longer officially usedHanja (Republic of Korea)  |

Figure 3: Countries using Chinese script

##  Principal Languages using the Script

As shown in the following non-exhaustive table, Chinese, Japanese and Korean are the three main languages using the Chinese script today but it does not imply that unlisted languages are less significant. For example, there are cases where a language may have a large population, but only a small part of it writes in Chinese script. Such languages are excluded from this list. In this list, all ISO 639-3 languages classed as "living" are included. They are taken from <http://www-01.sil.org/ISO639-3/codes.asp>, and codes may refer to a macro or an individual language.

|  |  |  |  |
| --- | --- | --- | --- |
| **Language** | **Language Code in ISO 639** | **Native Script Name** | **Locations** |
| Chinese | cdo (Min Dong Chinese)cjy (Jinyu Chinese)cmn (Mandarin Chinese)cpx (Pu-Xian Chinese)czh (Huizhou Chinese)czo (Min Zhong Chinese)gan (Gan Chinese)hak (Hakka Chinese)hsn (Xiang Chinese)mnp (Min Bei Chinese)nan (Min Nan Chinese)wuu (Wu Chinese)yue (Yue Chinese)zho (Chinese) | 汉字 Hanzi | China MainlandTaiwanHong KongMacaoSingaporeMalaysia |
| Japanese | jpn | 漢字 Kanji | Japan |
| Korean | kor | 한자 Hanja | Korea |

* Hanzi normally consists of two subsets, Simplified Chinese characters (Hans) and Traditional Chinese characters (Hant).
* Kanji is used in Japanese in addition to two other scripts (Hiragana and Katakana), together known as Jpan (ISO 15924 code).
* Hanja is used in Korean in addition to the Hangul script, together known as Kore (ISO 15924 code).

The relationship between Hanzi, Kanji and Hanja is as shown below, Hanzi (Hans & Hans), Kanji and Hanjia are all therefore covered by CGP.



Figure 4: Hanzi, Kanji & Hanja

#  Overall Development Process and Methodology

##  Previous work

In April 2004, the Joint Engineering Team (JET), a group composed of members of CNNIC, TWNIC, KRNIC, JPNIC as well as other individual experts, produced RFC 3743 "Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese and Korean". This guideline is intended for zone administrators, including but not limited to registry operators and registrars; and it includes information for all domain names holders on the administration of domain names that contain characters drawn from the Chinese, Japanese, and Korean scripts. It includes concepts for variant handling, such as bundling, atomic IDL Packages, and reserved variants. It also defines a standard table as well as an algorithm to generate the preferred variant and reserved variants. The key mechanisms of this specification utilize a three-column table, called a Language Variant Table, for each language permitted to be registered in the zone.

Collectively, the CDNC (Chinese Domain Name Consortium) has devised solutions to handle Chinese domain name variants. This includes bundling of Simplified Chinese (SC) and Traditional Chinese (TC) ("TC-SC Equivalence") domain names — as defined by the JET in RFC 3743 (April 2004) for the Chinese language as defined in RFC 4713 (October 2006) — and delegating the applied label, one preferred SC label and one preferred TC label to the same applicant. CDNC’s registration policy on handling TC-SC Equivalence is widely accepted. The [**CDNC IDN Table]**, developed by many Chinese linguistic and domain name experts over the last 10 years, is currently adopted by the Chinese, Taiwanese, Hong Kong, Macau and Singaporean governments, as well as by many new gTLD applicants. Over a decade of operating experience indicates that CDNC’s TC-SC Equivalence solution is a market-proven successful practice for handling Chinese variants in domain names.

Meanwhile, dotAsia, the registry of .ASIA and the member of CDNC, has extended the CDNC IDN Table by importing characters from HKSCS (Hong Kong Supplementary Character Set) and the Singapore set, developed its [**dotAsia IDN table]** under the framework of CDNC rules, to cover needs from the Hong Kong and Singaporean local communities.

A detailed analysis of the Chinese script had already been performed by the community in an earlier phase of the LGR program, which resulted in a **Chinese Case Study Team Report** (<https://archive.icann.org/en/topics/new-gtlds/chinese-vip-issues-report-03oct11-en.pdf>).

All these previous efforts made by the Chinese script community have been used as a basis for the current work, especially the Chinese Study Report and RFC 4713, in addition to other literature and the expertise available in the current task force.

##  Team Diversity

The current work is undertaken by experts from CDNC, who largely represent the Chinese language ccTLDs, as well as experts from a variety of backgrounds.

Geographically, the CGP has members from Chinese language regions across East Asia, including China mainland, Taiwan, Hong Kong, Macau, Singapore, Malaysia, as well as members from Europe and North America, totally 23 members belonging to 10 countries/regions listed in Appendix A.

The CGP consists of members with a diverse set of disciplines and very different perspectives. The members represent national and regional policy makers, the technical community directly working with the DNS, the security and law enforcement community, academia (technical and linguistic), and experience with local language computing using Unicode and specifically IDNs.

Besides, the CGP is pleased to have **Edmon CHUNG**, CEO of dotAsia and Co-Chair of the Universal Acceptance Steering Group, as the IDN advisor.

##  Work Process

The work has been carried out starting in September 2014, when the group was formed to put forward a "proposal for a generation panel for the Chinese script label generation rule set for the root zone". Since then, the Chinese Generation Panel (CGP) has held fortnightly conference calls, as well as 4 face-to-face meetings along with the CDNC meeting, in July 2015, March 2016, March 2017 and July 2017. In addition, the panel has been actively engaged on email, through the public mailing list of the task force.



The panel also maintains frequent communication with the JGP and KGP, to coordinate the Chinese code points and variant characters among three parties. The three Parties held 5 joint face-to-face meetings, in March 2015, May 2015, March 2016, September 2016 and November 2016, and had successive CJK joint sessions in ICANN meetings since ICANN 51 Los Angeles.

The work process includes the following steps:

* **Define and finalize the code point repertoire**

Within the scope set by the MSR, the CDNC and most CGP members urged to add CDNC characters into the CGP repertoire as much as possible, to reach consistency between the CDNC SLD operation and future TLD operation. Both CDNC IDN table and dotAsia IDN table have been taken into account.

In addition, the Table of General Standard Chinese Characters (TGSCC) published by China State Council in 2013, as well as the JGP repertoire and KGP repertoire were also referred to.

* **Define and finalize the code point variant sets**

CDNC has provided a market-proven variant set in the CDNC IDN table. Following CDNC rules, dotAsia extended CDNC repertoire and variant set to meet the requirements from the Hong Kong and Singaporean local communities. The CGP adopted CDNC variant rules directly and then made any necessary updates related to dotAsia variant rules.

The CGP recognizes that different panels (C, J and K) have different views on the variants corresponding to the same Chinese character; some CGP variant mappings conflict with KGP and JGP’s perception and practice. The CGP would work closely with JGP & KGP to make any necessary compromises to reach a consensus for all three parties and meet the [Procedure] requirement that "The variant mappings must agree for the same code point for all LGRs".

* **Define and finalize the whole label evaluation ruleset**

The CGP WLE follows the spirit of the CDNC ruleset, "TC-SC equivalence", which assigns all variant labels to the same applicant, while allocating the original applied label as well as only preferred SC label(s) and preferred TC label(s), generally no more than three labels, and blocks all other labels.

The CGP also acknowledges that some multiple preferred variant mappings work for SLD but may overproduce allocatable labels in the root zone. The CGP worked together with J, K and the IP to design an ideal solution to set applicants’ preferred labels allocatable as well as to limit the amount of allocatable labels to a reasonable number (for example, three).

* **Create XML LGR for Chinese script LGR proposal**

Considering the fact that the coordination on repertoire, variant mappings and WLE between CJK and IP is still in progress, the CGP work was carried out in a fast iteration model as indicated in the following figure:



Figure 5: Iteration model of CGP work process

#  Repertoire

##  Basic character set

**In 2004**, according to RFC 3743 and RFC 4713, the Chinese Domain Name Consortium (CDNC) drafted CDNC Chinese IDN Table. The CDNC Table has been used for second level domain (SLD) name registration under .CN, .TW, .HK and many CDN TLDs. In March, 2005, CNNIC and TWNIC submitted .CN Chinese Character Table[[1]](#footnote-1) and .TW Chinese Character Table[[2]](#footnote-2) separately, which included repertoire and variant mappings information.

**In 2012**, CDNC reviewed, proofread and published its combined IDN Table for the implementation of Chinese IDN registrations at gTLDs, including 37 ASCII code points and 19,520 Chinese characters (<http://www.cdnc.org/gb/research/file/CDNC_unicode.txt>).

