# Proposal for a Chinese Script Root Zone LGR

*LGR Version 1.7.2*

*Date: 2020-02-28*

*Document version:17*

*Authors: CGP (Chinese Generation Panel)*

# 1 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:

- proposal-chinese-lgr-28feb20-en.xml

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

- chinese-test-labels-28feb20-en.txt

All the appendices to the document can be found in the accompanying EXCEL and XML documents:

- **proposal-chinese-lgr-28feb20-en**, referred to as [Proposal] below
- Appendix A CGP Repertoire [201902].xlsx,
- Appendix B JGP Repertoire [201703].xlsx,
- Appendix C KGP Hanja Repertoire [201703].xlsx,
- Appendix D CGP Variant Mappings [201909].xlsx,
- Appendix E CGP Internal Review [202002].xlsx
- Appendix F IP External Review [201909].xlsx
- Appendix G.1 CDNC IDN Table 2018 in RFC3743 format.txt
- Appendix G.2 CDNC IDN Table 2018 in XML format.xml
- Appendix H KGP Hanja Variant Mappings [201703].xlsx,
- Appendix I CGP Variant Mappings Matching Existing Practice [201909].xlsx
- Appendix J CGP Variant Mappings Differ from Existing Practice [201909].xlsx
- Appendix K CGP Multiple Mappings [201906].xlsx,
- Appendix L.1 CDNC IDN Table 2005-2012 in RFC3743 format.txt
- Appendix L.2 CDNC IDN Table 2005-2012 in XML format.xml
- Appendix M.1 dotAsia IDN Table in RFC3743 format.txt
- Appendix M.2 dotAsia IDN Table in XML format.xml
- Appendix N CGP Internal Review.xml
- Appendix O out-of-repertoire variants [201909].xlsx
- Appendix P KGP Variant Groups.xml
- Proposed-LGR-Hani-1.7.2-20200224.xml
- Labels-Hani-20180824.txt

# 2 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-4

# 3 Background on Script and Principal Languages Using It

## 3.1 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.

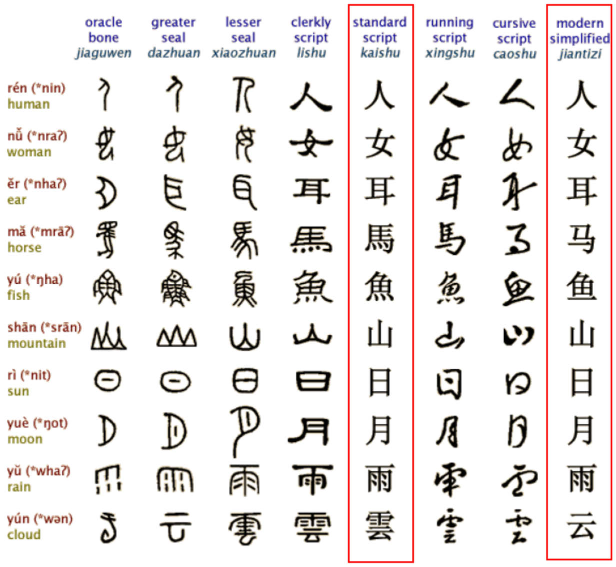


Figure 1: Evolution of Chinese Characters

Hanzi originated from inscriptions on bones or tortoise shells in the Shang Dynasty (c. 16th-11th century B.C.), known as the "Oracle" 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. Later, a pair of syllabaries known as hiragana and katakana, derived by simplifying Chinese characters selected to represent syllables of Japanese. Ultimately, modern Japanese writing uses a composite system, using kanji for word stems, hiragana for inflexional endings and grammatical words, and katakana to transcribe non-Chinese loanwords as well as serve as a method to emphasize native words.[1]

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.

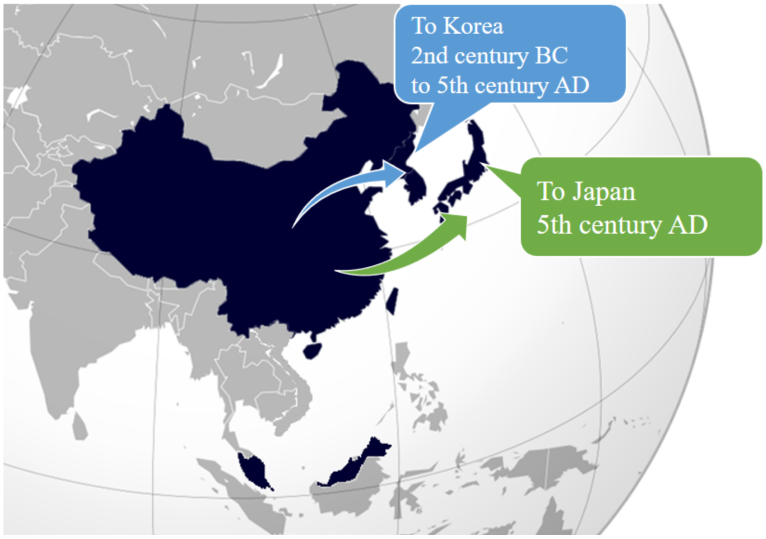
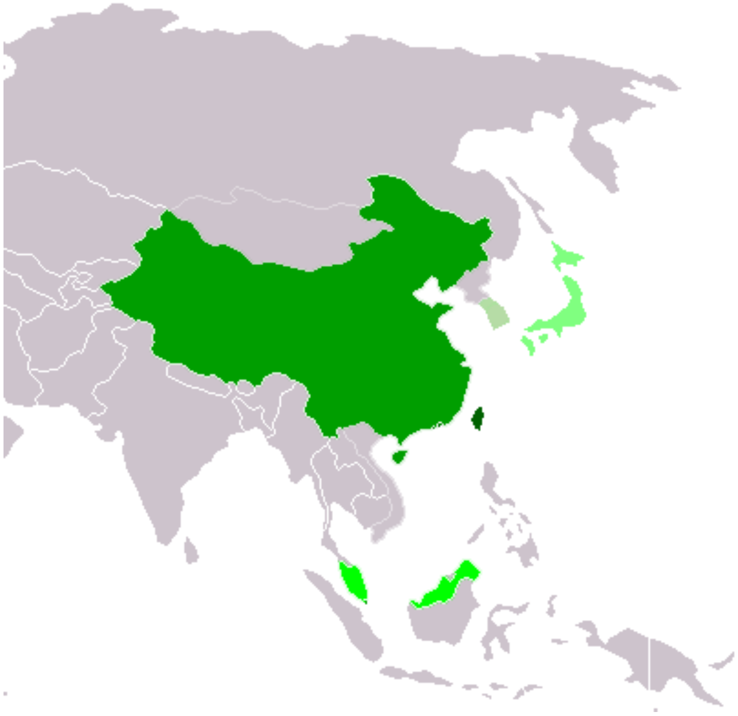


Figure 2: Chinese script spread to Japan and Korea

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.

## 3.2 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<br>(Malaysia) |
| 🟩 | Chinese script used with other systems of writing in the same language Kanji (Japan) |
| 🟩 | Chinese script daily used but no longer officially used<br>Hanja (Republic of Korea) |

Figure 3: Countries using Chinese script

## 3.3 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)<br>cjy (Jinyu Chinese) | 汉字 Hanzi | China Mainland |

| | | | |
|---|---|---|---|
| | cmn (Mandarin Chinese)<br>cpx (Pu-Xian Chinese)<br>czh (Huizhou Chinese)<br>czo (Min Zhong Chinese)<br>gan (Gan Chinese)<br>hak (Hakka Chinese)<br>hsn (Xiang Chinese)<br>mnp (Min Bei Chinese)<br>nan (Min Nan Chinese)<br>wuu (Wu Chinese)<br>yue (Yue Chinese)<br>zho (Chinese) | | Taiwan<br>Hong Kong<br>Macao<br>Singapore<br>Malaysia |
| Japanese | jpn | 漢字 Kanji | Japan |
| Korean | kor | 한자 Hanja | Korea |

- Hanzi normally consists of two overlapping 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), the common used Kanji and Hanja are all therefore included in CGP.

