Considerations concerning the Chinese Root LGR

**Last updated: July 2, 2016**

# Summary

This document analyzes the content of the current Chinese LGR as specified by the Chinese Generation Panel, and makes an initial set of recommendations to the CGP. Additional recommendations will be presented in a separate document.

The latest draft of that Chinese LGR (CLGR6) is represented by two files:

* CGP Proposal Draft 20160612.docx
* Appendix B CGP Repertoire.xlsx,
* Appendix C CGP Variants.xlsx
* Appendix D WLE 20160530.xml

The following items summarize the analysis:

* The repertoire is extremely close in code point size to the full MSR-2 Hanzi set (19 738 versus 19 850).
* Like MSR-2 and the .asia ZH set, it includes those Hong Kong core characters represented by the HK IICORE subset included in the Basic Multilingual Plane (BMP),
* It also includes all IICORE characters that are part of the BMP,
* However, it does not include the same category of HK characters (62 code points) included in the Supplementary Ideographic Plane (SIP),
* It also does not include two HK characters not in IICORE and therefore not in the MSR,
* By including all IICORE characters from the BMP, it includes a vast majority of the base Japanese Kanji core set (J0), leaving only 50 J0 code points excluded from the set (from a total of 6 356 J0 characters).

The main focus of this document is to identify the bounds, and analyze the composition of the contents, of the repertoire as a whole which CGP intends for inclusion in Hani, and to correlate and compare it to other sets, including MSR-2.

All Recommendations are presented separately and highlighted.

This first version of our commentary does not analyze the variant sets required, either as to the procedures for identifying and applying them, or their proposed contents when identified. This will be done in a future iteration of this document.

More attention will likewise be given in a future commentary to the constraints on relations between the Chinese, Japanese and Korean repertoires and variant sets. The precise roles for the different GPs and the Integration Panel will only emerge when the content of the three LGRs' repertoires and variant sets have been determined.

This document does not address in detail the overall format and presentation of the proposal. However, the IP notes that the format of the LGR draft does not use the common format recommended for LGR proposal.

Please refer to [https://community.icann.org/display/croscomlgrprocedure/Document+Repository](https://community.icann.org/display/croscomlgrprocedure/Document%2BRepository) , especially the items concerning ‘Requirements for LGR proposals’ and ‘LGR Proposal Template’. It may also be useful to look at a recent proposal such the Khmer LGR as a model. See: <https://www.icann.org/en/system/files/files/proposal-khmer-lgr-15apr16-en.pdf>.

**Integration Panel Recommendation**: *Follow the common format as far as practical.*

# Definitions

## 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.

## 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. .asia Japanese <https://www.iana.org/domains/idn-tables/tables/asia_ja_1.1.txt>
2. .asia 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 .asia Chinese repertoire is itself the result of a merge of various Chinese sources such as China PRC, Hong Kong SARs, and Taiwan.

Note that MSR-2 contains a few additional 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).

## IRG Sources

The Ideograph Rapporteur Group (IRG) is an entity tasked by ISO/IEC JTC1/SC2/WG2 to create proposals of unified ideographs for encoding in ISO/IEC 10646 and the Unicode Standard. The IRG tracks source information for all ideographs, the data being documented in the Unihan Database maintained by the Unicode Consortium. Chinese sources are prefixed with the letter “G”, Taiwanese sources with the letter “T”, and so on. For definitions of these, see UAX#38 at http://unicode.org/reports/tr38.

# Existing IDNA practice for Chinese repertoires

## IDNA .cn and .tw CJK Unified Ideograph repertoire

The IDNA Chinese repertoires defined by the .cn registry [CNNIC] and the .tw registry [TWNIC] are identical and contain 19 520 Hanzi ideographs from two Unicode blocks: CJK UNIFIED IDEOGRAPHS and CJK UNIFIED IDEOGRAPHS EXTENSION A. They include the core repertoires for both simplified and traditional Chinese expressed as follows (sources are not mutually exclusive between G, H, and T sources):

|  |  |  |  |
| --- | --- | --- | --- |
| **IRG Source name** | **size** | **Included in .cn and .tw** | **Excluded**  |
| G0 | 6 763 | 6 763 | 0 |
| G1 | 2 202 | 2 201 | 1 |
| HB0 | 10 | 10 | 0 |
| HB1 | 5 401 | 5 401 | 0 |
| HB2 |  7 650 | 7 650 | 0 |
| T1 | 5 414 | 5 413 | 0 |
| T2 | 7 650 | 7 650 | 0 |

Other secondary Hanzi sources, such as G3, G5, GE, GH, GK, and T3 are also widely represented in that repertoire.