##  Repertoire formation process

###  19561 Basic Repertoire

In October 2015, CDNC published the latest version of its IDN Table; 41 new Chinese characters were added into the character set as requested by HKIRC on behalf of the Hong Kong local community in 2013 and 2015, increasing the number of Chinese characters to 19,561.

Among the 41 Hong Kong characters, two characters (3A5C㩜 and 58B5墵) are out of scope of MSR2, which means only 19559 are included in MSR2. CGP has formally requested to add these two characters into MSR-3.

 

All these **19,561** code points form the basic set of the CGP repertoire. (CGP R0=CDNC IDN Table)

###  124 dotAsia characters

The dotAsia extended CDNC IDN Table 2012 to 19683 Chinese characters by adding 163 additional code points; of which 156 are part of HKSCS included in the IICORE collection; 4 are Non-IICORE and GS (Singapore Characters); 3 are Non-IICORE and part of various other Chinese sources that are necessary to insure full transitivity in variant processing. Together they make up the subset of the dotAsia repertoire contained in MSR-2, or 19,683 code points.
(<https://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt>)

39 of 163 are already included in CDNC IDN Table 2015, the remaining 124 extended the CGP repertoire up to **19,685** code points (CGP R1, including the dotAsia IDN Table).

####  55 characters located in the Basic Multilingual Plane as well as in IICORE:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unicode** | **Hanzi** | **CDNC** | **dotAsia** | **TGSCC** | **JGP** | **KGP** | **IICore** | **G** | **T** | **J** | **H** | **K** | **M** | **KP** | **S** |
| 65FF | 旿 |  | .ASIA | N |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 4C81 | 䲁 |  | .ASIA | V |  |  | IICORE |  | T4B |  |  |  |  |  | C |
| 5605 | 嘅 |  | .ASIA | V |  |  | IICORE |  |  |  | H1F |  | M1F |  | B |
| 6335 | 挵 |  | .ASIA | V |  |  | IICORE |  | T3B |  |  |  |  |  | C |
| 656D | 敭 |  | .ASIA | V |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 681E | 栞 |  | .ASIA | V | JGP | KGP | IICORE |  |  | J1A |  | K1C |  |  | B |
| 7460 | 瑠 |  | .ASIA | V | JGP | KGP | IICORE |  | T3D | J1A |  | K0A |  | P0A | A |
| 74C8 | 瓈 |  | .ASIA | V |  |  | IICORE |  | T3G |  |  |  |  |  | C |
| 9771 | 靱 |  | .ASIA | V | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 34E4 | 㓤 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3577 | 㕷 |  | .ASIA |  |  |  | IICORE |  | T3B |  |  |  |  |  | C |
| 35A1 | 㖡 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 35AD | 㖭 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 35BF | 㖿 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 35CE | 㗎 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1F |  | B |
| 35F3 | 㗳 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 35FE | 㗾 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 39F8 | 㧸 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 39FE | 㧾 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3A18 | 㨘 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3A52 | 㩒 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1F |  | B |
| 3A67 | 㩧 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3B39 | 㬹 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3DE7 | 㷧 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3DEB | 㷫 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3E74 | 㹴 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 3ED0 | 㻐 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  |  |  | P0A | C |
| 4065 | 䁥 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 406A | 䁪 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 40BB | 䂻 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 40DF | 䃟 |  | .ASIA |  |  |  | IICORE |  |  |  | H1E |  |  |  | C |
| 44EA | 䓪 |  | .ASIA |  |  |  | IICORE |  |  |  | H1D |  |  |  | C |
| 4606 | 䘆 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 47F4 | 䟴 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 4AB8 | 䪸 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  | K3D |  |  | C |
| 4C7D | 䱽 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 4C85 | 䲅 |  | .ASIA |  |  |  | IICORE |  | T4B |  |  |  |  |  | C |
| 4EEE | 仮 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 51B4 | 冴 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 5689 | 嚉 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 57DE | 埞 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 60E3 | 惣 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 62A6 | 抦 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 637F | 捿 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 6667 | 晧 |  | .ASIA |  | JGP | KGP | IICORE |  |  | J1A |  | K0A |  | P0A | A |
| 701E | 瀞 |  | .ASIA |  | JGP | KGP | IICORE |  | T3G | J1A |  | K0A |  | P0A | A |
| 7534 | 甴 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1C |  | B |
| 757A | 畺 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 7AC3 | 竃 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 8420 | 萠 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 9244 | 鉄 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 932C | 錬 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 98C7 | 飇 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 98E1 | 飡 |  | .ASIA |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 99C5 | 駅 |  | .ASIA |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |

####  7 characters located in the Basic Multilingual Plane, but not in IICORE

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unicode** | **Hanzi** | **CDNC** | **dotAsia** | **TGSCC** | **JGP** | **KGP** | **IICore** | **G** | **T** | **J** | **H** | **K** | **M** | **KP** | **S** |
| 39DB | 㧛 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 3BA3 | 㮣 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 43D3 | 䏓 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 4443 | 䑃 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 4882 | 䢂 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 4C9D | 䲝 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |
| 4C9E | 䲞 |  | .ASIA |  |  |  |  |  |  |  |  |  |  |  |  |

####  62 code points from Supplementary Ideographic Plane (Plane 2)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unicode** | **Hanzi** | **CDNC** | **dotAsia** | **TGSCC** | **JGP** | **KGP** | **IICore** | **G** | **T** | **J** | **H** | **K** | **M** | **KP** | **S** |
| 2070E | 𠜎 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1E |  | B |
| 20731 | 𠜱 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20779 | 𠝹 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1C |  | B |
| 20C53 | 𠱓 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20C78 | 𠱸 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20C96 | 𠲖 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20CCF | 𠳏 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20CD5 | 𠳕 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20D15 | 𠴕 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20D7C | 𠵼 |  | .ASIA |  |  |  | IICORE |  |  |  |  |  | M1E |  | C |
| 20D7F | 𠵿 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20E0E | 𠸎 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20E0F | 𠸏 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20E77 | 𠹷 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20E9D | 𠺝 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20EA2 | 𠺢 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20ED7 | 𠻗 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20EF9 | 𠻹 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1F |  | B |
| 20EFA | 𠻺 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20F2D | 𠼭 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20F2E | 𠼮 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20F4C | 𠽌 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20FB4 | 𠾴 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20FBC | 𠾼 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 20FEA | 𠿪 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 2105C | 𡁜 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 2106F | 𡁯 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 21075 | 𡁵 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 21076 | 𡁶 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 2107B | 𡁻 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 210C1 | 𡃁 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 210C9 | 𡃉 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 211D9 | 𡇙 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 220C7 | 𢃇 |  | .ASIA |  |  |  | IICORE |  |  |  | H1E |  |  |  | C |
| 227B5 | 𢞵 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22AD5 | 𢫕 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22B43 | 𢭃 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1F |  | B |
| 22BCA | 𢯊 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22C51 | 𢱑 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22C55 | 𢱕 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22CC2 | 𢳂 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22D08 | 𢴈 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22D4C | 𢵌 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22D67 | 𢵧 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 22EB3 | 𢺳 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 23CB7 | 𣲷 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 244D3 | 𤓓 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 24DB8 | 𤶸 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 24DEA | 𤷪 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 2512B | 𥄫 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 26258 | 𦉘 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 267CC | 𦟌 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1C |  | B |
| 269F2 | 𦧲 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 269FA | 𦧺 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 27A3E | 𧨾 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 2815D | 𨅝 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 28207 | 𨈇 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 282E2 | 𨋢 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1C |  | B |
| 28CCA | 𨳊 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 28CCD | 𨳍 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 28CD2 | 𨳒 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  |  |  | C |
| 29D98 | 𩶘 |  | .ASIA |  |  |  | IICORE |  |  |  | H1F |  | M1C |  | B |