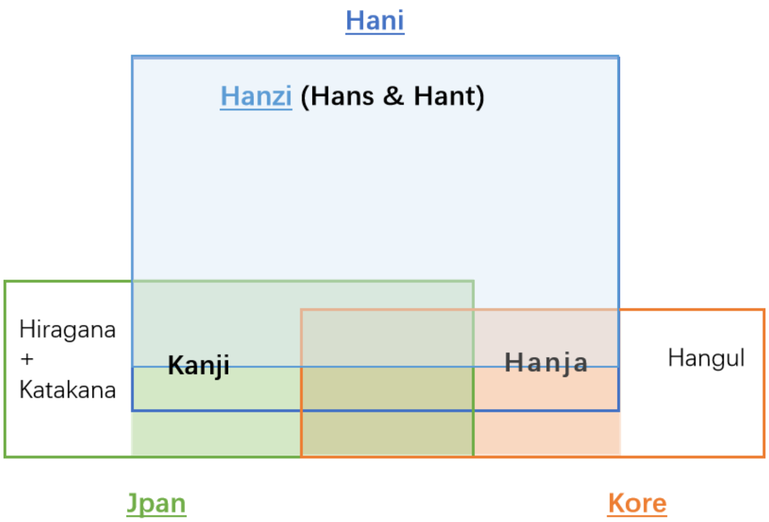


Figure 4: Hanzi, Kanji & Hanja

# 4 Overall Development Process and Methodology

## 4.1 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-name 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]**[2] (later named as version 1.0/2012) , 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. In accordance with [CDNC IDN Table], CNNIC and TWNIC generated and submitted .CN Chinese Character Table[3] and .TW Chinese Character Table[4] separately.

Meanwhile, dotAsia, the registry of .ASIA and a 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]**[5] under the framework of CDNC rules, to cover needs from the Hong Kong and Singaporean local communities.

Over a decade of operating experience has indicated that CDNC's TC-SC Equivalence solution is a market-proven successful practice for handling Chinese variants in domain names.

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.

## 4.2 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.

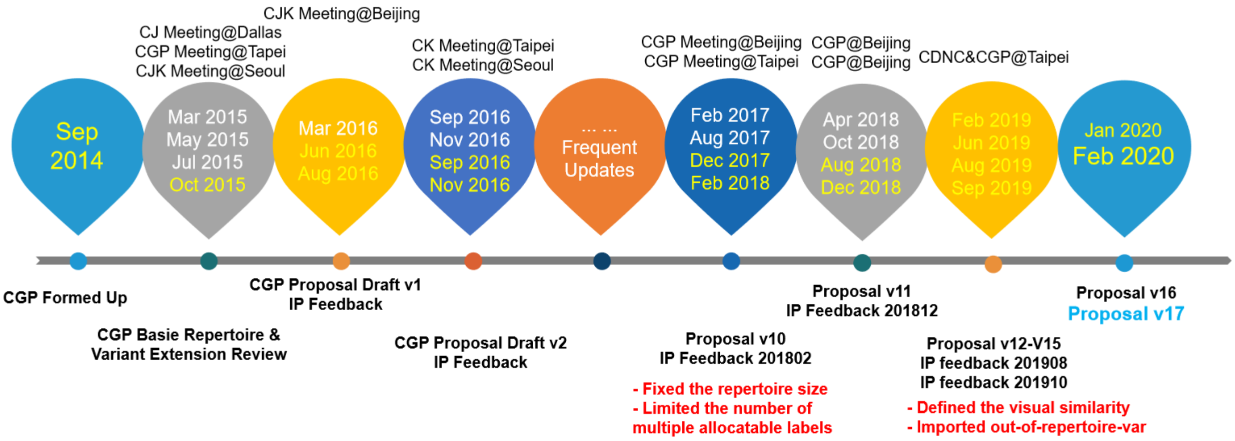
Geographically, the CGP has members from Chinese language regions across East Asia, including China mainland, Taiwan, Hong Kong, Macau, Singapore, Malaysia, as well as members from Europe and North America, totally 23 members belonging to 10 countries/regions listed in Appendix A.

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 its IDN advisor.

## 4.3 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 the face-to-face meetings along with the CDNC meeting, in July 2015, March 2016, Feb 2017, Aug 2017, April 2018 and Oct 2018. 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, November 2016, April 2018, 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.

- **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 variant mappings corresponding to the same Hanzi character, thus the CGP works closely with JGP, KGP and IP to generate the most compatible variant sets (e.g. to import J-only Kanji as out-of-repertoire variant).

- **Define and finalize the variant disposition**

The variant dispositions in CGLR  follow 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 while some multiple preferred variant mappings may work for SLD they 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 variant labels to a reasonable number (for example, Five).

Moreover, to minimize the risk of domain name abuse at the TLD level, the CGP defines the visual identical variants and the corresponding label disposition which do not exist in the CDNC or dotAsia SLD rules

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

The CGP creates the CLGR in XML format following the RFC7940. Considering the fact that the coordination on repertoire, variant mappings and variant label dispositions between CJK and IP required frequent feedback, the CGP work has been carried out in a fast iteration model as indicated in the following figure:

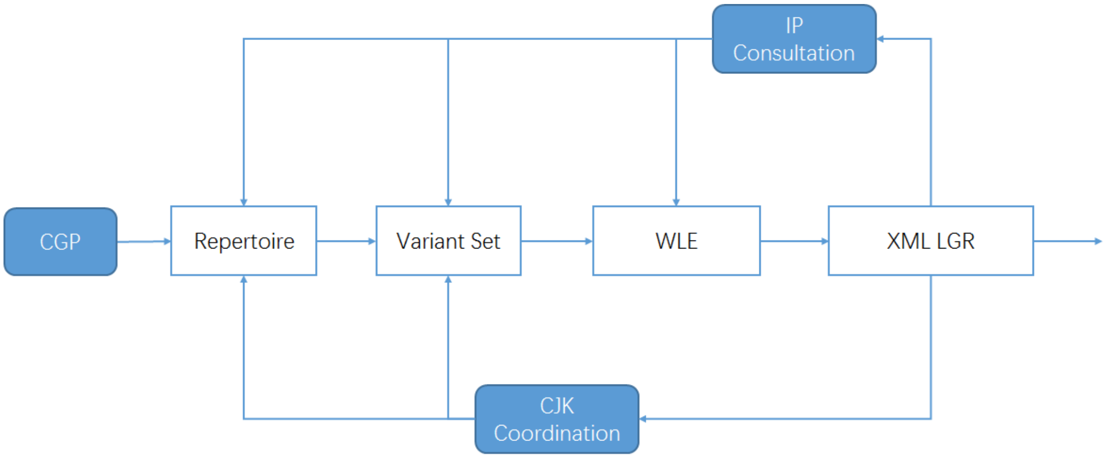


Figure 5: Iteration model of CGP work process

# 5 Repertoire

## 5.1 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[3] and .TW Chinese Character Table[4] 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).

## 5.2 Repertoire formation process

### 5.2.1 19563 Basic Repertoire

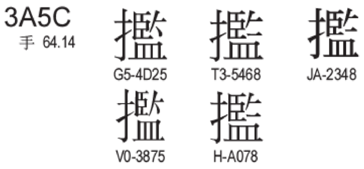
In October 2018, CDNC generated the latest version of its IDN Table [CDNC IDN Table 2.0/2018] as Appendix G or http://www.cdnc.asia/file/unicode-1-2.txt.

