# Considerations concerning the Chinese Root LGR

**Last updated: October 25, 2017**

**This document contains both an analysis of the most recent CLGR proposal from the perspective of the Integration Panel as well as a set of recommendations for changes, primarily in terms of documentation. The recommendations are extracted into labeled summaries and the preceding text can be read as background.**

## 1 Summary

This document analyzes the content of the current Chinese LGR as specified by the Chinese Generation Panel in terms of repertoire and variant sets. The latest draft of that Chinese LGR (CLGR9) is represented by the following files:

- CGP Proposal Draft 20170720.docx, [Proposal]
- Appendix I CGP LGR 20170720_AF1.xml, [CLGR9]
- Appendix I CGP LGR 20170720_AF1.html, (converted xml)

This document may also reference the previous draft of the Chinese LGR (CLGR8) which was not formally reviewed by the Integration Panel. That draft of that Chinese LGR (CLGR8) was represented by the following files:

- CGP Proposal Draft 20170118.docx, [PreviousProposal]
- Appendix I CGP LGR 20170118_AF1.xml, [CLGR8]
- Appendix I CGP LGR 20170118_AF1.html, (converted xml)

Note: The xml files have been modified from the file provided by CGP to make it syntactically correct but without changing the repertoire, references or the variant set definition.

In evaluating this proposal, this document compares it to the dotAsia ZH set (see 2.4) which, like the CLGR9, attempts to cover both simplified and traditional Chinese labels.

While previous versions of this document analyzed in detail the content of the repertoire and its size, there is no need at this point to repeat such analysis; the repertoire has matured and is now stable. However, there a concern remains that the repertoire contains some characters that might not be needed in the Chinese context, but are only included to complete variant sets. This concerns only a small subset of the repertoire (about 60 characters).

To facilitate the reading of these considerations, the detailed analysis of the variant set has been removed and will be covered in a separate document. A small subset of that detailed analysis is provided in Appendix A and B of this document.

The following list summarize the analysis:

- The repertoire is now made of 19,746 code points, very close to the size of MSR-2 Hanzi set (19,850). It is now a full superset of the dotAsia set, itself made of 19,683 code points.
- Of the 63 characters added to CLGR9 beyond dotAsia, 60 seem to be added solely to complete variant sets but lack a justification other than that, as they seem not to be used in the Chinese context. In that case, they should rather be defined as "out-of-repertoire" variants.
- The repertoire includes 2 characters not part of MSR-2, which will require a revision of MSR to include them. A formal request should be sent to IP when the GP is ready to do so.
- As identified in a previous version of this feedback, there is no explanation provided for the differences in variant sets between this LGR and, for example, second level LGRs. Most of these differences are still present and therefore should be addressed.
- In order to establish that there is a rational and linguistic basis for assigning variants, any deviations from existing practice should be justified. If it is felt necessary or useful, some face-to-face or virtual meetings may be setup between IP and CGP to make progress on this issue.
- While the origin of some variants can be traced to the .asia, .cn, and, .tw sets, no references or source information are provided for the modified or new variant sets. (The source for all variants, whether retained from second level or modified, should be unambiguously documented– perhaps not on a per-variant level, but globally, with any exceptions prominently marked).
- Because this set must be integrated with the rest of the CJK sets (i.e. the Japanese and Korean sets), it is important to get a version of this variant set which is agreeable to all concerned parties as soon as possible. The IP notes the progress that has been made in aligning the variant sets between CLGR and KLGR; however, the documentation of the result of this process could be improved.
- Some of the repertoire additions are ideographs that appear to be specific to Japanese or Korean. Their introduction to the CLGR9 necessarily creates additional variant mappings that will then affect the LGRs for Japanese and Korean. Unless an independent justification for their use in the Chinese context can be found they should be added as 'out-of-repertoire' variants instead.
- Finally creating too many 'trad' variants in a given variant set will overproduce allocatable labels, which raises a concern. Reducing or eliminating these multiple 'trad' variants should be explored.

**Outstanding Integration Panel recommendations:**

The following recommendations that were made in a previous version of these considerations do not seem to be addressed in CLGR9:

- Provide documentation for the origin of the proposed variant mappings in CLGR9, particularly where they differ from established second level practice.
- Review variant sets that differ from second level practice and provide a rationale for any differences.
- Review variant sets with multiple “trad” mappings to see if any of those could be changed to “blocked” to reduce the overproduction of allocatable variants.
- Document the specific requirements behind any decision to retain multiple “trad” variants.
- Provide a detailed rationale for inclusion of J-specific or K-specific code points in this C-specific LGR. Please address the implications for variant sets deriving from these additions.
- If J-specific and K-specific code points are not included as full members of this repertoire, they should be included as out-of-repertoire variants and variant sets created accordingly (see main text for details).
- Provide references to all variant mappings using available sources such as Unihan, dotAsia, and any other relevant sources, using the ‘ref’ attribute on the ‘var’ element.
- When presenting special cases and deviations for variant sets in the LGR proposal document, consider presenting these variant sets in term of sets, not as separate code-point-based entries, to ensure that the sets are fully transitive and reflexive. Note that the XML LGR file is the reference for the full definition of these variant sets

**New Integration Panel recommendations:**

The following recommendations are new in this version:

- Revise future LGR Proposal drafts so that they follow the LGR Proposal Template published by ICANN.
- Provide justification for the inclusion of 60 out of the 63 characters added on top of dotAsia; alternatively, change them to “out-of-repertoire” variants

# 2 Definitions

## 2.1 CLGR9

The term (CLGR9) represents the Proposed Chinese root LGR under review here, both in terms of repertoire and variant sets defined in the XML file. The terms CLGR7 and CLGR8 may be used to represent earlier versions of the Chinese LGR.