###  60 characters from TGSCC, JGP and KGP

####  18 TGSCC characters

After CDNC revealed Chinese Character Table version 2012, in 2013, China's State Council published **Table of General Standard Chinese Characters 通用规范汉字表 (TGSCC)** as
<http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm> (8105 characters).
To stay consistent with the official document, the CGP studied the TGSCC and found 18 characters outside CGP R1 that fall in the range of the MSR. The CGP imported these 18 characters into the repertoire, making the total number 19,703. (CGP R2)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unicode** | **Hanzi** | **TGSCC Index**  | **TGSCCType** | **JGP** | **KGP** | **IICore** | **G** | **T** | **J** | **H** | **K** | **M** | **KP** | **S** |
| 48BC | 䢼 | 6657, Page 34  | N |  |  | IICORE | G3D |  |  |  |  |  |  | C |
| 732F | 猯 | 7409, Page 38 | N | JGP |  |  |  |  |  |  |  |  |  |  |
| 9EB9 | 麹 | 7748, Page 39 | N | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 5227 | 刧 | Page 48 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 524F | 剏 | Page 47 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 6060 | 恠 | Page 54 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 74A2 | 璢 | Page 61 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 750E | 甎 | Page 55 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 754A | 畊 | Page 58 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 7ADA | 竚 | Page 70 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 8262 | 艢 | Page 84 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 88B5 | 袵 | Page 75 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 894D | 襍 | Page 47 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 8B0C | 謌 | Page 67 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 8F19 | 輙 | Page 78 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 945A | 鑚 | Page 59 | V | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 984B | 顋 | Page 67 | V | JGP |  |  |  |  |  |  |  |  |  |  |
| 9DC0 | 鷀 | Page 84 | V |  |  |  |  |  |  |  |  |  |  |  |

Note:
“N” means the character is the standard Normalized character listed in TGSCC
”V” means the character is not standard Normalized one but listed as suggest variant in TGSCC
All above information could be found in Appendix Z : TGSCC Full Table

####  42 CJK Coordination characters

Since the meaning is inherent to the symbol, the same logographic system can theoretically be used to represent different languages like Chinese, Japanese and Korean.

In the early 2000s, CDNC experts developed the CDNC IDN Table (CGP R0) in which some IICORE Hani characters used in the Japanese or Korean writing systems are excluded. Under the framework of CJK coordination, the Hani characters included in the JGP repertoire or KGP repertoire need to be reviewed as regards the variant relationships to the CGP repertoire, to reach a tripartite consensus on the characters and variant mappings.

To ensure that the CGP repertoire will not cause confusion or conflict for global Chinese character users and applicants at the root level, the CGP reviewed 99 IICORE characters in the MSR but not covered in CGP R2; of these, CGP found that 42 characters included in the JGP repertoire (version 201703, Appendix C) and KGP repertoire (version 201703, Appendix D) have variant relationships with code points in CGP R1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Unicode** | **Hanzi** | **CDNC** | **dotAsia** | **TGSCC** | **JGP** | **KGP** | **IICore** | **G** | **T** | **J** | **H** | **K** | **M** | **KP** | **S** |
| 967A | 険 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 7E4A | 繊 |  |  |  | JGP |  | IICORE |  | T3D | J1A |  |  |  |  | A |
| 9421 | 鐡 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 9D8F | 鶏 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 4FAD | 侭 |  |  |  | JGP |  | IICORE |  |  |  |  |  |  |  | A |
| 6442 | 摂 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 685F | 桟 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 7E4B | 繋 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 81D3 | 臓 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 8217 | 舗 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 9039 | 逹 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 9EBA | 麺 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 3960 | 㥠 |  |  |  |  | KGP | IICORE |  |  |  |  | K3D |  |  | C |
| 51E6 | 処 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 61F4 | 懴 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 685C | 桜 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 6D9C | 涜 |  |  |  | JGP |  | IICORE |  |  |  |  |  |  |  | A |
| 6E8C | 溌 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 731F | 猟 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 784F | 硏 |  |  |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 7C14 | 簔 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 7D9A | 続 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 8133 | 脳 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 86CD | 蛍 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 8E99 | 躙 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 91A4 | 醤 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 91C8 | 釈 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 96B2 | 隲 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 982C | 頬 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 98EE | 飮 |  |  |  | JGP | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 9A12 | 騒 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 9A13 | 験 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 9C2E | 鰮 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 9D0E | 鴎 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 9D2C | 鴬 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 56A2 | 嚢 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 663B | 昻 |  |  |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 7A36 | 稶 |  |  |  |  | KGP | IICORE |  |  |  |  | K0A |  | P0A | A |
| 7B86 | 箆 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |
| 839F | 莟 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 83B5 | 莵 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | C |
| 9A28 | 騨 |  |  |  | JGP |  | IICORE |  |  | J1A |  |  |  |  | A |

The CGP imported these 42 characters into the LGR, raising the total number of code points listed to 19,745 (CGP R3). However, all these 42 characters have reflexive mappings of type "out-of-repertoire-var" in Chapter 6. They are therefore not considered part of the “repertoire” and any labels containing one of these code points will be invalid. They are listed merely to provide the requisite entries for symmetric and transitive out-of-repertoire variants.

Finally, CGP generated the repertoire through the steps from 5.2.1 to 5.2.4, using the formation process illustrated in the following figure:



Figure 6: CGP R3 repertoire components

###  19,703 Characters’ source information

The CDNC IDN Table (version 2015) has 19,561 Hani/Hanzi characters, all included in CGP Repertoire.

The dotAsia IDN Table (version 1.1) has 19,683 Hani characters, all included in CGP Repertoire.

The TGSCC provided 10,986 Hani/Hanzi characters to CGP Repertoire, including 7846 general standard characters and related 3140 variant characters in comparison table/对照表.



Figure 7：Source of CGP repertoire

The JGP repertoire (version 201703, Appendix B) has 6358 Hani/Kanji characters. 6216 are overlapped characters in the CGP R2 repertoire. CGP will treat the other 142 as Japanese UNIQUE Hani/Kanji characters and will not seek to import any of them into the CGP repertoire. The JGP agreed to follow the CGP variant mappings of the overlapping 6216 characters as illustrated in Section 6.2.5.

The KGP repertoire (version 201703, Appendix C) has 4758 Hani/Hanja characters. 4744 are overlapping characters in the CGP R2 repertoire. The KGP agreed to follow the CGP variant mappings of the overlapping 4749 characters as illustrated in Section 6.2.5.



Figure 8：Source of CGP repertoire

CGP provides the detailed source information of every character in the repertoire in Appendix A.

##  Attempt to limit the size of the repertoire

In Section 5.2, the CGP generated a repertoire containing 19,703 code points / characters. It is remarkable that the CGP repertoire has such a large size compared with most other GPs. CGP would attribute it to the nature of the Chinese writing system, similar to other logographic writing systems with large repertoires.

Unlike a segmental writing system (e.g. alphabetic, Abjad, Abugida) which has limited [graphemes](https://en.wikipedia.org/wiki/Grapheme) to represent the [phonemes](https://en.wikipedia.org/wiki/Phoneme) (basic units of sound) of a language, or Syllabaries (e.g. Kana), which have limited graphemes to represent [syllables](https://en.wikipedia.org/wiki/Syllable) or [moras](https://en.wikipedia.org/wiki/Mora_%28linguistics%29), a logographic writing system has glyphs/logograms to represent [words](https://en.wikipedia.org/wiki/Word_%28linguistics%29) or [morphemes](https://en.wikipedia.org/wiki/Morpheme) rather than phonetic elements. In Chinese, a logogram is a single written character that represents a complete grammatical word (or, more precisely, a [morpheme](https://en.wikipedia.org/wiki/Morpheme)). As each character represents a single word, many logograms are required to write all the words of the language.

There are two reasons to explain why there are so many characters in the Chinese writing system. First, each Chinese character is an independent unit representing a word. 3000 years ago, the Shang [Dynasty](http://cn.bing.com/dict/search?q=Dynasty&FORM=BDVSP6&mkt=zh-cn) (16[th](http://cn.bing.com/dict/search?q=th&FORM=BDVSP6&mkt=zh-cn)-11[th](http://cn.bing.com/dict/search?q=th&FORM=BDVSP6&mkt=zh-cn) [century](http://cn.bing.com/dict/search?q=century&FORM=BDVSP6&mkt=zh-cn) [B](http://cn.bing.com/dict/search?q=B&FORM=BDVSP6&mkt=zh-cn).[C](http://cn.bing.com/dict/search?q=C&FORM=BDVSP6&mkt=zh-cn).) oracle bones included 3500-4500 characters already. During the course of history, more characters were invented to represent new words created along with social development. Second, massive numbers of variants occurred with the spread of Chinese characters and the development of written communication in the continent of East Asia. **Chinese variants are characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms.** In the Chinese writing system, variants are deemed as exchangeable, the classic case is simplified characters and traditional characters. Generally, each Chinese character has at least one non-reflexive variant character (in CDNC IDN Table, 1 non-reflexive variant on average, at most 7 non-reflexive variants).