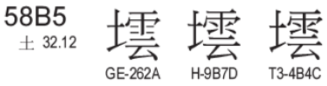
43 new Chinese characters were added into the character set as requested by HKIRC on behalf of the Hong Kong local community from 2013 to 2018, increasing the number of Chinese characters to 19,563.

| Unicode | Hanzi | CDNC | dotAsia | JGP | KGP | IICore | G | T | J | H | K | M | KP | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 52A4 | 劤 | 2013 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 52C5 | 勅 | 2013 | .ASIA | JGP | KGP | IICORE | | | J1A | | K0A | | P0A | A |
| 53DA | 叚 | 2013 | .ASIA | | | IICORE | | T3B | | | | | | C |
| 57D7 | 埗 | 2013 | .ASIA | | | IICORE | | | | H1E | | M1B | | B |
| 58DC | 壜 | 2013 | .ASIA | JGP | | IICORE | | | J1A | | | | | C |
| 5BD7 | 寗 | 2013 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 5FDF | 忟 | 2013 | .ASIA | | | IICORE | | | | H1F | | | | C |
| 617D | 慽 | 2013 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 6407 | 搇 | 2013 | .ASIA | | | IICORE | | | | H1F | | | | C |
| 64E1 | 擡 | 2013 | .ASIA | JGP | KGP | IICORE | | | | | K0A | | P0A | A |
| 64E5 | 擥 | 2013 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 661E | 昞 | 2013 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 6900 | 椀 | 2013 | .ASIA | JGP | KGP | IICORE | | | J1A | | K0A | | P0A | A |
| 6AC8 | 櫈 | 2013 | .ASIA | | | IICORE | | T3B | | | | | | C |
| 6C39 | 氹 | 2013 | .ASIA | | | IICORE | | | | H1E | | M1A | | B |
| 3A5C | 㩜 | 2013 | | | | | | | | | | | | |
| 58B5 | 墵 | 2013 | | | | | | | | | | | | |
| 6EDD | 滝 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 734F | 獏 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | C |
| 758E | 疎 | 2015 | .ASIA | JGP | KGP | IICORE | | | J1A | | K0A | | P0A | A |
| 764E | 癎 | 2015 | .ASIA | | KGP | IICORE | | | | H1F | K0A | | P0A | A |
| 767A | 発 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 76CC | 盌 | 2015 | .ASIA | | | IICORE | | | J1A | | | | | C |
| 7B6F | 筯 | 2015 | .ASIA | | | IICORE | | T3B | | | | | | C |
| 7B92 | 箒 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | C |
| 7C83 | 粃 | 2015 | .ASIA | JGP | KGP | IICORE | | | | | K0A | | P0A | A |
| 7DDC | 緜 | 2015 | .ASIA | JGP | | IICORE | | T3B | | | | | | C |
| 8117 | 脗 | 2015 | .ASIA | | | IICORE | | | | H1F | | | | C |
| 84DA | 蓚 | 2015 | .ASIA | JGP | KGP | IICORE | | | | | K0A | | P0A | A |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8597 | 薗 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 89A9 | 覩 | 2015 | .ASIA | JGP | KGP | IICORE | | | | | K0A | | P0A | A |
| 8EE2 | 転 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 994D | 饍 | 2015 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 9D44 | 鵄 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | C |
| 9F62 | 齢 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 68C5 | 棅 | 2015 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 6A53 | 橓 | 2015 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 7200 | 爀 | 2015 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 73E1 | 珡 | 2015 | .ASIA | | | IICORE | | T3G | | | | | | C |
| 73E4 | 珤 | 2015 | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 8FBA | 辺 | 2015 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |
| 681E | 栞 | 2018 | .ASIA | JGP | KGP | IICORE | | | J1A | | K1C | | | B |
| 99C5 | 駅 | 2018 | .ASIA | JGP | | IICORE | | | J1A | | | | | A |

Among the 43 Hong Kong characters, two characters (3A5C㩜 and 58B5 墵) were out of scope of MSR-2.





As requested by CGP, they have been formally accepted into MSR-4.

Thus, all **19563** CDNC code points form up the basic set of the CGP repertoire. .

### 5.2.2 122 dotAsia characters

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. These 19563 code points are all included in MSR-4. (https://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt)

41 of the newly added 163 characters were already included in CDNC IDN Table 2018, the remaining 122 extend the CGP repertoire up to **19,685** code points.

The following subsections break down these 122 characters into three sections.

#### *5.2.2.1 53 characters located in the Basic Multilingual Plane as well as in IICORE:*

| Unicode | Hanzi | CDNC | dotAsia | JGP | KGP | IICore | G | T | J | H | K | M | KP | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 65FF | 旿 | | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 4C81 | 䲁 | | .ASIA | | | IICORE | | T4B | | | | | | C |
| 5605 | 嘅 | | .ASIA | | | IICORE | | | | H1F | | M1F | | B |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6335 | 挵 | | .ASIA | | | IICORE | | T3B | | | | | | C |
| 656D | 敭 | | .ASIA | | KGP | IICORE | | | | | K0A | | P0A | A |
| 7460 | 瑠 | | .ASIA | JGP | KGP | IICORE | | T3D | J1A | | K0A | | P0A | A |
| 74C8 | 瓈 | | .ASIA | | | IICORE | | T3G | | | | | | C |
| 9771 | 靱 | | .ASIA | 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 |

### 5.2.2.2 7 characters located in the Basic Multilingual Plane, but not in IICORE

| Unicode | Hanzi | CDNC | dotAsia | 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 | | | | | | | | | | | |

### 5.2.2.3 62 code points from Supplementary Ideographic Plane (Plane 2)

| Unicode | Hanzi | CDNC | dotAsia | 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 |

### 5.2.3 CLGR Repertoire of 19685 characters

Finally, CGP generated the repertoire (19685 characters) through the steps from 5.2.1 to 5.2.2.3, using the formation process illustrated in the following figure:

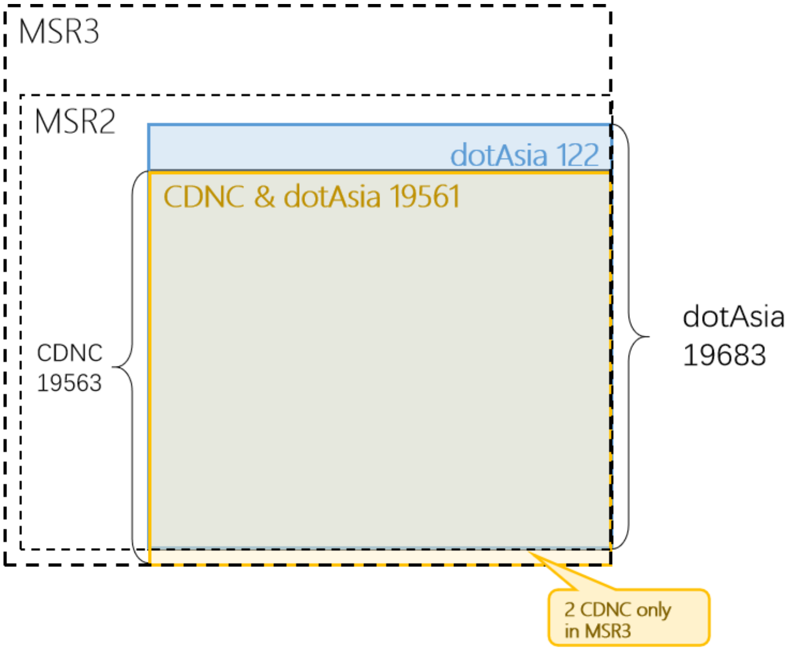


Figure 6: CGP repertoire components

The CDNC IDN Table (version 2018) has 19563 Hani/Hanzi characters, all included in CGP Repertoire.

The dotAsia IDN Table (version 1.1) has 19683 Hani characters, all included in CGP Repertoire.