## 2.2 IICORE collection

The International Ideographs Core (IICORE) is a fixed collection of CJK Ideographic code points deemed essential to all IRG Asian constituencies except Vietnam (a total of 7 sources). It contains 9 810 code points and is part of both ISO/IEC 10646 and Unicode. It was created by IRG based on priority (A to C, A being the highest) among its 7 sources.

## 2.3 MSR-2 CJK repertoire

The CJK repertoire in MSR-2 consists of 19 850 CJK Unified Ideographs, corresponding to the union of the following sub-repertoires:

1) dotAsia Japanese https://www.iana.org/domains/idn-tables/tables/asia_ja_1.1.txt
2) dotAsia Chinese https://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt
3) IICORE as defined in Unicode 6.3
4) Code point U+9DC0.

The dotAsia Chinese repertoire is itself a union of repertoires from various Chinese sources such as China PRC, Hong Kong SARs, and Taiwan.

Note that MSR-2 also contains a few code points that have the 'Han' extended script property but are not considered CJK Ideographs (for example U+3005 IDEOGRAPHIC ITERATION MARK and U+3006 IDEOGRAPHIC CLOSING MARK).

## 2.4 dotAsia LGR

A transcription of the dotAsia (ZH) domain name definition available at https://www.iana.org/domains/idn-tables/tables/asia_zh_1.1.txt into the XML-format is publicly available at https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-chinese-15may16-en.xml. This transcription was created as part of reference for 2nd level domain. It shares many features with the proposed root Chinese LGR. The dotAsia table (or its XML transcription) represents an important set that can be used as a reference point for the comparison in terms of both the repertoire and the variant sets. It contains 19 684 Han ideographs and 3 505 variant sets. In comparison, the current Chinese Root Zone LGR draft (CLGR9) contains 19 746 Han ideographs and 3 477 variant sets.

Technically, the transcription of the dotAsia (ZH) domain contains one more CJK ideograph than dotAsia's original table: U+9DC0 that was added to complete a variant set. The original IDN table contains 19 683 Han ideographs. In following comparisons of CLGR7 and dotAsia repertoire, the original 19 683 code points defined in dotAsia should be used.

The dotAsia repertoire is fully included in CLGR9. CLGR9 contains an additional 63 code points not included in dotAsia.

## 2.5 Unihan

The Unihan database at http://www.unicode.org/charts/unihan.html is a Unicode Standard component containing information related to all CJK Ideographs. That information includes sources, variants, dictionaries, etc. As such it is an extremely useful tool to validate the CLGR9 content.

# 3 Consideration concerning the format of the [Proposal]

It is important that the Chinese LGR proposal follow the guidelines provided in the following link: https://community.icann.org/display/croscomlgrprocedure/Document+Repository?preview=/43989034/56133334/LGR-Proposal-Template.docx

In particular, the following information **must** be provided in the document preamble:

- LGR version,
- Date
- Document version
- Authors

It would facilitate review and integration process if the breakdown into sections were to follow the common scheme (and the missing information were to be supplied for section 2)

**1 General Information** (section 0 in current [Proposal]

**2 Script for which the LGR is proposed** (not declared in current [Proposal])

**3 Background on Script and Principal Languages Using It** (section 1 in current [Proposal])

**4 Overall Development Process and Methodology** (section 2 in current [Proposal])

**5 Repertoire** (partially covered in section 3 of current [Proposal])

**6 Variants** (partially covered in section 4 of current [Proposal]

**7 WLE rules** (section 5 in current [Proposal])

In addition, much of the detail about development process for the repertoire and the variant sets should be moved to annexes. Section 5, 6, and 7 as specified above should be limited to the formal definition of the repertoire, variant sets and WLE rules.

**Integration Panel recommendations:**

Revise future LGR Proposal Drafts to follow the LGR Proposal Template as indicated above.

# 4 Repertoire considerations

The [Proposal] in section 3 describes the development of the repertoire, including detailing the various iterations from the initial repertoire to the current repertoire. The current repertoire can also be simply derived from the dotAsia repertoire by adding the following list of 63 characters:

- 2 HKSCS characters that were left out when processing HKIRC request
- 18 characters from the Table of General Standard Chinese Characters (TGSCC formerly known as Normalized Hanzi List for Common Use (NHCU) in previous drafts of CLGR)
- 43 IICORE characters from J source and K sources part of Chinese based variant sets. These are also part of draft versions of the JGP and KGP repertoires.

These additions are also described in sections 3.2.1, 3.2.3, and 3.2.4 respectively in the [Proposal].

The 2 HKSCS characters (U+3A5C  and U+58B5 墜), although not part of MSR-2, are acceptable in principle. However, a revision of MSR will be required to include these 2 characters and to make them available to the Root Zone process. Both code points are members of CLGR9 variant sets.

The 18 characters from TGSCC can be described as follows (class values from section 3.2.3 of the [Proposal], IRG Sources (GE sources excluded), IICORE value, and Unihan variants from Unihan, CLGR9 variants from [CLGR9]). 17 code points (out of 18) are members of CLGR9 variant sets.