Statistically, "[Text](http://cn.bing.com/dict/search?q=Text&FORM=BDVSP6&mkt=zh-cn) [Notes](http://cn.bing.com/dict/search?q=Notes&FORM=BDVSP6&mkt=zh-cn) [and](http://cn.bing.com/dict/search?q=and&FORM=BDVSP6&mkt=zh-cn) [Word](http://cn.bing.com/dict/search?q=Word&FORM=BDVSP6&mkt=zh-cn) [Explanations](http://cn.bing.com/dict/search?q=Explanations&FORM=BDVSP6&mkt=zh-cn) 说文解字/說文解字" in Han Dynasty (202 B.C.-220 A.D.) includes 9,353 characters, "Lei Pian 类篇/類篇" in Song Dynasty (960-1279 A.D.) includes 31,319 characters. In 1710, [Emperor](http://cn.bing.com/dict/search?q=Emperor&FORM=BDVSP6&mkt=zh-cn) [Kangxi](http://cn.bing.com/dict/search?q=Kangxi&FORM=BDVSP6&mkt=zh-cn) released the "Kangxi [Dictionary](http://cn.bing.com/dict/search?q=Dictionary&FORM=BDVSP6&mkt=zh-cn) 康熙字典" including 47,035 characters. In 1959, the Japanese scholar Tetsuji Morohashi compiled "Dai Kan-Wa Jiten大漢和辞典" covering 49,964 characters. In 1994, the Chinese Zhonghua Book Company published "Zhonghua Zihai 中华字海" containing 87,019 characters. In 2004, the [Taiwan Ministry of Eduction released](http://dict.variants.moe.edu.tw/main.htm)"[Dictionary of Chinese Character Variants異體字字典"](http://dict.variants.moe.edu.tw/main.htm)containing 106,230 characters.

It is obvious that, among the tens of thousands of Chinese characters, not all are frequently used in modern society. The Chinese Ministry of Education requires that students be able to handle 3500 characters after [nine](http://cn.bing.com/dict/search?q=nine&FORM=BDVSP6&mkt=zh-cn) [year](http://cn.bing.com/dict/search?q=year&FORM=BDVSP6&mkt=zh-cn)s of [compulsory](http://cn.bing.com/dict/search?q=compulsory&FORM=BDVSP6&mkt=zh-cn) [education](http://cn.bing.com/dict/search?q=education&FORM=BDVSP6&mkt=zh-cn), the number is 3500-4500 in Taiwan and 3500 in Hong Kong. However, everyday Chinese script users are able to "write" and "read" many more characters than what they actually learned in school due to two reasons.

The first reason is that Chinese variant characters have the same pronunciation. Because of that, modern internet users who have received compulsory education prefer to use phonetic-based input methods (e.g. Pinyin拼音 in China mainland, Zhuyin注音 in Taiwan, Jyutping粤拼 in Hong Kong), which allow users to input [phonetic](http://cn.bing.com/dict/search?q=phonetic&FORM=BDVSP6&mkt=zh-cn) [symbol](http://cn.bing.com/dict/search?q=symbol&FORM=BDVSP6&mkt=zh-cn)s and select characters/labels from the alternative variant characters/labels with the same pronunciation in the selection box. Moreover, a few users prefer other input methods like shape-based input methods (e.g. Wubi五笔 in China, Simplified Tsang-jei 速成 in Hong Kong), handwriting recognition or speech recognition, however, most of them provide a phonetic-based selection box as a basic function to enable users to input variants with [no](http://cn.bing.com/dict/search?q=no&FORM=BDVSP6&mkt=zh-cn) [barriers](http://cn.bing.com/dict/search?q=barriers&FORM=BDVSP6&mkt=zh-cn).

The second reason is, a set of Chinese variant characters generally share the same radical or components, and thus have a certain degree of visual similarity, allowing educated readers to recognize the variant relationship easily. For example, the character for "fight" (a morpheme pronounced dòu ) has 6 variants with similar visual forms, 鬪(9B2A)闘(95D8)鬥(9B25)鬦(9B26)鬬(9B2C)鬭(9B2D). More importantly, hardly any variant character appears alone in any domain label: they occur together with other characters in a word or phrase, providing semantic context and helping the readers to recognize the meaning of domain labels more effectively and conveniently. (For example 头发/头髪tóufǎ ‘hair’ and 发展/發展fāzhǎn ‘development’)

The above two natural characteristics give Chinese variant characters great acceptability, usability and exchangeability in real life, especially in information systems. Hence, the [development](http://cn.bing.com/dict/search?q=developement&FORM=BDVSP6&mkt=zh-cn) and popularization of the [internet](http://cn.bing.com/dict/search?q=internet&FORM=BDVSP6&mkt=zh-cn) promote Chinese character usage in cyberspace. In terms of [Computer](http://cn.bing.com/dict/search?q=Computer&FORM=BDVSP6&mkt=zh-cn) [Coding](http://cn.bing.com/dict/search?q=Coding&FORM=BDVSP6&mkt=zh-cn) [Standards](http://cn.bing.com/dict/search?q=Standards&FORM=BDVSP6&mkt=zh-cn), the early Taiwan BIG5 standard includes 13,053 characters, the current Taiwan state standard CNS11643(4.0) includes 76,067 characters. China GB2312 standard included 6763 characters, while the latest standard GB18030 included 20,912 characters. The current Unicode standard (as of 10 October 2015), including CJK Unified Ideographs Extensions A-E, contains 80,388 characters. In terms of internet application and daily usage, in 2007, the paper "**A Survey on the Usage of Chinese Characters and Phrases in the Newspapers, Radio, TV, and Web**" in Applied Linguistics [1003-5397(2007)01-0029-09] shows 8128 independent characters are used in daily life. Another paper in 2010, "**Survey on Chinese Weblog Wording" in Journey of Xianning University** [1006-5342(2010)01-0076-03], shows 20923 characters are used.

The most symbolic event occurred in 2016: China’s [Ministry](http://cn.bing.com/dict/search?q=Ministry&FORM=BDVSP6&mkt=zh-cn) [of](http://cn.bing.com/dict/search?q=of&FORM=BDVSP6&mkt=zh-cn) [Civil](http://cn.bing.com/dict/search?q=Civil&FORM=BDVSP6&mkt=zh-cn) [Affairs](http://cn.bing.com/dict/search?q=Affairs&FORM=BDVSP6&mkt=zh-cn) issued **Notification 2016[33]**, requiring government departments to update the naming-related information system in public service and administration areas, to cover the characters in national standard GB13000 (20,902 chars) or GB18030 (70,244 chars). The two standards cover the CGP repertoire entirely.
**http://www.gov.cn/xinwen/2016-05/09/content\_5071481.htm**

Actually, most of the above concerns were taken into account when the CDNC generated its Chinese IDN Table in the early 2000s. To create an IDN Table with broad applicability and [backwards](http://cn.bing.com/dict/search?q=backwards&FORM=BDVSP6&mkt=zh-cn) [compatibility](http://cn.bing.com/dict/search?q=compatibility&FORM=BDVSP6&mkt=zh-cn), the CDNC referred to multiple source files about Chinese characters and variants, including:

1. Complete List of Simplified Characters 简化字总表 (2235 chars)
2. List of Commonly Used Characters in Modern Chinese现代汉语通用字表 (7000 chars)
3. China National Standard GB2312 (6763 chars)
4. Taiwan standard BIG5 (13,053 chars)
5. Chinese Variants Collation Table 第一批异体字整理表 (810 variant groups)
6. Chinese Big Dictionary 汉语大字典 (54,678 chars)
7. Chinese Relationship Table for Unihan Project
8. International Standard Chinese Big Dictionary 国际标准汉字大辞典
9. Unicode 3.2
10. Unihan Database and extension A (20,992 + 6,582 chars)

The CDNC took Reference 1 – Reference 4 as sources to set up a fundamental character set, then imported variant characters from Reference 5 – Reference 8 to develop variant mappings, generated the CDNC IDN table with 19520 Chinese characters. All fall in the range of Reference 9 (Unicode 3.2) and Reference 10 (Unihan Database and extension A).

In the early stage of developing the GP repertoire, CGP members attempted to replace the CDNC IDN table with a smaller character set, hoping the reduction would help decrease the [computational](http://cn.bing.com/dict/search?q=computational&FORM=BDVSP6&mkt=zh-cn) [complexity](http://cn.bing.com/dict/search?q=complexity&FORM=BDVSP6&mkt=zh-cn) of the LGR and speed up the coordination work with J & K. The CGP generated a reduced repertoire called MSS (Minimum Shared Set) of 12563 characters, most of them are historically registered in SLD under .CN/.TW/.HK/.网址 (7722 chars) or come from the Table of General Standard Chinese Characters (4612 chars) published by China PRC State Council in 2013.