### 5.2.4 Source information and "out of repertoire" characters

To better illustrate the source information of each character in CGP repertoire, the CGP provided the info of whether the character is included in CDNC table, dotAsia table, JGP, KGP and  IICore set, as well as its IICore region value.

Given that the variant mappings are required to respect the implicit definitions for each LGR's native users, and that each integrated LGR will implement the superset of the variant sets applicable to the LGR, thus the CGP reviewed the J-only and K-only Hani characters as regards their variant mappings to the characters in CGP repertoire, and decided to import the necessary "out-of-repertoire-var" code points into CLGR.

The JGP repertoire (version 201703, Appendix B) has 6356 Hani/Kanji characters, among which 6212 are overlapped characters in the CGP repertoire. Of the remaining 144 Kanji characters, CGP imported 76 of them into the CLGR as "out-of-repertoire-var" code points as expressed in Section 6.2.2.

The KGP repertoire (version 201703, Appendix C) has 4758 Hani/Hanja characters, among which 4744 are overlapped with characters in the CGP repertoire. Similarly, CGP imported 4 additional K-only Hanja characters as "out-of-repertoire-var" code points into CLGR as expressed in Section 6.2.2 (one out-of-repertoire-var character is shared between J and K).

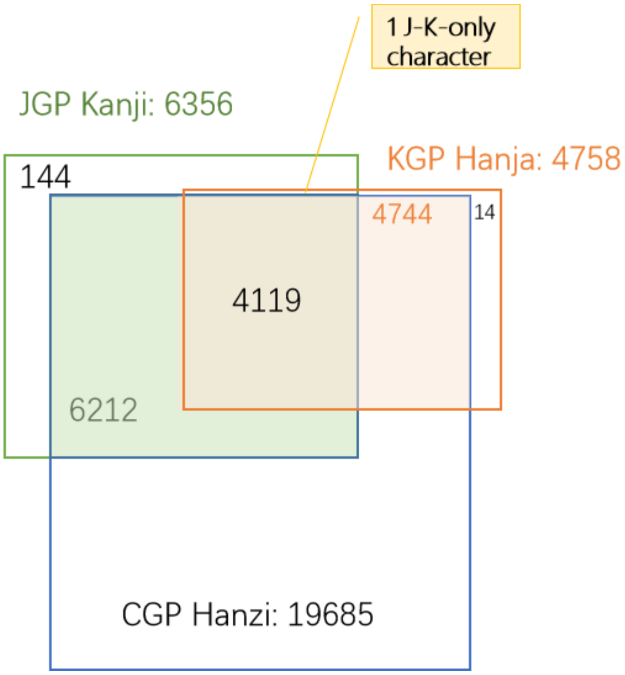


Figure 8 : Source of CGP repertoire

CGP provides the detailed source information of CGP repertoire in Appendix A, and lists out the source information of "out-of-repertoire-var" code points in Appendix O.

## 5.3 Attempt to limit the size of the repertoire

In Section 5.2, the CGP generated a repertoire containing 19,685 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 a limited repertoire of graphemes to represent the phonemes (basic units of sound) of a language, or a syllabary (such as Kana), which has a limited repertoire of graphemes to represent syllables or moras, a logographic writing system has glyphs/logograms to represent words or morphemes rather than phonetic elements. In Chinese, a logogram is a single written character that represents a complete grammatical word (or, more precisely, a 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 oracle bones of the Shang Dynasty (16th-11th century B.C.) already included 3500-4500 characters. 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[1] variant character (in CDNC IDN Table, 1 non-reflexive variant on average, at most 7 non-reflexive variants).

Statistically, "Text Notes and Word Explanations 说文解字/說文解字" from the Han Dynasty (202 B.C.-220 A.D.) includes 9,353 characters, and "Lei Pian 类篇/類篇" in Song Dynasty (960-1279 A.D.) includes 31,319 characters. In 1710, Emperor Kangxi released the "Kangxi Dictionary 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 Education released "Dictionary of Chinese Character Variants 異體字字典" 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 3,500 characters after nine years of compulsory education, the number is 3,500-4,500 in Taiwan and 3,500 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 symbols 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 barriers.

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 )

[1] A reflexive variant maps the character to itself. See Section (6.1).

has 6 variants with similar visual forms, 鬪(9B2A)閗(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/头髮 tóufǎ 'head hair' and 发展/發展 fāzhǎn 'develop & expand')

The above two natural characteristics give Chinese variant characters great acceptability, usability and exchangeability in real life, especially in information systems. Hence, the development and popularization of the internet promote Chinese character usage in cyberspace. In terms of Computer Coding Standards, 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 6,763 characters, while the latest standard GB18030 included 20,912 characters. The current Unicode standard (as of 10 October 2019), including CJK Unified Ideographs Extensions A-F, contains 87,887 CJK Unified ideographs. 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 8,128 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 20,923 characters are used.

The most symbolic event occurred in 2016: China's Ministry of Civil Affairs 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**

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 compatibility, 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 sets)
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, to generate 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 complexity of the LGR and speed up the coordination work with J & K. In 2015, 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. Some other registries, like dotAsia, extended CDNC IDN Table by adding more local characters from Unicode Supplementary Ideographic Plane (Plane 2). 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 sets.

For all above reasons, the CGP decided to define CGP Repertoire as the conflation of CDNC table and dotAsia table, a character set with high capacity and compatibility, to ensure the consistency of user experiences, registry practices as well as the local regulations .

# 6 Variants

## 6.1 Variant definition in CLGR

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 Section 4.1, CHINESE (CHARACTER) VARIANTS are defined as:

**"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, which means, the code point is its own preferred variant. In others, a code point has no reserved variant.

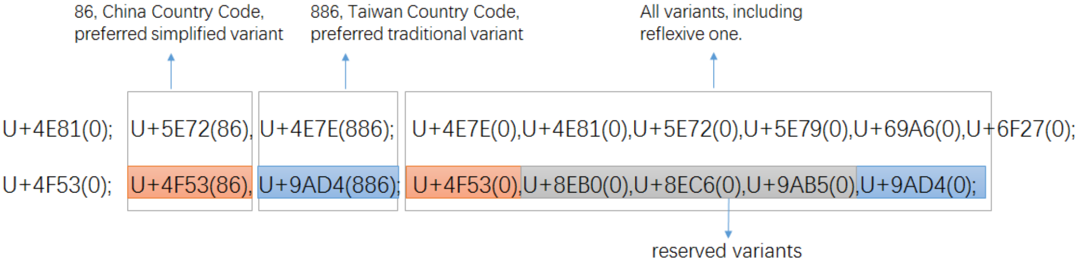

86, China Country Code,
preferred simplified variant
886, Taiwan Country Code,
preferred traditional variant
All variants, including
reflexive one.
U+4E81(0); U+5E72(86), U+4E7E(886); U+4E7E(0),U+4E81(0),U+5E72(0),U+5E79(0),U+69A6(0),U+6F27(0);
U+4F53(0); U+4F53(86), U+9AD4(886); U+4F53(0),U+8EB0(0),U+8EC6(0),U+9AB5(0),U+9AD4(0);
reserved variants