| No | UCS | Glyph | Class | IRG Sources | IICORE | Unihan variants | CLGR9 variants |
|---|---|---|---|---|---|---|---|
| 1 | 48BC | 邽 | N | G3,T4 | | | |
| 2 | 732F | 猫 | N | G5,T3,J0,K2 | | | 8C92 |
| 3 | 9EB9 | 麭 | N | J0 | AJ | | 9EB4 |
| 4 | 5227 | 刧 | V | T3,J0,K1,H | | 5226, 52AB | 5226, 523C, 52AB |
| 5 | 524F | 剋 | V | T3,J0,K2,H | | 5231 | 521B, 5231, 5259, 5275 |
| 6 | 6060 | 恮 | V | T3,J0,K1 | | 602A | 602A |
| 7 | 74A2 | 瑂 | V | T3,J0,K2 | | 7409 | 7409. 7460 |
| 8 | 750E | 甄 | V | T3,J0,K1,H | | 78DA | 7816, 78DA |
| 9 | 754A | 畊 | V | T3,J0,K1,H | | 8015 | 8015 |
| 10 | 7ADA | 竽 | V | T3,J0,K2 | | 4F47 | 4F2B, 4F47 |
| 11 | 8262 | 艦 | V | T3,J0,K2,HB2 | | 6AA3 | 6A2F, 6AA3 |
| 12 | 88B5 | 衹 | V | T3,J0,K2,H | | 887D | 887D |
| 13 | 894D | 褋 | V | T3,J0,K1,H | | 96DC | 6742, 96D1, 96DC, 96E5 |
| 14 | 8B0C | 謌 | V | T3,J0,K1,H | | 6B4C | 6B4C |
| 15 | 8F19 | 輙 | V | T3,J0,K2,H | | 8F12 | 8F12, 8F84 |
| 16 | 945A | 鑽 | V | T3,J0,K2 | CJ | 947D | 9246, 9409, 947D, 94BB |
| 17 | 984B | 顋 | V | T3,J0.K1,H | | 816E | 816E |
| 18 | 9DC0 | 鷁 | V | G1,T4,J3,K2,H | | | 9DBF, 9E5A |

Findings concerning TGSCC additions:

- 16 characters have J0 sources (core Japanese Kanji set), among these 16 J0 sources 14 have the variant class (V) according to section 3.2.18 of the [Proposal]. Another character (code point U+9DC0) is part of a variant set (9DBF 鷲, 9DC0 鷁, 9E5A 鷗), although it is not recorded in Unihan (Unihan only recognizes a variant relationship between U+9DBF and U+9E5A).
- All Unihan variants are included in CLGR9 but the latter add many members.
- 15 characters have the variant class (V) according to section 3.2.18 of the [Proposal]. Therefore, they may have been only incorporated into TGSCC for that purpose.
- One character (code point U+9EB9), has only the J0 source, has the normalized class (N) according to section 3.2.18 of the [Proposal], and is not in a variant set according to Unihan.

- That leaves only one character (code point U+48BC) with solid evidence for inclusion. The other characters may add unnecessary variant set members, especially because so many are J0 source characters which are by definition part of the Japanese root LGR.

These last 43 characters are IICORE characters from J source and K sources part of Chinese based variant sets. They are not used in Chinese, as far as the IP understands. The 43 characters can be described as follows (class values from section 3.2.4 of the [Proposal], IRG Sources (GE sources excluded), IICORE value, and Unihan variants from Unihan, CLGR9 variants from [CLGR9]). By definition, all code points are members of CLGR9 variant sets.

| No | UCS | Glyph | IRG Sources | IICORE | Unihan variant | CLGR9 variants |
|---|---|---|---|---|---|---|
| 1 | 3960 | 惰 | G5,T3,JA,K3 | CK | | 8ADD, 8C1E |
| 2 | 4FAD | 侭 | J0 | AJ | 5118 | 5118, 5C3D, 76E1 |
| 3 | 51E6 | 処 | J0,K2 | AJ | 458F, 8655 | 5904, 8655 |
| 4 | 56A2 | 嚢 | J0 | AJ | 56CA | 56CA |
| 5 | 61F4 | 懴 | J0,K2 | CJ | 61FA | 5FCF, 61FA |
| 6 | 6442 | 摂 | J0 | AJ | 651D | 6315, 6444, 651D |
| 7 | 663B | 昴 | T3,K0 | AKP | 6602 | 6602 |
| 8 | 685C | 桜 | J0,K2 | AJ | 6AFB | 6A31, 6AFB |
| 9 | 685F | 桟 | J0 | AJ | 68E7 | 6808, 68E7, 8F4F |
| 10 | 6D9C | 涜 | J0,K2 | AJ | 7006 | 6E0E, 7006 |
| 11 | 6E8C | 溌 | J0,K2 | AJ | 6F51 | 6CFC, 6F51 |
| 12 | 731F | 猟 | J0,K2 | AJ | 7375 | 730E, 7375 |
| 13 | 784F | 研 | T3,J3,K0 | AKP | | 63C5, 7814 |
| 14 | 7A36 | 穢 | T4,K0 | AKP | | 7A22 |
| 15 | 7B86 | 箆 | T3,J0,K2 | AJ | | 7BE6 |
| 16 | 7C14 | 簔 | TF,J0 | CJ | 7C11 | 7C11, 84D1 |
| 17 | 7D9A | 続 | J0,K2 | AJ | 7E8C | 7E8C, 7EED |
| 18 | 7E4A | 纎 | J0,K2 | ATJ | 7E96 | 5B45, 7E34, 7E8E, 7E96, 7EA4 |
| 19 | 7E4B | 繁 | J0,K2 | AJ | 7E6B | 7E6B |
| 20 | 8133 | 脳 | J0 | AJ | 8166 | 8111, 8166 |
| 21 | 81D3 | 臓 | J0 | AJ | 81DF | 810F, 81DF, 9AD2 |
| 22 | 8217 | 舗 | J0 | AJ | 92EA | 8216, 92EA, 94FA |
| 23 | 839F | 莟 | G3,T3,J0,K1 | CJ | | 83E1 |
| 24 | 83B5 | 菟 | J0 | CJ | 83DF | 83DF |
| 25 | 86CD | 蛍 | J0 | AJ | 87A2 | 8424, 87A2 |
| 26 | 8E99 | 躙 | G5,T3,J0,K1 | CJ | | 8E8F, 8EAA |
| 27 | 9039 | 達 | J0 | CJ | 9054 | 8FBE, 8FD6, 9054 |
| 28 | 91A4 | 醤 | J0 | AJ | 91AC | 9171, 91AC |
| 29 | 91C8 | 釈 | J0,K2 | AJ | 91CB | 91CA, 91CB |
| 30 | 9271 | 鉱 | J0,K2 | AJ | 7926 | 77FF, 7926, 945B |