The CGP generated the MSS and expected that this limited repertoire could significantly decrease the complexity and workload of coordination between CJK, however, this reduction attempt caused a heated discussion among the CGP members, especially for those registry representatives who had already adopted the CDNC IDN Table for second level registrations.

The core issue is that many members tend to believe that it is the variant mapping rules, not the repertoire size that directly affects the computational complexity of the LGR. The storage capacity and processing power of the modern computer is much more than what is needed to deal with a repertoire of about 20,000 characters. Since the 2000s, many IDN registries have adopted the CDNC IDN Table and developed IDN registration systems without decreasing the computational performance of the EPP service. Considering the SLD market acceptance of the existing CDNC IDN Table (adopted by over 5 ccTLDs and 20 new gTLDs) and the continuity of registries’/registrars’/registrants’ experience, many CGP members suggested that the characters of the CDNC IDN Table be included to the maximum extent possible.

Moreover, CJK coordination work shows that the JGP has no discrepancy with the CGP repertoire and variant mappings. The KGP has no discrepancy with the CGP repertoire either, but only concerns the mapping relationships of specific 258 variant groups.

For [all](http://cn.bing.com/dict/search?q=all&FORM=BDVSP6&mkt=zh-cn) the above [reasons](http://cn.bing.com/dict/search?q=reasons&FORM=BDVSP6&mkt=zh-cn)[,](http://cn.bing.com/dict/search?q=%2C&FORM=BDVSP6&mkt=zh-cn) the CGP decided to keep CGP R2 as the final repertoire with high capacity and [compatibility](http://cn.bing.com/dict/search?q=compatibility&FORM=BDVSP6&mkt=zh-cn). This is used to ensure consistency in user experience, registry practice and the local regulation.

#  Variants

##  Variant definition in CGP

In the Chinese writing system, there are two types of variants:

The first type is created by regional variations in the standard writing system. There are now two common writing systems: Simplified Chinese and Traditional Chinese. Both writing systems use different subsets of the same Unicode Han script, but their repertoires are not mutually exclusive.

The second type is the generic variant. Several Chinese characters are visually different in form but treated equally with universal interchangeability. This relationship of interchangeability is much stronger than the relationship between the Traditional and Simplified forms.

In the Chinese Case Study Team Report mentioned in 2.1, CHINESE (CHARACTER) VARIANTS are:

**"characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts."**

This understanding and variants mapping rule has been reflected in the CDNC IDN Table, and inherited by the current CGP LGR document.

In alignment with RFC 4713 and CDNC practice, generally, every code point in the CGP repertoire has its preferred/allocatable simplified variant(s), preferred/allocatable traditional variant(s), and reserved/blocked variant(s). In some cases, a code point has a reflexive preferred variant. In others, a code point has no reserved variant.



Figure 9: variant setting in CDNC IDN Table

Once transformed into XML-format defined in RFC 7940 all preferred variant char(s) are "allocatable", all reserved variant char(s) are "blocked", with sub-types as:

|  |  |  |
| --- | --- | --- |
| Sub-Type | Type | Comment |
| simp | Allocatable | preferred simplified variant char; |
| r-simp | Allocatable | reflexive preferred simplified variant char; |
| trad | Allocatable | preferred traditional variant char |
| r-trad | Allocatable | reflexive preferred traditional variant char |
| both | Allocatable | preferred simplified and traditional variant chars are the same |
| r-both | Allocatable | reflexive preferred simp and trad variant chars are the same |
| r-neither | Blocked | Non-allocatable reflexive/original char |
| blocked | Blocked | Non-allocatable variant char  |

Following the XML schema, the two variant mappings in Figure 8 will be transformed into the following XML source text:

<char cp="4F53" tag="sc:Hani" >

<var cp="4F53" type="r-simp" comment="identity" />

<var cp="8EB0" type="blocked" />

<var cp="8EC6" type="blocked" />

<var cp="9AB5" type="blocked" />

<var cp="9AD4" type="trad" />

</char>

<char cp="4E81" tag="sc:Hani" >

<var cp="4E7E" type="trad" />

<var cp="4E81" type="r-neither" comment="identity" />\

<var cp="5E72" type="simp" />

<var cp="5E79" type="blocked" />

<var cp="69A6" type="blocked" />

<var cp="6F27" type="blocked" />

 </char>

Considering that there are some characters borrowed/imported from JGP and KGP, which are not really used in the modern Chinese language region, the sub-type of "out-of-repertoire-var" is needed for the reflexive mappings of these characters.

|  |  |  |
| --- | --- | --- |
| Sub-Type | Type | Comment |
| out-of-repertoire-var | Invalid | Non-CGP chars imported from other GPs |

Original Simplified Traditional All Variants

硏(784F) 研(7814) 研(7814) 揅(63C5)研(7814)硏(784F)

[JGP] [CGP] [CGP] [CGP][CGP][JGP]

<char cp="784F" tag="sc:Hani" >

<var cp="63C5" type="blocked" />

<var cp="7814" type="both" />

<var cp="784F" type="out-of-repertoire-var" comment="identity" />

</char>

Note: To eliminate the overproduction of allocatable labels caused by multiple allocatable variant mappings, CGP designed some new sub-types under the type of "Allocatable", the related definition and WLE will be illustrated in Section 6.2.2 and Section 6.2.3.

##  Variant Mappings formation process

###  18974 basic variant mappings from CDNC-2005

CDNC IDN Table (repertoire and variant mappings), generated in early 2000s and extended in 2015, along with RFC3743 and RFC 4713, is the most wildly accepted rules for Chinese domain name registration at the second level, and has been applied to .CN, .TW, .MO, .HK, .SG for decades. The dotAsia IDN Table imports 99.5% of its variant mappings from CDNC-2005 IDN Table. Taking all these factors into account, CGP borrowed the variant mappings in the CDNC IDN Table directly and developed the basic variant mappings table corresponding to CGP Repertoire 0 (19,561 characters).

However, among the 19,561 characters and their variant mappings, there are a few variant mappings changed later due to the further coordination work from Section 6.2.2 to Section 6.2.4.

In this CGP proposal, 18911 variant mappings are kept as same as CDNC-2005 and dotAsia, 63 are the same as CDNC-2005 but different with dotAsia . These **18,974** (18,911 + 63) variant mappings forms up the basic CGP variant mappings table as Appendix D [Sheet 6.2.1-18974].

###  126 more variant mappings from dotAsia

In the early 2000s, when drafting the IDN table, CDNC experts focused on modern frequently-used characters in China mainland, Taiwan and Hong Kong, and excluded some specific locally-used characters and rarely-used IICORE characters from CDNC IDN Table 2015 (CGP R0). dotAsia extended the CDNC IDN table, by adding 124 new regional characters and modifying some old variant mappings according to local requirement. By comparing the two tables, we found a slight difference between dotAsia and CDNC-2015: 108 different variant mappings of the previously included characters, and 124 new variant mappings along with 124 newly-added characters.

CGP and Edmon CHUNG, the CEO of dotAsia, discussed the issue of inconsistency between CDNC-2015 variant mappings and dotAsia variant mappings, and agreed that the dotAsia table was created as an experiment for Hong Kong local characters, but the intent has always been to merge it and make it consistent with CGP rules once it is integrated for root zone and gTLD purpose. Thus, dotAsia agreed to synchronize and update the IDN table in IANA once the CGP rules are finalized in 5 steps:

**Step 1**, keep consistent with CDNC-2015 IDN table by giving up the 108 different variant mappings corresponding to the previously included characters.

**Step 2**, in September 2015, CGP & CDNC held joint meetings and invited linguistic experts from China mainland, Taiwan and Hong Kong, reviewed 55 new added dotAsia characters in section 5.2.2.1, together with 18 TGSCC characters in section 5.2.3 and 42 CJK coordination characters in section 5.2.3.2. reset the variant mappings of them as Appendix E.

**Step 3**, in May 2016, city of Haikou, CGP & CDNC joint meeting reviewed 7 unique dotAsia Hanzi characters in section 5.2.2.2. These Hanzi characters are not included in the CDNC-2005 IDN table, nor in TGSCC, nor in IICORE, but only exist in the dotAsia IDN table submitted to IANA. The variant mappings of the 7 characters were reset as Appendix E:

**Step 4**, for the 62 dotAsia code points from Unicode Plane 2 as in section 5.2.2.3, CGP directly accepted their variant mappings from dotAsia IDN Table into CGP rules.