Figure 9: variant setting in CDNC IDN Table

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

| 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 |

According to the XML schema, the two variant mappings in Figure 8 will be transformed into the following XML 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>
```

Note: A reflexive variant maps the code point to itself. The type of the reflexive mapping determines which category of allocatable variant labels the code point may be part of. In the XML format, reflexive variant mapping types for the CLRG by convention use an "r-" prefix.

Note: To eliminate the overproduction of allocatable labels caused by multiple allocatable variant mappings, CGP created some new sub-types of the "allocatable" variant type, the related definitions and variant dispositions are illustrated in Section 6.3.2 and Section 6.3.3.

## 6.2 Variant Mappings formation process

### 6.2.1 CGP internal coordination

#### *6.2.1.1 19498 basic variant mapping entries from CDNC-2018*

The CDNC IDN Table (repertoire and variant mappings), generated in the early 2000s and extended in 2018 (Appendix G), along with RFC3743 and RFC 4713, represents 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 [CDN IDN Table]. Considering all these factors, CGP directly borrowed all 19563 variant mapping entries in the CDNC IDN Table to generate the CLGR variant mappings.

However, among all 19,563 CDNC characters and their variant mapping entries, there are a few variant mappings that were changed later due to the CGP internal review in Section 6.2.1.3 and IP feedback in Section 6.2.1.4.

Finally, in this proposed LGR, 19432 variant mapping entries are kept the same as CDNC-2018 and dotAsia, while 66 entries are the same as CDNC-2018 but different from dotAsia. These **19,498** (19432+66) variant mappings constitute the basic CGP variant mappings table listed in Appendix D [Sheet "6.2.1.1-19432"] and Appendix D [Sheet "6.2.1.1-66"].

#### *6.2.1.2 143 unique variant mapping entries 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. dotAsia extended the CDNC IDN table 2012, by adding 163 new regional characters and modifying some existing variant mappings according to local requirement.

43 of them having been supplemented into CDNC-2018, only 122 are dotAsia unique characters. CGP also adopted these 122 characters and their variant mapping entries into this LGR proposal.

CGP and Edmon CHUNG, the CEO of dotAsia, discussed the issue of inconsistency between CDNC 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.

In September 2015, CGP & CDNC held joint meetings and invited linguistic experts from China mainland, Taiwan and Hong Kong, reviewed 172 IICORE characters, including 53 unique dotAsia characters in section 5.2.2.1 and , then reset the variant mappings of them as Appendix E [Sheet "172 chars reviewed in 2015"] . 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-2015 IDN table, 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 [Sheet "7 .asia chars reviewed 2016"]. Correspondingly, CGP rechecked and altered some 56 additional variant mapping entries related to the above 62 (55+7) characters.

For the other 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.

In this proposed LGR, there are 19575 variant mapping entries that are the same as the dotAsia IDN table, including 19432 entries that are the same as CDNC-2018 and dotAsia, as well as 143 unique dotAsia entries. Among these 143 entries, 38 are kept the same as dotAsia but different from CDNC-2018, while 43 Non-CDNC-2018 dotAsia entries are unchanged, and a further 62 Unicode Plane B dotAsia entries are also unchanged.

These **143** (38+43+62) variant mappings constitute the unique dotAsia variant mappings table listed in Appendix D [Sheet "6.2.1.2-38"], Appendix D [Sheet "6.2.1.2-43"] and Appendix D [Sheet "6.2.1.2-62"].

#### *6.2.1.3 38 variant mappings revised by CGP internal review team*

In September 2015, CGP invited linguistic experts from mainland China, Taiwan and Hong Kong, to review 172 IICORE characters (including 55 unique dotAsia Hanzi, 93 J-only Kanji, 13 K-only Hanja and 1 J&K character) as Appendix E [Sheet "172 chars reviewed in in 2015"].

In May 2016, CGP reviewed additional 7 dotAsia characters as Appendix E [Sheet "7 .asia chars reviewed in 2016"].

In Dec 2019, CGP reviewed the rest 50 J-only kanji characters as Appendix E [Sheet "50 J-only reviewed in 2019"].

Among all 229 reviewed characters, 62 dotAsia unique Hanzi and 118 associated variant mapping entries were reviewed (Appendix E [62 dotAsia+ 56 associated]. Among the 118 variant mapping entries, 80 are unchanged or the same as dotAsia or CDNC-2018, while the remaining 38 variant mapping entries should be considered as the result of CGP internal review work, listed in Appendix D [Sheet "6.2.1.3-38"].

#### 6.2.1.4 6 variant mappings changed by IP review

IP reviewed the CGP Proposal draft in 2019 and proposed the 28 variant mapping entries listed as "Appendix F IP External Review", among which CGP accepted 17 entries and rejected 11 entries. The accepted 11 entries are the same as dotAsia IDN Table, therefore there are 6 entries listed as changed by the IP review in Appendix D [Sheet "6.2.1.4-6"].

### 6.2.2 CJK coordination and 80 "out-of-repertoire" variants

A coordination mechanism among three parties is needed to realize unified Chinese script generation rules in the DNS root zone. At the CDNC meeting in Shanghai (May, 2014), CJK agreed to take the below principles of 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 shall be discussed across 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 since the IETF Dallas meeting in 2015, trying to define a unified variant-mapping table for Chinese scripts.

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 set 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 sets into the JGP ones. Thus, both parties eliminated the potential conflict caused by variant set inconsistency. The CGP would like to express its appreciation for the JGP's openness, tolerance and compromise. Reciprocally, CGP reviewed all 144 J-only characters and imported 76 (56 in 2015, 20 in 2019) Kanji as "out-of-repertoire-var" code points in the CLGR XML document and as listed in Appendix E [Sheet "76 OORV imported from JGP"].

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 this CGP proposal, from the most conservative perspective, CGP also reviewed 14 K-only characters and imported 5 Hanja as “out-of-repertoire-var” code points in the CLGR XML document and as listed in Appendix E EXCEL document [Sheet "5 OORV imported from KGP"]. Because one of this out-of-repertoire-var code points is shared between Kanji and Hanja, this results in only 80 out-of-repertoire-var code points.

The disposition of “out-of-repertoire-var ” will be invalid and blocked as the following example:

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 51E6 | 処 | 51E6 | 処 | ≡ | out-of-repertoire-var | | identity |
| 51E6 | 処 | 5904 | 处 | ↔ | blocked | | |
| 51E6 | 処 | 8655 | 處 | ↔ | blocked | | |
| 5904 | 处 | 5904 | 处 | ≡ | r-simp | | identity |
| 5904 | 处 | 8655 | 處 | →<br>← | trad<br>simp | | |
| 8655 | 處 | 8655 | 處 | ≡ | r-trad | | identity |

After importing 51E6 処 from JGP and setting up variant mappings to the existing CGP variant set (5904 处 and 8655 處), the XML description of these 3 code points will become:

```
<char cp="5904" tag="sc:Hani" ref="0 100 200 300 600">
        <var cp="5904" type="r-simp" comment="identity" ref="101 201"/>
        <var cp="8655" type="trad" ref="101 201"/>
        <var cp="51E6" type="blocked"/>
</char>
<char cp="51E6" tag="sc:Hani" ref="400" comment="not part of repertoire">
        <var cp="51E6" type="out-of-repertoire-var" comment="identity"/>
        <var cp="5904" type="blocked"/>
        <var cp="8655" type="blocked"/>
</char>
<char cp="8655" tag="sc:Hani" ref="0 100 200 300 400 500 600">
        <var cp="5904" type="simp" ref="101 201"/>
        <var cp="8655" type="r-trad" comment="identity" ref="101 201"/>
        <var cp="51E6" type="blocked"/>
</char>
```

### 6.2.3 19685 Variant mappings' source information

After all above steps from Section 6.2.1 to 6.2.2, CGP finalized a list of 19685 CLGR15 variant mappings corresponding to the  CLGR15 repertoire as Appendix D [Sheet All Variant Mappings], consisting of 19498 basic variant mappings, 143 unique dotAsia variant mappings, 38 variant mappings from the CGP internal review, and 6 from IP review proposal.

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

- 19498 variant mapping entries consistent with the existing practice of CDNC-2018.
- 19575 variant mapping entries consistent with the existing practice of dotAsia.
- 115 variant mapping entries consistent with CGP internal review
- 165 variant sets (NOT the variant mapping) consistent with KLGR.

- 20 variant mapping entries same as IP's suggestions

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.

- 187 variant mappings different from the existing practice of CDNC (including 122 Non-CDNC chars)
- 110 variant mappings different from the existing practice of dotAsia (including 2 Non-dotAsia chars)

## 6.3 Effort to reduce the number of multiple allocatable labels

In the Chinese writing system, quite a few characters 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 | Allocatable Simplified Char | Allocatable Traditional Char |
|---|---|---|---|
| 4E30 | 丰(4E30) | 丰(4E30) | 丰(4E30)豐(8C50) |
| 8C50 | 豐(8C50) | 丰(4E30) | 豐(8C50) |
| Unicode | Original Char | Preferred Simplified Char | Preferred Traditional Char |
| 53F0 | 台(53F0) | 台(53F0) | 台(53F0)檯(6AAF)臺(81FA)颱(98B1) |
| 6AAF | 檯(6AAF) | 台(53F0) | 檯(6AAF) |
| 81FA | 臺(81FA) | 台(53F0) | 臺(81FA) |
| 98B1 | 颱(98B1) | 台(53F0) | 颱(98B1) |

Multiple preferred variant mapping examples

"丰台"is the simplified geo name of a district under Beijing Municipality, it has 8 allocatable traditional labels according to the CGP variant mappings in Section 6.2.

Original Label: 丰台(4E30)(53F0)

Allocatable Simplified Label: 丰台(4E30)(53F0)

Allocatable Traditional Labels: 丰台(4E30)(53F0)丰檯(4E30)(6AAF)丰臺(4E30)(81FA)丰颱(4E30)(98B1)
豐台(8C50)(53F0)豐檯(8C50)(6AAF)豐臺(8C50)(81FA)豐颱(8C50)(98B1)

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 selection is complete, all the other allocatable labels are reserved, the reserved allocatable labels could be reactivated later at the request of an applicant, to make sure the applicant could get all his desired labels.

However, unlike the SLD practice, according to Root Zone LGR framework, there are no provisions for "human select" or "reserve and reactivate", 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 to the maximum degree.

CGP checked the variant mappings in Appendix D, found that there are only 194 multiple allocatable variant mappings out of all 19685 cases, 10 with 2 ASVs (Allocatable Simplified Variants), 173 with 2

ATVs (Allocatable Traditional Variants), 9 with 3 ATVs and 2 with 4 ATVs. These 164 multiple variant mappings are listed in Appendix K. Having analyzed the 194 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.

### 6.3.1 Identify and classify the "redundant/muted" variant mappings

The 194 multiple allocatable variant mappings are divided into 7 categories:

| Category | number | Original | Allocatable Simp | Allocatable Trad |
|---|---|---|---|---|
| Case 1 | 7 | A | AB | A |
| Case 2 | 1 | A | AB | C |
| Case 3 | 2 | A | BC | A |
| Case 4 | 146 | A | A | AB |
| Case 5 | 27 | A | A | BC |
| Case 6 | 9 | A | A | ABC |
| Case 7 | 2 | A | A | ABCD |

Examples:

| Category | Original | Allocatable Simp | Allocatable Trad |
|---|---|---|---|
| Case1 | 乾(4E7E) | 乾(4E7E)干(5E72) | 乾(4E7E) |
| Case 2 | 麽(9EBD) | 麽(9EBD)么(4E48) | 麼(9EBC) |
| Case 3 | 餘(9918) | 馀(9980)余(4F59) | 餘(9918) |
| Case 4 | 丰(4E30) | 丰(4E30) | 丰(4E30)豐(8C50) |
| Case 5 | 冲(51B2) | 冲(51B2) | 沖(6C96)衝(885D) |
| Case 6 | 升(5347) | 升(5347) | 升(5347)昇(6607)陞(965E) |
| Case 7 | 台(53F0) | 台(53F0) | 台(53F0)檯(6AAF)臺(81FA)颱(98B1) |

By case analysis and simulation computation, CGP found that these 194 variant mappings could be transferred to the following format without causing negative impact to TLD applicant.

| Case | number | Original | Allocatable Simp | Allocatable Trad |
|---|---|---|---|---|
| Case1 | 7 | A | A (muted, reflexive )<br>B | A |
| Case2 | 1 | A | A (muted, reflexive )<br>B | C |
| Case3 | 2 | A | B (simp type1)<br>C (simp type2) | D |
| Case4 | 146 | A | A | A (muted, reflexive)<br>B |
| Case5 | 27 | A | A | B (trad type1)<br>C (trad type2) |
| Case6 | 9 | A | A | A (muted, reflexive)<br>B (trad type1) |

| | | | | |
|---|---|---|---|---|
| | | | | C (trad type2) |
| Case7 | 2 | A | A | A (muted, reflexive)<br>B (trad type1)<br>C (trad type2)<br>D (muted, infrequent) |

The "muted"&"reflexive" variant is deemed redundant, the label that contains it is the same as the original label, and hence is safe to be BLOCKED.

The "muted"&"infrequent" variant is rarely used character, not included either in Modern Chinese Common Used Table in China mainland or Common used Chinese standard table in Taiwan. Therefore, the label that contains it is safe to be BLOCKED.

The "simp type1" and "simp type2" variant characters will be treated as two mutually exclusive subgroups, which means, the mixture of "simp type1" and "simp type2" will be BLOCKED as redundant. If a specific mixed label happens to be the desired one, the applicant could input this specific label as the original label, at the cost of losing some "less desired" simplified label.

The "trad type1" and "trad type2" variant characters will be treated as two mutually exclusive subgroups, which means, the mixture of "trad type1" and "trad type2" will be BLOCKED as redundant. If a specific mixed label happens to be the desired one, the applicant could input this specific label as the original label, at the cost of losing some "less desired" traditional label.

### 6.3.2 Create 4 new types for multiple mapping variants

According to the design in Section 6.3.1 and 6.3.2, CGP proposed 8 new sub-types to identify the corresponding variant characters in multiple mappings in Appendix K.

In practice, only 4 new types are created as initially proposed. The other 4 are mapped to the sub-types defined in Section 6.1, because their dispositions are functionally equivalent to the existing sub-types; a "comment" value is used to show a variant's proposed sub-type and the reason of "being muted"[2], as well as be a reference to guide other teams to make their own label generation rules at SLD in the future.

| Proposed | Disposition | Description | Implemented |
|---|---|---|---|
| sub-type:<br>r-both-ms | Blocked | *r-both-ms indicates that for a given code point, its reflexive type is inherently r-both, but there is at least another 'simp' type (or other simplified types), and therefore it is preferred in a traditional context. Therefore, it is to be treated as a 'r-trad'.*<br>A simp label containing "r-both-ms" char is BLOCKED | sub-type:<br>r-trad<br>comment:<br>r-both-ms |
| | Allocatable | A trad label containing "r-both-ms" char is ALLOCATABLE<br>An original label containing "r-both-ms" is ALLOCATABLE | |

[2] "muted" means the variant is allocatable but not preferred and therefore shall not be used to generate allocatable labels. This is normally indicated by a "-m" in the name of the subtype, but, as explained, these subtypes are functionally equivalent to other, already existing subtypes. Therefore, the equivalent subtype is used and the original name with "-m" is given in a comment.

| Proposed | Disposition | Description | Implemented |
|---|---|---|---|
| sub-type:<br>r-both-mt | Blocked | *r-both-mt indicates that for a given code point, its reflexive type is inherently r-both, but there is at least another 'trad' type (or other traditional types), and therefore it is preferred in a simplified context. Therefore, it is to is treated as a 'r-simp'.*<br>A trad label containing "r-both-mt" char is BLOCKED | sub-type:<br>r-simp<br>comment:<br>r-both-mt |
| | Allocatable | A simp label containing "r-both-mt" char is ALLOCATABLE<br>An original reflexive label containing "r-both-mt" is ALLOCATABLE | |
| sub-type:<br>r-simp-ms | Blocked | *r-simp-ms indicates that for a given code point, its reflexive type is inherently r-simp, but there is at least another 'simp' type (or other simplified types), along with another 'trad' type and therefore it is never preferred in any variant labels. Therefore, it is to be treated as a 'r-neither'.*<br>A simp label containing "r-simp-ms" is BLOCKED | sub-type:<br>r-neither<br>comment:<br>r-simp-ms |
| | Allocatable | An original label containing "r-both-ms" is ALLOCATABLE | |
| sub-type:<br>trad-mt | Blocked | *Allocatable trad is rarely used, not in Modern Chinese Common Used Table in China mainland, nor Common used Chinese standard table in Taiwan. Set the allocatable trad as "trad-mt" (muted) and treat it as a 'blocked'.*<br>A trad label containing "trad-mt" is BLOCKED | sub-type:<br>blocked<br>comment:<br>trad-mt |
| sub-type:<br>simp-1 | Allocatable | *Among the multiple allocatable simplified variants, set the allocatable simp with the smallest hex-code as "simp-1"*<br>A simp label containing "simp-1" is ALLOCATABLE | sub-type:<br>simp-1 |
| | Blocked | A simp label containing BOTH "simp-1" and "simp-2" is BLOCKED | |
| sub-type:<br>simp-2 | Allocatable | *Among the multiple allocatable traditional variants, set the allocatable simp with the largest hex-code as "simp-2"*<br>A simp label containing "simp-2" is ALLOCATABLE | sub-type:<br>simp-2 |
| | Blocked | A simp label containing BOTH "simp-1" and "simp-2" is BLOCKED | |
| sub-type:<br>trad-1 | Allocatable | *Among the multiple allocatable traditional variants, set the allocatable trad with the smallest hex-code as "trad-1"*<br>A trad label containing "trad-1" is ALLOCATABLE | sub-type:<br>trad-1 |
| | Blocked | A trad label containing BOTH "trad-1" and "trad-2" is BLOCKED | |
| sub-type:<br>trad-2 | Allocatable | *Among the multiple allocatable traditional variants, set the allocatable trad with the largest hex-code as "trad-2"*<br>A trad label containing "trad-2" is ALLOCATABLE | sub-type:<br>trad-2 |
| | Blocked | A trad label containing BOTH "trad-1" and "trad-2" is BLOCKED | |

Theoretically, given any valid input label, the optimized rules will generate at most 5 ALLOCATABLE labels -- the original label, one all simp1 label, one all simp2 label, one all trad-1 label and one all trad-2 label.

CGP provides two examples to illustrate how the mechanism reduces the number of allocatable labels in Appendix R. In the majority of cases for which there aren't multiple trad or multiple simp variants, the rules will only generate 3 ALLOCATABLE labels.

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 164 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.

### 6.3.3 Create new actions to assign variant label dispositions

According to the designs in Section 6.3.2, CGP created some new "actions" to reduce the number of multiple allocatable labels, keeping the number to a maximum of five (5).