| No | UCS | Glyph | IRG Sources | IICORE | Unihan variant | CLGR9 variants |
|---|---|---|---|---|---|---|
| 31 | 9421 | 鐵 | J0,K2 | CJ | 9435 | 9244, 9295, 9435, 94C1 |
| 32 | 967A | 險 | J0,K2 | AJ | 96AA | 7877, 78B1, 7906, 9669, 96AA, 9E7C |
| 33 | 96B2 | 隲 | T3,J0 | CJ | 9A2D | 9A2D, 9A98 |
| 34 | 982C | 頰 | J0 | AJ | 9830 | 9830, 988A |
| 35 | 98EE | 飲 | J0,K0 | AKP | 98F2 | 98F2, 996E |
| 36 | 9A12 | 騷 | J0 | AJ | 9A37 | 9A37, 9A9A |
| 37 | 9A13 | 驗 | J0,K2 | AJ | 9A57 | 9A57, 9A8C |
| 38 | 9A28 | 驊 | J0 | AJ | 9A52 | 9A52 |
| 39 | 9C2E | 鱷 | T3,J0,K1 | CJ | 9CC1 | 9C1B, 9CC1 |
| 40 | 9D0E | 鴎 | J0 | AJ | 9DD7 | 9DD7, 9E25 |
| 41 | 9D2C | 鴦 | J0 | AJ | 9DAF | 83BA, 9DAF |
| 42 | 9D8F | 鶏 | J0 | AJ | 96DE | 96DE, 9CEE, 9DC4, 9E21 |
| 43 | 9EBA | 麺 | J0 | AJ | 9EB5 | 9762, 9EAA, 9EB5 |

Findings concerning IICORE additions:

- As the [Proposal] states, out of the 99 IICORE characters in the MSR but not covered in previous versions of CGP, the CGP added "*43 characters included in the JGP repertoire (version 201703. Appendix C) and KGP repertoire (version 201703, Appendix D) with variant relationships with CGP R2*".
- All Unihan variants but one (U+458F, a member of the variant set associated with U+51E6) are included in CLGR9 but the latter add many members.

Normally, the inclusion of J-specific or K-specific code points in a Chinese LGR would appear to serve no purpose. From a repertoire perspective, it would only make sense if there was a requirement to apply for labels that combine these code points with some Chinese-only code points. Absent such a requirement, it is doubtful that the inclusion of these code points can be justified on repertoire considerations only. The Conservatism Principle demands that the repertoire selection be conservative – only the necessary code points should be included.

However, where these code points have variant relations with other code points that are in the CLGR repertoire, the issue becomes more interesting. Even if, under conservative design, a code point is only present in the Japanese LGR, for example, it might be possible to apply for a label that is seen, by Chinese users, as a variant of some other Chinese label. This cross-repertoire variant relation is similar to the cross-script variant issue in alphabetic scripts. In both cases, to allow for blocking the variant label, it is required to add the out-of-repertoire code point to the repertoire.

However, because notionally these code points are not part of the required set for labels, they should be given a reflexive variant mapping of type "out-of-repertoire-var" and variant mappings of type "blocked" to (and from) all code points that are variants of it in the repertoire.

The Default actions that are defined in MSR-2 and that form part of every Root Zone LGR create a disposition of “invalid” for any label containing a code point with reflexive variant of type “out-of-repertoire-var”. In this manner, the full variant sets can be specified for cross-repertoire variants without having to allow these code points to be usable for labels (as required by the Conservatism principle).

Finally, the table of the 43 characters from JGP and KGP in section 3.2.4 of the [Proposal] has many errors in its IICORE content. It should be corrected to align with the same IICORE content exposed in the table shown in section 4.2.2 of the same document.

**Integration Panel recommendations:**

Fix the table content in section 3.2.4. Justify why 60 out of the 63 additions made over dotAsia are required in Chinese context if that is the case. If not, and if only required to complete variant sets, add them as out-of-repertoire variants. This means that their reflexive mapping should be ‘out-of-repertoire-var’ and mapping from and to these characters should be ‘blocked’. The code points corresponding to the 3 acceptable characters are U+3A5C, U+58BC, and U+48BC. Adding the first two characters will require a revision of MSR-2.

# 5 Variant considerations

## 5.1 General

There are three groups of variant set issues:

1) CLGR9 variant set differs from dotcn/dottw definition (and DotAsia common subset). In previous versions of CLGR, this used to be a small set (4 members). It is now expected to be much bigger following the discussions with KGP. The IP has still not reviewed this group.

2) CLGR9 variant set differs from DotAsia for code points not part of original dotcn/dottw. In previous versions of CLGR, this was a set of 47 elements. DotAsia may not be a perfect reference and may be superseded. However, in the absence of existing second level implementations, the IP require some justification for the choices made. This is of particular concern in cases where other sources (such as Unihan) would have supported the variant mappings as found in DotAsia and not CLGR.

Section 4.2.3 of the [Proposal] ‘69 dotAsia unique variants review’ dotAsia characters attempts to explain these differences. The analysis found in that section has been part of previous CLGR drafts (at least CLGR7 and [CLGR8]) and has been reviewed as being erroneous by the Integration Panel in previous feedback. However, it remains unchanged. The IP feels that the following points still need to be addressed:

- Only 7 characters out of the 69 characters are listed,
- Even among those seven characters, one is missing from the enumeration: 陉 (U+49CD),
- In the table, the left part does not actually correspond to dotAsia variant sets, contrary to what is implied,

- In the same table, the right part does not correspond to the current CLGR9 variant set (it may correspond to a previous version of CLGR).
- The last paragraph states:

  *The CGP and Edmon CHUNG discussed the issue of inconsistency between the CGP 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.*

  The issue with the statement is that now there is now a reference 2nd level LGR table for Chinese based on the dotAsia set, available at https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-chinese-15may16-en.xml

  In the [previousProposal] there was a section 4.5 'Variant inconsistency between .asia [viz dotAsia] and CGP', which contained a list of 69 BMP characters with different variant mappings between dotAsia and CLGR. It contained the following terms: "*The inconsistency reflects the regional cultural individuality application difference on SLD registration in the early days*."