**Step 5**, recheck and alter 32 variant mappings of previously encoded CDNC-2015 characters related to the above 124 newly-added characters.

However, some variant mappings were changed once again due to the further coordination work from Section 6.2.3 to Section 6.2.5.

In this CGP proposal, we have 27 variant mappings kept the same as dotAsia but different from CDNC-2015, 99 Non-CDNC-2005 dotAsia variant mappings unchanged. These **126** (27+99) variant mappings forms up the dotAsia variant mappings table as Appendix D [Sheet 6.2.2-126].

###  202 variant mappings from review team

####  17 variant mappings from TGSCC

Section 5.2.3 listed 18 characters in TGSCC which are not covered by CDNC IDN Table or dotAsia IDN Table. In September 2015, CGP & CDNC invited linguistic experts from Mainland China, Taiwan and Hong Kong, reviewed all these 18 TGSCC characters and gave the variant mappings of them as Appendix Review.

However, 1 variant mapping was changed due to the coordination between CGP and KGP in Section 6.2.4. In the final CGP proposal, we kept 17 variant mappings in Appendix D [Sheet 6.2.3.1-17].

####  41 variant mappings from JGP and KGP

Section 5.2.4 provided 42 characters imported from JGP and KGP as out-of-repertoire-var. In September 2015, CGP & CDNC invited linguistic experts from China mainland, Taiwan and Hong Kong, reviewed all these 42 JGP and KGP characters and gave them the reflexive mappings of type "out-of-repertoire-var".

However, one J&K variant mappings was changed due to the coordination between CGP and KGP in Section 6.2.5. In the final CGP proposal, we kept 41 variant mappings in Appendix D [Sheet 6.2.3.2-42].

####  144 variant mappings interacted

With the increase of variant mappings from Section 6.2.1 to Section 6.2.4, some newcomers influence the original CDNC-2015 variant mappings and dotAsia variant mappings, according to the CGP & CDNC joint review work in September 2015, there are 144 original related variant mappings got affected. The new variant mappings are listed in Appendix D [Sheet 6.2.3.3-144].

###  441 Variant mappings from CJK coordination

####  Principle and Framework

A coordination mechanism among three parties is needed to realize unified Chinese script generation rules in the DNS root zone. During the CDNC meeting in Shanghai (May, 2014), the IP proposed the basic principles of the coordination scheme:

* Each CJK panel creates an LGR and each LGR includes a repertoire and variants.
* If an LGR includes Han characters, the variant mappings must agree for all three panels.
* The variant types may be different (blocked or allocatable), so that the variant types do not have to agree as across LGRs.

Based on the principles above, the CGP, JGP and KGP started coordination work at the IETF Dallas meeting in 2015, trying to define a unified variant-mapping table for Chinese scripts, then define each party’s variant types/sub-types (e.g., allocatable or blocked) for characters contained in this table. According to the consensus at the IETF Dallas meeting, the JGP initiated work called "CJK Integration Procedure" as follows:

Step 1: Each CJK GP generates its own LGR (hereinafter, LGR-alpha)

Step 2: CJK GPs collectively generate a merged table of each LGR-alpha (hereinafter, LGR-M)

Step3: Each CJK GP extracts its original repertoire with integrated variants from LGR-M.

Step 4: Each CJK GP adds "Out of Repertoire" code points for symmetry.

Step 5: Each CJK GP merges WLE in LGR-alpha into one.

Step 6: Each CJK GP generates an integrated LGR (hereinafter, LGR-beta).



Figure 10: Framework of CJK LGR integration for Han characters, by JGP

####  Coordination between C and J

Some Kanji characters are in a simplified form (called the "new character form"), derived from the traditional imported form (called the "old character form"). In the Japanese language environment and writing system, it is appropriate to distinguish NEW and OLD forms as different and independent characters instead of pure variants. This understanding has been reflected in the IANA IDN table developed by the .JP registry, JPRS, in which no variants are identified for Kanji.

Some characters in a CGP variant group have the same pronunciations and meanings, but have different meanings in Japanese language environments. For example, (U+673机) means [desk, small table] and (U6A5F機) means [machine] in Japanese, but both mean [machine] in Chinese.

The JGP showed great openness and agreed to import all CGP variant mappings into the JGP ones. Thus, both parties eliminated the potential conflict caused by variant inconsistency. The CGP would like to express its appreciation for the JGP’s openness, tolerance and compromise.

####  Pre-integration between C and K

Hanja characters are no longer used in official documents (a law enacted on April 14th, 2011 orders all ROK official government documents to be written only in Hangul; Hanja or other scripts can only be written within parentheses if allowed by presidential decree), but are still sometimes used by a few Korean people in daily life. In August 2016, the KGP generated its first version of its LGR, and raised 258 Hanja/Hanzi variant groups whose variant mappings in the CGP LGR are NOT acceptable to the KGP (Appendix F).

Unlike the JGP, the KGP adopted another strategy and approach to resolve differences on variant mappings, asking the CGP to review and reset the variant mapping related to the 258 variant groups case by case.

As requested, the CGP performed a statistical analysis of all the disputed 258 variant groups, including the number of registered labels containing disputed characters under .CN/.TW/.HK/.网址, and the semantics of all these registered labels. The CGP separated 258 variant groups into 5 types and 2 categories, the variant groups the CGP would RESERVE, the ones the CGP would give up and split into INDEPENDENT characters.

|  |  |  |
| --- | --- | --- |
| Reserved 1 | Total registered number of variants is not zero, but the actual meaning of the variants are the same; suggest to follow CGP rules | 139 |
| Reserved 2 | One variant’s registered number of variants is close to zero, and the actual meanings of the variants are the same: suggest to follow CGP rules |
| Independent 1 | Total registered number of variants is zero, but not commonly used in domain registration; suggest dealing with these independently | 119 |
| Independent 2 | Total registered number of variants is not zero, but meanings of variants are different: suggest dealing with these independently |
| Independent 3 | One variant’s registered number of variants is close to zero, not commonly used in domain registration: suggest dealing with these independently |

The KGP invited linguistic experts from the local community and separated the 258 groups into 2 categories: 149 groups the KGP would KEEP and 109 groups the KGP would DROP.

|  |  |  |
| --- | --- | --- |
| Keep | Some CGP variant mappings don’t make any sense in the Korean Hanja environment, instead, these characters are used independently with their own semantic meanings. Hence, KGP intended to keep them as independent characters. | 149 |
| Drop | Some CGP variant mappings are weakly related in the Korean Hanja environment; the KGP intends to drop their independence and accept the CGP’s variant mappings. | 109 |



Figure 11: coordination before C&K Taiwan meeting

In September 2016, the KGP and CGP held their first-round coordination meeting and reduced the number of variant groups in disagreement to 60. In the IETF meeting, in November 2017, the KGP and CGP reduced the number to 2. Finally, in December 2017, the KGP and CGP solved all 258 variant groups and reached complete agreement on the variant mappings. For the 258 unacceptable variant groups, 112 will go with CGP rules, the remaining 146 will go with KGP rules.



Figure 12: C&K coordination process

Accordingly, CGP updated 445 variant mappings related to 146 groups as Appendix G, and KGP generated its Hanja variant mappings table as Appendix H.

However, through the processes from 6.2.1 to 6.2.3, there were 2 variant mappings that remained the same as the original CDNC rules, and 2 that remained as original dotAsia rules: so only 441 variant mappings could be considered as the result of KGP pre-integration as Appendix D [Sheet 6.2.3.1-17].

After all above steps from Section 6.2.1 to 6.2.4, CGP finalized the whole 19,745-strong CGP variant mapping table corresponding to CGP R2 as Appendix D [Sheet All Variant Mappings], including 18,974 basic variant mappings, 126 dotAsia variant mappings, 202 variant mappings from TGSCC, JGP and KGP, and 441 KGP pre-integration variant mappings.

###  19745 Variant mappings’ source information

To illustrate the relationship between the CGP variant mappings and the existing SLD practice and some other variant mappings work, CGP provides the reference/source information of every variant mapping in the XML document as well as in the EXCEL appendix I.

* The variant mappings consistent with the existing practice of CDNC.
* The variant mappings consistent with the existing practice of dotAsia.
* The variant mappings consistent with the CGP review settings.
* The variant mappings changed with the pre-integration with KLGR.