```
<action disp="allocatable" only-variants="simp r-simp both r-both simp-1" comment="all simplified label type 1" />
<action disp="allocatable" only-variants="simp r-simp both r-both simp-2" comment="all simplified label type 2" />

<action disp="allocatable" only-variants="trad r-trad both r-both trad-1" comment="all traditional label type 1"/>
<action disp="allocatable" only-variants="trad r-trad both r-both trad-2" comment="all traditional label type 2"/>

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

# 7 Visually Identical Characters

The term 'identical' can be understood both as semantically identical and visually identical. Being 'exchangeable from the point of view of the user community' may be based on semantic identity (as traditional versus simplified in Hanzi) but also on visual identity (for example in the case of U+0259 ə and U+01DD ǝ in the Latin script). Both types of identity can be understood as making the code points 'identical' from a user point of view when using domain labels

CGP acknowledges that the concept of visual variant has been considered by many other GPs working on Root Zone LGRs. However, traditionally, the Chinese language and script community regard only "semantically identical characters" as exchangeable variants. The corresponding Chinese Variant definition is formally stated in the Chinese Case Study Team Report (https://www.icann.org/news/announcement-2011-10-03-en), Section 2.1:

**"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."**

Based on this principle and definition, Chinese community carried out the research and practice work since 1998, generated the Chinese variant character list as CDNC IDN table and dotAsia IDN table.

Due to the nature of the Chinese writing system, 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. The semantically identical Chinese variants generally have visually similar forms (like 4443䑃 and 6726 朦), but the reverse is not true, some visually similar Chinese characters have totally different semantics (like 58AB[墫]58FF[壿]), typically, the Chinese script users don't treat these characters as exchangeable variants.

In the CGP-CDNC joint meetings, April 2018 and Oct 2018, the Chinese community members disputed the concept of visual variants, CGP finally acknowledges the lack of visual similarity rules would trigger the risk of massive domain name abuse, and the necessity to generate the rules to minimize such risk. CGP appreciates the Unicode consortium's confusable list (https://www.unicode.org/Public/security/11.0.0/confusables.txt) has listed 45 visually confusable pairs, including 6 Han-Han pairs, whose 12 characters fall in the range of CGP repertoire.

Following the discussion with IP during ICANN'64 and further feedback from IP, CGP proposes to handle the 6 Han-Han pairs as below:

Three pairs including characters out of the scope of "List of Commonly Used Characters in Modern Chinese 现代汉语通用字表"will be treated as visual identical variants

-- 676E 杮 & 67FF 柿、8D7F 赿 & 8D86 趆、58AB 墫 & 58FF 壿

The other three pairs will be treated as unrelated singletons

-- 571F 土 & 58EB 士、9E42 鹂 & 9E43 鹃、53E3 口 & 56D7 囗

Accordingly, the variant set {U+8D7F, U+8D86} has been defined as follows (assuming U+8D86 is more common than U+8D7F)

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 8D7F | 赿 | 8D7F | 赿 | ≡ | r-neither | | identity |
| 8D7F | 赿 | 8D86 | 趆 | → | both | | visual-similarity |
| | | | | ← | blocked | | |
| 8D86 | 趆 | 8D86 | 趆 | ≡ | r-both | | identity |

The variant set {U+676E, U+67BE, U+67FF} resulting of incorporating U+676E into the pre-existing variant set {U+67BE, U+67FF} has been defined as follows: (assuming U+67FF is more common than U+676E)

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 676E | 杮 | 676E | 杮 | ≡ | r-neither | | identity |
| 676E | 杮 | 67BE | 枾 | ↔ | blocked | | visual-similarity |
| 676E | 杮 | 67FF | 柿 | → | both | | visual-similarity |
| | | | | ← | blocked | | |
| 67BE | 枾 | 67BE | 枾 | ≡ | r-neither | | identity |
| 67BE | 枾 | 67FF | 柿 | → | both | | |
| | | | | ← | blocked | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 67FF | 柿 | 67FF | 柿 | ≡ | r-both | | identity |

The joint variant set {58AB, 58FF, 6A3D, 7F47, 8E72} resulting of combining {58AB, 6A3D, 7F47} and {58FF, 8E72} has been defined as follows:

| Source | Glyph | Target | Glyph | | Type | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 58AB | 墫 | 58AB | 墫 | ≡ | r-both | | identity |
| 58AB | 墫 | 58FF | 壿 | ↔ | blocked | | visual-similarity |
| 58AB | 墫 | 6A3D | 樽 | ↔ | blocked | | |
| 58AB | 墫 | 7F47 | 罇 | ↔ | blocked | | |
| 58AB | 墫 | 8E72 | 蹲 | ↔ | blocked | | visual-similarity |
| 58FF | 壿 | 58FF | 壿 | ≡ | r-trad | | identity |
| 58FF | 壿 | 6A3D | 樽 | ↔ | blocked | | visual-similarity |
| 58FF | 壿 | 7F47 | 罇 | ↔ | blocked | | visual-similarity |
| 58FF | 壿 | 8E72 | 蹲 | →<br>← | simp<br>blocked | | |
| 6A3D | 樽 | 6A3D | 樽 | ≡ | r-both | | identity |
| 6A3D | 樽 | 7F47 | 罇 | →<br>← | blocked<br>simp | | |
| 6A3D | 樽 | 8E72 | 蹲 | ↔ | blocked | | visual-similarity |
| 7F47 | 罇 | 7F47 | 罇 | ≡ | r-trad | | identity |
| 7F47 | 罇 | 8E72 | 蹲 | ↔ | blocked | | visual-similarity |
| 8E72 | 蹲 | 8E72 | 蹲 | ≡ | r-both | | identity |

In addition, as suggested by IP, the variant set {U+5B0E, U+5B14} represents another pair of visually confusable characters and has been defined as follows (assuming U+5B14 is more common than U+5B0E).