  The statements made in the [Proposal] and [previousProposal] raise some concerns because they imply a lack of stability for the variant set that could be very damaging for deployed domain names. In addition, while dotAsia may revise its practice, there is now a separate reference which is not bound by the same principle.

  In general, the IP would welcome a more detailed explanation of the differences. For example, consider the variant pair [U+3A18, U+64E4] and mapping in CLGR9 (1st) and dotAsia (2nd)

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 3A18 | 揹 | 3A18 | 揹 | ≡ | r-trad | | identity |
| 3A18 | 揹 | 64E4 | 擃 | → | simp | | |
| | | | | ← | blocked | | |
| 64E4 | 擃 | 64E4 | 擃 | ≡ | r-both | | identity |
| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
| 3A18 | 揹 | 3A18 | 揹 | ≡ | r-both | | identity |
| 3A18 | 揹 | 64E4 | 擃 | ↔ | blocked | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 64E4 | 擤 | 64E4 | 擤 | ≡ | r-both | | identity |

Unihan in the entry for U+3A18 mentions in its kDefinition field that it is “(same as U+64E4 擤) to blow the nose with the fingers; (Cant.) to scour; to rebuke; to hit with a ball”; but it has no traditional/simplified mapping. Therefore, while a semantic variant is implied, the traditional/simplified mapping added by CLGR9 is not supported.

(this example with few more are provided in Appendix B)

Overall the whole variant set has to be documented/referenced on external evidences, especially when it differs from previous practice such as dotcn/dottw/dotAsia, and any additions/deviations have to be based upon clear research material like Unihan or similar sources.

On a positive note, the IP notes that the variant mapping for Plane 2 characters part of dotAsia is now endorsed by CGP.

3) CLGR9 variant set differs because of characters addition beyond DotAsia. This was an identified set of 60 elements in earlier versions of CLGR. In CLGR9, the count is updated to 62 (all but one of the 63 additions have variant relationship in CLGR9). Based on the repertoire review it appears that all these 62 code points may be only needed to complete variant sets but are not used in Chinese context. This means that their reflexive mapping should be ‘out-of-repertoire-var’ and mapping from and to these characters should be ‘blocked’.

**Integration Panel recommendations:**

The variant set has to be documented/referenced on external evidences, especially when it differs from previous practice such as dotcn/dottw/dotAsia, and any additions/deviations have to be based upon clear research material like Unihan or similar sources.

The inclusion of J-specific or K-specific code points was done to handle such out-of-repertoire variant issues. Therefore, their mapping types should be updated so that they match the expected types for an out-of-repertoire code point as described above. If, instead, the GP desires them to be included as full members, the IP expects a documented justification for their inclusion as full members of the repertoire, based on their demonstrated use in Chinese establishing a requirement for support in IDNs.

Finally, the section 4.2.3 of the [Proposal] needs to be rewritten to describe the rationale for these differences between dotAsia and CLGR.

## 5.2 Consideration on multiple variant character mappings and multiple allocatable variants

Because of some earlier feedback from the Integration Panel on that topic, the section 4.3 of the [Proposal] goes into great length in exploring in how to reduce allocatable variants, by using multiple

variant sub-type (ending by '-m'), which are not implemented in the repertoire part of the XML file [CLGR9].

It also describes a multiple LGR execution or an extra process to request 'reserved label activation', concluding by the following text:

> *In CGP LGR 201703 (Appendix J), the CGP decided to introduce the 6 new subtypes ("simp-m", "trad-m", "r-simp-m", "r-trad-m", "r-both-m" and "both-m") and, correspondingly, new rules for these sub-types into the LGR. But the new sub-types will not be actually tagged to any chars before either of the above two solutions was accepted by the IP.*

The fundamental issue with that approach is that any process concerning multiple LGR execution and/or extra process concerning activation of reserved label is out of scope for the Integration Panel. Therefore the IP cannot either accept or decline such approach.

**IP Recommendation**

The discussion in section 4.3 of the current [Proposal] should be removed (not only is it out of scope, but it contradicts the formal specification of the LGR via the XML document).

The IP continues to have concerns about high multiplicity of allocatable variant labels, but considers that any solution must be found in the scope of the RZ-LGR work. Therefore, the IP requests that the CGP review and provide feedback on the alternative solution presented in section 6 below.

# 6 Variant Mappings that may overproduce allocatable labels

**Note:** The following text is largely unchanged from the previous IP considerations (dated October 28, 2106); the IP plans to provide additional comments and analysis on the variant set. The overall concern is whether there is a sound linguistic or other foundation for the assignment of variants in the current set, and if so, how to document that fact., .

In the Chinese LGR, the variant mappings and WLE rules are designed with the assumption that given any valid input label, there would be at most three resulting allocatable labels -- the original label, an all-simplified label, and an all-traditional label. This is achieved using variant mappings of having at most one instance in each of the following set of types:

a. trad, r-trad, both, r-both
b. simp, r-simp, both, r-both

However, in CLGR7, there are 196 code points (attached file CLGR-Overproducing-Variants-20160530.txt) with variant types that violate the above constraint. This would lead to overproduction of variant labels with an "allocatable" status.