The variant mappings different from existing practice of CDNC or dotAsia as also indicated in reference/source information in XML document as well as in the EXCEL appendix J

* 587 variant mappings different from the existing practice of CDNC as Appendix J [Sheet CDNC]
* 646 variant mappings different from the existing practice of dotAsia as Appendix J [Sheet dotAsia]

##  Effort to reduce the number of multiple allocatable labels

In the Chinese writing system, there are a quite a few characters that have multiple simplified variant characters or multiple traditional variant characters. These multiple allocatable variant mappings might lead to overproduction of allocatable labels.

|  |  |  |  |
| --- | --- | --- | --- |
| Unicode | Original Char | Preferred Simplified Char | Preferred Traditional Char |
| 5925 | 夥 | 夥(5925) 伙(4F19) | 夥(5925) |
| 647A | 摺 | 摺(647A) 折(6298) | 摺(647A) |
| 9EBD | 麽 | 麽(9EBD) 么(4E48) | 麼(9EBC) |
| 5347 | 升 | 升(5347) | 升(5347) 昇(6607) 陞(965E) |
| 53F0 | 台 | 台(53F0) | 台(53F0 )檯(6AAF) 臺(81FA) |
| 590D | 复 | 复(590D) | 复(590D) 復(5FA9) 複(8907) |
| 5C40 | 局 | 局(5C40) | 局(5C40) 侷(4FB7) 跼(8DFC) |
| 5E76 | 并 | 并(5E76) | 并(5E76) 並(4E26) 併(4F75) |

Multiple preferred variant mapping examples

To avoid the overproduction issue, in SLD practice, CDNC members and dotAsia designed a ranking selection function or human interaction mechanism, to enable the applicants to SELECT at most one all-simplified and at most one all-traditional label from the multiple alternatives. Once [the](http://cn.bing.com/dict/search?q=the&FORM=BDVSP6&mkt=zh-cn) [selection](http://cn.bing.com/dict/search?q=selection&FORM=BDVSP6&mkt=zh-cn) is complete, all the other allocatable labels are reserved, the reserved allocatable labels could be reactivated later at the request of an applicant later, to make sure the applicant could get all his desired labels.

However, unlike the SLD practice, according to Root Zone LGR framework, the "human select" or "reserve and reactive" are not allowed in LGR, all generated labels are either ALLOCATABLE or BLOCKED, and the blocked labels will never be activated. So a new mechanism is needed to generate a limited number of allocatable labels, as well as to satisfy the applicant’s requirement in maximum degree.

CGP checked the variant mappings in Appendix D, found that there are only 136 multiple allocatable variant mappings out of all 19,745 cases, 3 with 2 PSVs (Preferred Simplified Variants), 127 with 2 PTVs (Preferred Traditional Variants), 5 with 3 PTVs and 1 with 4 PTVs. These 136 multiple variant mappings are listed in Appendix K. Having analyzed the 136 variant mappings one by one, CGP proposed an engineering method to optimize generation rules, with the aim of reducing the number of allocatable labels without eliminating multiple mappings.

###  Find out "redundant" variants

The 136 multiple allocatable variant mappings are divided into 6 categories:

|  |  |  |  |
| --- | --- | --- | --- |
| number | Original | Allocatable Simp | Allocatable Trad |
| 2 | A | AB | A |
| 1 | A | AB | C |
| 103 | A | A | AB |
| 23 | A | A | BC |
| 5 | A | A | ABC |
| 1 | A | B | CD |
| 1 | A | A | ABCD |

By case analysis and simulation computation, CGP found that these 136 variant mappings could be transferred to the following format without causing any serious problem for TLD applicant.

|  |  |  |  |
| --- | --- | --- | --- |
| number | Original | Allocatable Simp | Allocatable Trad |
| 2 | A | A (r-both-ms – muted, reflexive )B | A |
| 1 | A | A (muted, reflexive )B | C |
| 103 | A | A | A (muted, reflexive )B |
| 23 | A | A | B (trad-1, no mixed with trad-2)C (trad-2, no mixed with trad-1) |
| 5 | A | A | A (muted, reflexive )B (trad-1, no mixed with trad-2)C (trad-2, no mixed with trad-1) |
| 1 | A | B | C (trad-1, no mixed with trad-2)D (trad-2, no mixed with trad-1) |
| 1 | A | A | A (muted, reflexive )B (muted, not in Modern Chinese Common Used Table in China mainland, nor Common used Chinese standard table in Taiwan)C (trad-1) D (trad-2) |

The "muted" variant characters will be treated as "BLOCKED". The muted variants are redundant, either reflexive, or rarely used (not covered by Modern Chinese Common Used Table in China mainland or Common used Chinese standard table in Taiwan).

The "trad-1" and "trad-2" variant characters will be treated like two independent sub groups, which means, the mixture of "trad-1" ad "trad-2" will be blocked as redundant ones. If a specific mixed label happens to be the desired one, the applicant is asked to input this specific label as the original label, at the cost of losing some "less desired" traditional label.

Given any valid input label, the optimized rules will generate at most 4 allocatable labels -- the original label, an all-simplified label, an all-traditional label-1 and an all-traditional label-2. The engineering mechanism includes three steps.

The merit of the above mechanism is that it retains the same simplified and traditional mappings as the existing SLD as far as possible. It does not change the simplified type or traditional type of any variant character of these 136 variant mappings, Instead, it subdivides them into common simplified/traditional ones and extra simplified/traditional ones, and generates extra disposition rules. The disadvantage is that it doesn’t guarantee that the applicant could get any specific label from an infinite allocatable label list, but allows the applicant to replace the original input label with one specific desired variant label. CGP regards this as an acceptable trade-off to reduce the number of multiple allocatable labels.

###  Create new sub-types for multiple mapping variants

According to the designs in Section 6.3.1, CGP created 6 new sub-types, to identify the corresponding variant characters in multiple mappings.

|  |  |  |
| --- | --- | --- |
| Sub-Type | Type | Comment |
| "r-both-ms" | Allocatable | Reflexive preferred simp and trad variant chars are the sameReflexive preferred simp is one of multiple simp variant mappings Set reflexive preferred simp as "r-both-ms" |
| "r-both-mt" | Allocatable | Reflexive preferred simp and trad variant chars are the sameReflexive preferred trad is one of multiple trad variant mappings Set reflexive preferred trad as "r-both-mt" |
| "r-simp-m" | Allocatable | Reflexive preferred simp is one of multiple simp variant mappings Set reflexive preferred simp as "r-simp-m" |
| "trad-m" | Allocatable | Preferred trad is rare used, not in Modern Chinese Common Used Table in China mainland, nor Common used Chinese standard table in TaiwanSet the preferred trad as "trad-m" |
| "trad-1" | Allocatable | Among the multiple preferred traditional variantsSet the preferred trad with the smallest hex-code as "trad-1" |
| "trad-2" | Allocatable  | Among the multiple preferred traditional variantsSet the preferred trad with the largest hex-code as "trad-2" |

###  Create new evaluation rules

According to the designs in Section 6.3.1, CGP created some new evaluation rules to reduce the number of multiple allocatable labels, keep the number to a maximum of four (4).

<action disp="allocatable" only-variants="simp r-simp both r-both r-both-mt" comment="all simplified label" />

<action disp="allocatable" only-variants="trad r-trad both r-both r-both-ms trad-1" comment="all traditional label category one"/>

<action disp="allocatable" only-variants="trad r-trad both r-both r-both-ms trad-2" comment="all traditional label category two"/>

<action disp="blocked" any-variant="simp trad both trad-m trad-1 trad-2" comment="block any other mixed labels" />

<action disp="allocatable" all-variants="r-neither r-trad r-simp r-simp-m r-both r-both-mt r-both-ms" comment="original label" />

#  Whole Label Evaluation Rules (WLE)

##  Delegating all simplified, all tradition and original applied labels

There is a "TC-SC Equivalence" rule in RFC4713, which means **delegating the original applied-for label, all simplified labels and all traditional labels to the same applicant, blocking all the other labels**. To remain consistent with this rule, the CGP generates its own XML table of CGP repertoire and variant mappings according to the XML-format specifications in RFC 7940 and marks every variant mapping with one of the following types:

"r-simp", "r-trad", "r-both", "simp", "trad", "both", "r-neither", "blocked", "out-of-repertoire-var"

 A direct implementation of the rules in RFC 4713 would lead to the following definitions of “action” elements in the LGR:<rules>