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 5B0E | 嬎 | 5B0E | 嬎 | ≡ | r-neither | | identity |
| 5B0E | 嬎 | 5B14 | 嬔 | →<br>← | both<br>blocked | | visual-similarity |
| 5B14 | 嬔 | 5B14 | 嬔 | ≡ | r-both | | identity |

In practice, besides setting the type value of the above variants as "both" or "blocked", CGP has set the comment value as "visual-similarity" to indicate the visual identical relationship, as well as to guide the other registries who refer to it on SLD label generation rules in the future.

**For example:**

```
<char cp="8D7F" tag="sc:Hani" ref="0 100 200">
        <var cp="8D7F" type="r-both" comment="identity" />
</char>
<char cp="8D86" tag="sc:Hani" ref="0 100 200">
        <var cp="8D86" type="r-both" comment="identity" />
</char>
```

transferred to >>

```
<char cp="8D7F" tag="sc:Hani" ref="0 100 200">
        <var cp="8D7F" type="r-neither" comment="identity" />
        <var cp="8D86" type="both" comment="visual-similarity">
</char>
<char cp="8D86" tag="sc:Hani" ref="0 100 200">
        <var cp="8D7F" type="blocked" comment="visual-similarity"/>
        <var cp="8D86" type="r-both" comment="identity"/>
</char>
```

# 8 Assigning Dispositions to Variant Labels

## 8.1 Delegating all simplified, all traditional and original applied-for 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 other variant labels**. To remain consistent with this rule, when the CGP generates its own XML table of CGP repertoire and variant mappings according to the XML-format specifications in RFC 7940 it marks every variant mapping with one of the following types (or its subtypes as described in 8.2 and 8.3):

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

These variant types are then used in determining a disposition for each variant label, based on which variant mappings were used to derive it. The evaluation is performed using "action" elements.
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>
```

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

To limit the number of allocatable labels to at most five(5), CGP created 4 new sub-types for variants in sets that have multiple allocatable variant mappings, and marks corresponding variant mappings with one the following types (see Section 6.3.2):
"simp-1", "simp-2", "trad-1", "trad-2".

Using these new subtypes, 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 simp-1" comment="all simplified label type 1 " />
<action disp="allocatable" only-variants="simp r-simp both r-both simp-2" comment="all simplified label type 2 " />

<action disp="allocatable" only-variants="trad r-trad both r-both trad-1" comment="all traditional label type 1 "/>
<action disp="allocatable" only-variants="trad r-trad both r-both trad-2" comment="all traditional label type 2 "/>

<action disp="blocked" any-variant="simp trad both simp-1 simp-2 trad-1 trad-2" comment="block any other mixed labels" />
<action disp="allocatable" all-variants="r-neither r-trad r-simp r-both" comment="original label" />
<action disp="valid" comment="catch all (default action)"/>
</rules>
```

In other words, a variant label can be of simp-1 or simp-2 type, but cannot contain a mix of simp-1 and simp-2. Likewise for trad-1 and trad-2.

## 8.3 Blocking or allocating visual similar labels

To block or allocate the labels containing visual identical characters, CGP created NO new rules for variants within visual similarity, but only introduced new comment value for variants. (see Chapter 7)

# 9 Contributors

List of CGP expert team

| Name | Organization | Country/Region | Language Expertise |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Chao QI | Teleinfo | 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 | | Malaysia | Chinese |
| Jean-Jacques Subrenat | ATLAC ICANN | France | French, English, Chinese, Japanese. |
| Jenifer CHUNG | Dot Asia | USA/Hongkong | Chinese |
| Jiagui XIE | Teleinfo | China | Chinese |
| Jonathan SHEA | TraxCom | 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 | Institute of Computing Technology, CAS | China | Chinese |
| Yuxiao LI | Chinese Academy of Cyberspace | 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, Pitinan KOOARMORNPATANA, Jianchuan ZHANG |

# 10 References

[1] Coulmas, Florian (1991). The writing systems of the world. Blackwell. ISBN 978-0-631-18028-9.

[2] http://www.cdnc.org/gb/research/file/CDNC_unicode.txt

[3] http://www.iana.org/domains/idn-tables/tables/cn_zh-cn_4.0.html

[4] http://www.iana.org/domains/idn-tables/tables/tw_zh-tw_4.0.1.html

[5] https://www.iana.org/domains/root/db/asia.html

[6] [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

[7] [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

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

[9] [CDNC-2018] CDNC IDN Table http://www.cdnc.asia/file/unicode-1-2.txt

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

[11] [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

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

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

[14] [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/

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

[16] [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>

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

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

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

[20] [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

[21] [OutOfRepertoireVariants] Out of Repertoire Variants in Root-Zone LGR and Proposals, https://www.icann.org/en/system/files/files/root-zone-lgr-repertoire-variants-25sep17-en.pdf

# 11 Appendix

**Appendix A: CGP Repertoire**
The EXCEL document includes 19685 CGP Unicode code points and their source information.

**Appendix B: JGP Repertoire**
The EXCEL document includes 6533 JGP Unicode code points, 6356 of which are Han/Kanji characters.

**Appendix C: KGP Repertoire**
The EXCEL document includes KGP 4758 Han/Hanja Characters and their Unicode code points.

**Appendix D: CGP Variant Mappings Table**
The EXCEL document includes 19785 CGP characters and their variant mapping entries

**Appendix E: CGP Internal Review**
The EXCEL document includes the CGP internal review on 172 IICORE characters and 7 dotAsia characters.

**Appendix F: Appendix F IP External Review**
IP proposed to change the mappings of 9 variant sets in May 2019, CGP accepted 7 of them and refused 2.

**Appendix G: CDNC IDN Table 2018**
The document of the latest CDNC IDN Table which has been adapted by CN and TW, and has been submit to IANA. In both XML and TXT format.

**Appendix H: KGP Hanja Variant Mappings**
The EXCEL document containing Han/Hanja variant mappings in KLGR proposal

**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 L: CDNC IDN Table 2005-2012**
The first version of the CDNC IDN Table generated in 2005 and used until 2012. In both XML and TXT format.

**Appendix M: dotAsia IDN Table**
The XML format document of dotAsia IDN Table 2015. Also in TXT format.

**Appendix N: CGP Internal Review**
The XML format document of 105 variant mappings generated by CGP internal review. Also in TXT format.

**Appendix O: out-of-repertoire variants**
The EXCEL document of 144 out-of-repertoire variants imported from JGP and 14 Hanja characters not included in CLGR.

**Appendix P: KGP Variant Sets**
The XML document of Hanja variant sets proposed by KGP in March 2017.

**Appendix R: Examples of Reducing Multiple Allocatable labels**
Two examples to illustrate how to reduce the number of allocatable labels by adopting the new types created in Section 6.3.