An example would be:

```
<char cp="53F0" tag="sc:Hani" ref="0 100 101 102 103 104" >
```

```
    <var cp="53F0" type="r-both" comment="identity,reflexive" />
    <var cp="6AAF" type="trad" />
    <var cp="7C49" type="block" />
    <var cp="81FA" type="trad" />
    <var cp="98B1" type="trad" />
  </char>
```

Using the one-line notation in the attached file, the above is represented as:

53F0[台] trad=> 6AAF[檯] trad=> 81FA[臺] r-both=> 53F0[台] trad=> 98B1[颱]

An input label of 台湾 (53F0 6E7E) would result in 5 allocatable variant labels (action numbers indexed per sequence order of the <action> elements in the XML file: 0 to 5):

- **Variant: (檯灣) (6AAF 7063): [trad] ==> allocatable due to Action[2]**
- Variant: (檯湾) (6AAF 6E7E): [trad r-simp] ==> blocked due to Action[4]
- **Variant: (臺灣) (81FA 7063): [trad] ==> allocatable due to Action[2]**
- Variant: (臺湾) (81FA 6E7E): [trad r-simp] ==> blocked due to Action[4]
- **Variant: (台灣) (53F0 7063): [trad r-both] ==> allocatable due to Action[2]**
- **Variant: (台湾) (53F0 6E7E): [r-both r-simp] ==> allocatable due to Action[1]**
- **Variant: (颱灣) (98B1 7063): [trad] ==> allocatable due to Action[2]**
- Variant: (颱湾) (98B1 6E7E): [trad r-simp] ==> blocked due to Action[4]
- Variant: (黨灣) (7C49 7063): [trad block] ==> blocked due to Action[0]
- Variant: (黨湾) (7C49 6E7E): [r-simp block] ==> blocked due to Action[0]

Some of the above "allocatable" labels are unnecessary from a semantic standpoint.

In at least some of these 196 code points, the variant type assignments appear to be due to a simplified code point having multiple traditional variants. If so, it may be an acceptable trade-off to eliminate the multiple traditional mappings, and let registrants who need a specific traditional variant label apply for the specific traditional label.

If that was the argument, most of these cases (except perhaps for a few cases such as "Taiwan") can be fixed by not having multiple traditional mappings. Registrants who want a specific traditional label should apply for the traditional string, which should give the right simplified string, and won't over-generate.

An example would be a label involving two of the code points that exhibit this issue:

66F2[曲] r-both=> 66F2[曲] trad=> 9EB4[麴]

9709[霉] r-both=> 9709[霉] trad=> 9EF4[黴]

The label 红曲霉 (7EA2 66F2 9709) "red yeast" would yield 5 allocatable labels:

- Variant: (红麴霉) (7EA2 9EAF 9709): [r-both block r-simp] ==> blocked due to Action[0]

- Variant: (红麴黴) (7EA2 9EAF 9EF4): [trad block r-simp] ==> blocked due to Action[0]
- **Variant: (红曲霉) (7EA2 66F2 9709): [r-both r-simp] ==> allocatable due to Action[1]**
- Variant: (红曲黴) (7EA2 66F2 9EF4): [trad r-both r-simp] ==> blocked due to Action[4]
- Variant: (红麹霉) (7EA2 9EB9 9709): [r-both block r-simp] ==> blocked due to Action[0]
- Variant: (红麹黴) (7EA2 9EB9 9EF4): [trad block r-simp] ==> blocked due to Action[0]
- Variant: (红麴霉) (7EA2 9EB4 9709): [trad r-both r-simp] ==> blocked due to Action[4]
- Variant: (红麴黴) (7EA2 9EB4 9EF4): [trad r-simp] ==> blocked due to Action[4]
- Variant: (紅麴霉) (7D05 9EAF 9709): [trad r-both block] ==> blocked due to Action[0]
- Variant: (紅麴黴) (7D05 9EAF 9EF4): [trad block] ==> blocked due to Action[0]
- **Variant: (紅曲霉) (7D05 66F2 9709): [trad r-both] ==> allocatable due to Action[2]**
- **Variant: (紅曲黴) (7D05 66F2 9EF4): [trad r-both] ==> allocatable due to Action[2]**
- Variant: (紅麹霉) (7D05 9EB9 9709): [trad r-both block] ==> blocked due to Action[0]
- Variant: (紅麹黴) (7D05 9EB9 9EF4): [trad block] ==> blocked due to Action[0]
- **Variant: (紅麴霉) (7D05 9EB4 9709): [trad r-both] ==> allocatable due to Action[2]**
- **Variant: (紅麴黴) (7D05 9EB4 9EF4): [trad] ==> allocatable due to Action[2]**

Instead, if the variant types could be amended to the following:

66F2[曲] r-both=> 66F2[曲] blocked=> 9EB4[麴]

9709[霉] r-both=> 9709[霉] blocked=> 9EF4[黴]

The same input label 红曲霉 (7EA2 66F2 9709) "red yeast" would yield 2 allocatable labels (omitting output labels that have been assigned a "blocked" disposition):

- **Variant: (红曲霉) (7EA2 66F2 9709): [r-both r-simp] ==> allocatable due to Action[1]**
- **Variant: (紅曲霉) (7D05 66F2 9709): [trad r-both] ==> allocatable due to Action[2]**

This may not be desirable because 紅麴霉 (7D05 9EB4 9709) is perhaps more appropriate. In that case, it can be the applied-for label, which would then yield the following 2 allocatable labels:

- **Variant: (红曲霉) (7EA2 66F2 9709): [simp r-both] ==> allocatable due to Action[1]**
- **Variant: (紅麴霉) (7D05 9EB4 9709): [r-trad r-both] ==> allocatable due to Action[2]**

Conclusion: Under the conservatism principle, LGRs should strive to minimize allocatable variants. The IP would like to urge the CGP to change the variant types of the affected code points to mitigate this issue, and/or provide strong evidence for the need of including exceptional cases with multiple allocatable variants.