<!--Action elements - order defines precedence-->

<action disp="invalid" any-variant="out-of-repertoire-var" comment="action for imported variant" />

<action disp="blocked" any-variant="blocked" comment="default action for blocked variant" />

<action disp="allocatable" only-variants="simp r-simp both r-both" comment="all simplified label" />

<action disp="allocatable" only-variants="trad r-trad both r-both" comment="all traditional label"/>

<action disp="blocked" any-variant="simp trad both r-simp r-trad r-both r-neither" comment="block any simp&trad mixed labels" />

<action disp="allocatable" only-variants="r-simp r-trad r-both r-neither" comment="original label"/>

<action disp="allocatable" comment="catch-all" />

</rules>

##  Blocking redundant all-simplified or all-traditional labels

To limit the number of allocatable labels to four (4), CGP created new sub-types for variants within multiple allocatable variant mappings, and marks corresponding variant mappings with one the following types (see Section 6.3.2):

"r-both-ms", "r-both-mt", "r-simp-m", "trad-m", "trad-1", "trad-2"

Consequently, the ”action” elements in the LGR are changed and extended as follows:

<rules>

<!--Action elements - order defines precedence-->

<action disp="invalid" any-variant="out-of-repertoire-var" comment="action for imported variant" />

<action disp="blocked" any-variant="blocked" comment="default action for blocked variant"/>

<action disp="allocatable" only-variants="simp r-simp both r-both r-both-mt" comment="all simplified label" />

<action disp="allocatable" only-variants="trad r-trad both r-both r-both-ms trad-1" comment="all traditional label category one"/>

<action disp="allocatable" only-variants="trad r-trad both r-both r-both-ms trad-2" comment="all traditional label category two"/>

<action disp="blocked" any-variant="simp trad both trad-m trad-1 trad-2" comment="block any other mixed labels" />

<action disp="allocatable" all-variants="r-neither r-trad r-simp r-simp-m r-both r-both-mt r-both-ms" comment="original label" />

<action disp="valid" comment="catch all (default action)"/>

</rules>

#  Contributors

List of CGP expert team

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Organization | Country/Region | Language Expertise |
| Chao QI | CNNIC | China | Chinese |
| Chris DILLON | University College London | UK | Chinese, Japanese, Korean |
| Connie Hon | IP Mirror | Singapore | Chinese |
| Di MA | ZDNS | China | Chinese |
| Guoying LI | Beijing Normal University | China | Chinese |
| Holmes LEONG | MONIC | Macao | Chinese |
| James SENG | 21ViaNet Group Limited | Malaysia | Chinese |
| Jean-Jacques Subrenat | ATLAC ICANN | France | French, English, Chinese, Japanese. |
| Jenifer CHUNG | Dot Asia | USA/Hongkong | Chinese |
| Jiagui XIE | CONAC | China | Chinese |
| Jonathan SHEA | HKIRC | Hong Kong | Chinese |
| Joseph YEE | Afilias | Canada | Simplified Chinese, Traditional Chinese, (Familiar with Japanese) |
| Kenny HUANG(Co-Chair) | TWNIC | Taiwan | Chinese |
| Linlin ZHOU | CNNIC | China | Chinese |
| Lu QIN | Hong Kong Polytechnic University | Hong Kong | Chinese |
| Nai-Wen HSU | TWNIC | Taiwan | Chinese |
| Ryan TAN | SGNIC | Singapore | Chinese |
| Shutian CUI | Ministry of Industry and Information Technology | China | Chinese |
| Wei WANG(Co-Chair) | CNNIC | China | Chinese |
| Xiaodong LEE | CNNIC | China | Chinese |
| Yuxiao LI | Beijing University of Posts and Telecommunications | China | Chinese |
| Zheng WANG |  | China | Chinese |
| Zhiwei YAN | CNNIC | China | Chinese |
| Zhoucai ZHANG | UniHan Digital Tech., Ltd. | China | Chinese mainly |

Advisor Edmon CHUNG

ICANN Staff Dr. Sarmad HUSSAIN

#  References

[Procedure] Internet Corporation for Assigned Names and Numbers, "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels." (Los Angeles, California: ICANN, March, 2013) <http://www.icann.org/en/resources/idn/variant-tlds/draft-lgr-procedure-20mar13-en.pdf>

[Requirements] Integration Panel "Requirements for LGR Proposals from Generation Panels" available online as <https://www.icann.org/en/system/files/files/Requirements-for-LGR-Proposals-20150424.pdf>

[UCD] The Unicode Consortium, Unicode Character Database, available online as <http://www.unicode.org/Public/UCD/latest/>

[CDNC-2015] CDNC IDN Table <http://www.cdnc.org/gb/research/file/unicode.txt>

[DotAsia] DotAsia ZH IDN Table <http://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt>

[TGSCC] Chinese Character Set China's State Council Table of General Standard Chinese Characters (TGSCC)
<http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm>

[IICORE] Chinese Character Set International Ideographs Core
<http://appsrv.cse.cuhk.edu.hk/~irg/irg/IICore/IRGN1067R2_IICore22_MappingTable.txt>

[RFC4290] Klensin, J., "Suggested Practices for Registration of Internationalized Domain Names (IDN)", RFC 4290, December 2005. https://datatracker.ietf.org/doc/rfc4290/

[RFC3743] Konishi, K., Huang, K., Qian, H., and Y. Ko, "Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean", RFC 3743, April 2004. https://datatracker.ietf.org/doc/rfc3743/

[RFC4713] Lee, X., Mao, W., Chen, E., Hsu, N., and J. Klensin, "Registratio and Administration Recommendations for Chinese Domain Names", RFC 4713, October 2006. https://datatracker.ietf.org/doc/rfc4713/

[IDNCJK] Seng, J., Yoneya, Y., Huang, K., and Kyongsok, K., "Han Ideograph (CJK) for Internationalised Domain Names", Internet Draft. Available at <http://tools.ietf.org/html/draft-ietf-idn-cjk-01>

[RFC7940] K. Davies, A. Freytag, Representing Label Generation Rulesets using XML, https://datatracker.ietf.org/doc/rfc7940/

[ICANN Documents] Guidelines for the Implementation of Internationalised Domain Names (2003). <http://www.icann.org/en/general/idn-guidelines-20jun03.htm>

[ICANN Documents] New gTLD draft Applicant Guidebook. 2011, <http://www.icann.org/en/topics/new-gtlds/rfp-clean-19sep11-en.pdf>

[ICANN Documents] Chinese Case Study Team Report, Report on Chinese Variants in Internationalized Top-Level Domains, 2011, <https://archive.icann.org/en/topics/new-gtlds/chinese-vip-issues-report-03oct11-en.pdf>

# Appendix A: CGP Repertoire

CGP Repertoire EXCEL document includes 19745 Unicode code points, as well as their utf-8 characters and source information.

# Appendix B: JGP Repertoire

JGP Repertoire EXCEL document includes 6533 Unicode code points, 6356 of which are Hani/Kanji characters.

# Appendix C: KGP Repertoire

KGP Repertoire EXCEL document includes 4758 Hani/Hanja Characters and their Unicode code points.

# Appendix D: CGP Variant Mappings Table

# Appendix E: CGP Variant Review on new added characters

CGP & CDNC joint review team worked on 172 IICORE characters and 7 dotAsia characters.

# Appendix F: KGP Unacceptable 258 Variant Groups

KGP proposed 258 variant groups whose variant mappings are unacceptable to Korean community in August 2016, CGP and KGP coordinated on these 258 variant groups (445 variant mappings).

# Appendix D: KGP CGP Pre-Integration on 445 Variant Mappings

The output of coordination work on 258 unacceptable variant groups proposed by KGP.

# Appendix I: CGP Variant Mappings Matching Existing Practice

CGP provides the reference/source information of every variant mapping that consistent with the existing practice of CDNC, dotAsia, as well as with the CGP review output and KLGR pre-integration.

# Appendix J: CGP Variant Mappings Differ from Existing Practice

The variant mappings different from existing practice of CDNC or dotAsia.

# Appendix K CGP Multiple Mappings

3 multiple allocatable simplified mappings and 103 multiple allocatable traditional mappings

# Appendix Z TGSCC Full Table 通用规范汉字表

8105 normalized Chinese characters in TGSCC and the related suggested variant characters.

1. <http://www.iana.org/domains/idn-tables/tables/cn_zh-cn_4.0.html> [↑](#footnote-ref-1)
2. <http://www.iana.org/domains/idn-tables/tables/tw_zh-tw_4.0.1.html> [↑](#footnote-ref-2)