**Appendix, Some cases of variant set differences.**

**A – variant set differences introduced by new repertoire additions**

1. This variant set has one added member U+3960.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 3960 | 慣 | 3960 | 慣 | ≡ | r-neither | | identity |
| 3960 | 慣 | 8ADD | 譋 | → | trad | | |
| | | | | ← | blocked | | |
| 3960 | 慣 | 8C1E | 谖 | → | simp | | |
| | | | | ← | blocked | | |
| 8ADD | 譋 | 8ADD | 譋 | ≡ | r-trad | | identity |
| 8ADD | 譋 | 8C1E | 谖 | → | simp | | |
| | | | | ← | trad | | |
| 8C1E | 谖 | 8C1E | 谖 | ≡ | r-simp | | identity |

The code point U+3960 has G, T, J, and K source and is part of the IICORE set (value CK, meaning low priority, Korean usage).


Unihan kDefinition field indicates that this is a variant of U+8ADD 譋. As such the proposed mappings would be adequate if U+3960 was required for Chinese usage. However, if instead its justification is purely from an integration scenario, its mapping to other code points should be all ‘blocked’ and its own type should be ‘out-of-repertoire-var’.

2. This variant set has one added member U+3A5C (not in dotAsia.) from the HKSCS set. In addition, U+39DB and U+64E5 (both in dotAsia) are also included in the CLGR9 (1st and 2nd) and are mapped differently from dotAsia (3rd). This case is a hybrid of this category (one code point added not in dotAsia) and the next category (two code points already in dotAsia but treated differently). The red highlighting in both tables reflects all differences between the two LGRs.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 39DB | 𪜀 | 39DB | 𪜀 | ≡ | r-both | | identity |
| 39DB | 𪜀 | 64E5 | 黫 | ↔ | blocked | | |
| 64E5 | 黫 | 64E5 | 黫 | ↔ | r-both | | identity |
| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
| 3A5C | 𣭚 | 3A5C | 𣭚 | ≡ | r-both | | identity |
| 3A5C | 𣭚 | 63FD | 揽 | ↔ | blocked | | |
| 3A5C | 𣭚 | 652C | 攬 | ↔ | blocked | | |
| 63FD | 揽 | 63FD | 揽 | ≡ | r-simp | | identity |
| 63FD | 揽 | 652C | 攬 | → | trad | | |
| | | | | ← | simp | | |
| 652C | 攬 | 652C | 攬 | ≡ | r-trad | | identity |
| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
| 39DB | 𪜀 | 39DB | 𪜀 | ≡ | r-simp | | identity |
| 39DB | 𪜀 | 63FD | 揽 | ↔ | blocked | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 39DB | 㚔 | 64E5 | 擏 | → | trad | | |
| | | | | ← | simp | | |
| 39DB | 㚔 | 652C | 攬 | ↔ | blocked | | |
| 63FD | 揽 | 63FD | 揽 | ≡ | r-simp | | identity |
| 63FD | 揽 | 64E5 | 擏 | ↔ | blocked | | |
| 63FD | 揽 | 652C | 攬 | → | trad | | |
| | | | | ← | simp | | |
| 64E5 | 擏 | 64E5 | 擏 | ≡ | r-trad | | identity |
| 64E5 | 擏 | 652C | 攬 | ↔ | blocked | | |
| 652C | 攬 | 652C | 攬 | ≡ | r-trad | | identity |

The code point U+3A5C has G, T, H, J, and V (Vietnam) source.


Unihan kDefinition field indicates that this is a variant of U+652C 攬. It also has semantic variants association with U+64E5 擏, and U+3A2B 揽(not in CLGR9) is listed as a simplified variant. While the mapping for U+3A5C is acceptable (and correspond to an earlier feedback from IP), there is no justification for changing the mapping for the pair (U+64E5, U+39DB). In fact, the table in section 4.2.3 supports the dotAsia mapping for these 2 characters.

3. This variant set has one added member U+7ADA.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 4F2B | 仒 | 4F2B | 仒 | ≡ | r-simp | | identity |
| 4F2B | 仒 | 4F47 | 伫 | → | trad | | |
| | | | | ← | simp | | |
| 4F2B | 仒 | 7ADA | 竚 | → | blocked | | |
| | | | | ← | simp | | |
| 4F47 | 伫 | 4F47 | 伫 | ≡ | r-trad | | identity |
| 4F47 | 伫 | 7ADA | 竚 | → | blocked | | |
| | | | | ← | trad | | |
| 7ADA | 竚 | 7ADA | 竚 | ≡ | r-neither | | identity |

The code point U+7ADA has G, H, T, J, and K sources and is part of the Normalized Hanzi list for Common Use.


Unihan kSemanticVariant field indicates that this is a variant of U+4F47 伫. As such the proposed mappings would be adequate. However, if instead the justification for adding U+7ADA is purely from an integration scenario, its mapping to other code points should be all ‘blocked’ and its own type should be ‘out-of-repertoire-var’.

**B- cases where there are differences between CLGR9 and doAsia**

1. The code point U+3A18 was included in CLGR9 because of its membership in IICORE but is has been assigned different types for its variant mappings between CLGR9 (1st) and dotAsia (2nd)

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 3A18 | 揹 | 3A18 | 揹 | ≡ | r-trad | | identity |
| 3A18 | 揹 | 64E4 | 擓 | → | simp | | |
| | | | | ← | blocked | | |
| 64E4 | 擓 | 64E4 | 擓 | ≡ | r-both | | identity |
| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
| 3A18 | 揹 | 3A18 | 揹 | ≡ | r-both | | identity |
| 3A18 | 揹 | 64E4 | 擓 | ↔ | blocked | | |
| 64E4 | 擓 | 64E4 | 擓 | ≡ | r-both | | identity |

Unihan in the entry for U+3A18 mentions in its kDefinition field that it is “(same as U+64E4 擓) to blow the nose with the fingers; (Cant.) to scour; to rebuke; to hit with a ball”; but it has no traditional/simplified mapping. Therefore, while a semantic variant is implied, the traditional/simplified mapping added by CLGR9 is not supported.

2. The code point U+3A52 was included in CLGR9 because of its membership in IICORE but is treated differently between CLGR9 (table follows) and dotAsia (where it is a singleton reflexive variant of ‘r-both’)

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 3A52 | 㯳 | 3A52 | 㯳 | ≡ | r-trad | | identity |
| 3A52 | 㯳 | 64D2 | 㰢 | → | simp | | |
| | | | | ← | blocked | | |
| 64D2 | 㰢 | 64D2 | 㰢 | ≡ | r-both | | identity |

The simplified mapping between U+3A52 and U+64D2 in CLGR9 is not supported by Unihan and looks doubtful. Unihan kDefinition field for U+3A52 indicates that this is a variant of U+64D2 but without simplified mapping. In Unihan U+64D2 has itself a semantic variant relationship with U+6366 捦, not supported by either CLGR9 or dotAsia.

3. In dotAsia (table follows), U+4E11 and U+919C have a variant relationship. In CLGR9, both are singleton reflexive variant of type ‘r-both’.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 4E11 | 丱 | 4E11 | 丱 | ≡ | r-both | | identity |
| 4E11 | 丱 | 919C | 醜 | → | trad | | |
| | | | | ← | simp | | |
| 919C | 醜 | 919C | 醜 | ≡ | r-trad | | identity |

Unihan supports the dotAsia mapping, both code points are part of KLGR.

4. In dotAsia (table), U+4E18, U+4E20, and U+5775 have a variant relationship. In CLGR9, only U+4E18 and U+4E20 are variants, U+5775 is a singleton reflexive variant of type ‘r-both’.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 4E18 | 丘 | 4E18 | 丘 | ≡ | r-both | | identity |
| 4E18 | 丘 | 4E20 | 北 | → | blocked | | |
| | | | | ← | both | | |
| 4E18 | 丘 | 5775 | 坵 | → | blocked | | |
| | | | | ← | simp | | |
| 4E20 | 北 | 4E20 | 北 | ≡ | r-neither | | identity |
| 4E20 | 北 | 5775 | 坵 | ↔ | blocked | | |
| 5775 | 坵 | 5775 | 坵 | ≡ | r-trad | | identity |

Unihan supports the dotAsia mapping, both U+4E18 and U+5775 are part of KLGR.

5. In dotAsia (table), U+4E26, U+4F75, U+5002, U+5E76, U+5E77, and U+7ADD have a variant relationship. In CLGR9, only U+4E26, U+4F75, U+5002, U+5E76 and U+5E77 are variants, U+7ADD is a singleton reflexive variant of type ‘r-both’, other mappings are unchanged.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 4E26 | 並 | 4E26 | 並 | ≡ | r-trad | | identity |
| 4E26 | 並 | 4F75 | 併 | ↔ | blocked | | |
| 4E26 | 並 | 5002 | 倂 | ↔ | blocked | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4E26 | 並 | 5E76 | 并 | → | simp | | |
| | | | | ← | trad | | |
| 4E26 | 並 | 5E77 | 幷 | ↔ | blocked | | |
| 4E26 | 並 | 7ADD | 竝 | → | blocked | | |
| | | | | ← | trad | | |
| 4F75 | 併 | 4F75 | 併 | ≡ | r-trad | | identity |
| 4F75 | 併 | 5002 | 倂 | → | blocked | | |
| | | | | ← | trad | | |
| 4F75 | 併 | 5E76 | 并 | → | simp | | |
| | | | | ← | trad | | |
| 4F75 | 併 | 5E77 | 幷 | ↔ | blocked | | |
| 4F75 | 併 | 7ADD | 竝 | ↔ | blocked | | |
| 5002 | 倂 | 5002 | 倂 | ≡ | r-neither | | identity |
| 5002 | 倂 | 5E76 | 并 | → | simp | | |
| | | | | ← | blocked | | |
| 5002 | 倂 | 5E77 | 幷 | ↔ | blocked | | |
| 5002 | 倂 | 7ADD | 竝 | ↔ | blocked | | |
| 5E76 | 并 | 5E76 | 并 | ≡ | r-both | | identity |
| 5E76 | 并 | 5E77 | 幷 | → | blocked | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | ← | both | | |
| 5E76 | 并 | 7ADD | 竝 | → | blocked | | |
| | | | | ← | simp | | |
| 5E77 | 幷 | 5E77 | 幷 | ≡ | r-neither | | identity |
| 5E77 | 幷 | 7ADD | 竝 | ↔ | blocked | | |
| 7ADD | 竝 | 7ADD | 竝 | ≡ | r-neither | | identity |

Unihan has variant mapping between U+4E26, U+5E76, U+5E77, and U+7ADD. The code points U+5002, U+5E77, and U+7ADD are part of KLGR.

6. The code point U+53DA was included in CLGR9 because of its membership in IICORE but is treated differently between CLGR9 (1st) and dotAsia (2nd). In one case, it is a variant of U+6BB5, in the other a variant of U+5047. The code points U+6BB5 and U+5047 are members of both CLGR9 and dotAsia with 'r-both' mapping.

| Source | Glyph | Target | Glyph | | Type(s) | Ref | Comment |
|---|---|---|---|---|---|---|---|
| 53DA | 叚 | 53DA | 叚 | ≡ | r-both | | identity |
| 53DA | 叚 | 6BB5 | 段 | ↔ | blocked | | |
| 6BB5 | 段 | 6BB5 | 段 | ≡ | r-both | | identity |
| **Source** | **Glyph** | **Target** | **Glyph** | | **Type(s)** | **Ref** | **Comment** |
| 5047 | 假 | 5047 | 假 | ≡ | r-both | | identity |
| 5047 | 假 | 53DA | 叚 | ↔ | blocked | | |
| 53DA | 叚 | 53DA | 叚 | ≡ | r-both | | identity |

Unihan does not bring any clarification either way (it describes a variant relationship between U+5047 and U+4EEE part of CLGR9 but not included here). This needs further investigation. The code points U+5047 and U+6BB5 are part of KLGR.

---