**Note**: The one exception to the inclusion of the core sets in .asia is the code point U+9DC0 (鷀), which is part of the G1 set. To recap the discussion in the MSR, this code point should have been the traditional variant of U+9E5A (鹚), but has been commonly replaced by U+9DBF (鶿) in that role. It is included in the current version of the Chinese LGR (CLGR6).

## DotAsia .asia ZH repertoire

Because the repertoire defined by .cn and .tw does not fully cover the needs of some Chinese constituencies such as Hong Kong SAR and Singapore, a new IDNA repertoire was created by DotAsia (.asia registry) for Chinese use (.asia ZH). It is a superset of these sets. It adds 163 Hanzi ideographs, of which 156 are part of HKSCS (Hong Kong Supplementary Character Set) included in the IICORE collection, 4 are GS (Singapore Characters), and the remaining 3 are part of various other Chinese sources that are necessary to insure full transitivity in variant processing.

This makes a total of 19 683 Hanzi ideographs for .asia ZH.

**Note**: This LGR contains 62 code points from the block: CJK UNIFIED IDEOGRAPHS EXTENDED B, added through the inclusion of the 156 characters from HKSCS.

**Note**: By being such a large union of various Hanzi ideographs, this repertoire includes almost all the core Japanese Kanji set (6 212 Kanji ideographs out of a total of 6 356 for the J0: JIS X 0208-1990 set).

# Current Chinese Root LGR 2016-06-12 (CLGR6, section 3.2)

The latest repertoire proposed by the Chinese Generation (CLGR6) Panel has a simple relationship to the .asia zh set:

It includes all code points from the latter, **except** the 62 code points from the IICORE HKSCS subset which are located in CJK UNIFIED IDEOGRAPHS EXTENSION B.

**Note**: A previous version of the CLGR (CLGR5) was missing an additional 7 code points (U+39DB, U+3BA3, U+43D3, U+4443, U+4882, U+4C9D, and U+4C9E) which are now included in CLGR6.

On top of the .asia ZH repertoire reduced by 62 code points, CLGR6 adds 117 characters for a total of 19 738 code points (which is only 112 less than the full MSR-2 Hanzi repertoire). These 117 code points are a mix of IICORE characters (102) and what China includes in their ‘Normalized Hanzi list for Common Use (15). A vast majority of these characters are also part of the J0 set.

**Note**: A previous version of the Integration Panel Considerations erroneously assumed that all J0 characters were part of the IICORE set. A total of 50 J0 characters are not part of IICORE and are also not part of the CLGR6.

The CGP LGR document explains in the section 3.2 ‘Repertoire formation process’ how the current set was formed and it is extremely useful for the analysis of the set.

### Code points not in MSR-2

In the second paragraph of Section 3.2 in the LGR proposal, it is mentioned that out of ‘41 new Chinese characters’ from HKIRC, ‘two of them [were] out of scope of MSR’.

After further investigation and consultation with experts from Hong Kong it was determined that these code points were U+3A5C and U+58B5.





These two code points were not included in MSR originally because they were not contained in any IDN tables known to the Integration Panel when MSR-1 and MSR-2 were created and they were not part of the IICORE set.

If the Chinese GP can provide evidence of their need for IDN, the Integration Panel would consider adding them in a future version of the MSR in timely fashion to allow the creation of a Chinese LGR including these 2 characters. Such a revision might be advisable if the addition were to result in removing or reducing a regional bias to the Root Zone by better supporting Hong Kong requirements.

### Code points not in the BMP

The CGP LGR proposal does not provide a rationale for the exclusion of the non BMP characters. These are all HKSCS characters and are specific to Cantonese. The main reason for being in Extension B (instead of the main block or Extension A) is because they were processed later by the Ideographic Rapporteur Group which is in charge of preliminary research on CJK character encoding. If the only rationale for excluding them is because of their encoding in a Supplementary Plane, this needs to be specifically discussed and justified by the Chinese Generation Panel.

Because these 62 characters are also part of the same IICORE set, there is no other rationale for exclusion beyond their code point value at this point, and their exclusion would tend to add a regional bias to the Root Zone.

### Other Issues and Recommendations

Finally, there is apparently a process issue in the creation of Appendix B. Five entries: 5CCC, 6CCC, 7CCC, 8CCC, and 9CCC are presented as 5MSS, 6MSS, 7MSS, 8MSS, and 9MSS in the code point column in the spreadsheet. They are correctly represented in the XML file.

**Integration Panel Recommendation**: *Please provide a rationale for the non-inclusion of the IICORE characters included in the Supplemental Ideographic Plane. In addition, provide these 2 missing code point values from the HKIRC set and fix the five entries noted above in Appendix B.*

# Repertoire coordination with JGP and KGP (should be section 3.3)

While the text in the proposal correctly shows that there are ‘6306 overlapped characters’ between the Kanji characters in the JGP repertoire, the following figure 7 is slightly misleading by offering counts of the totality of the JGP repertoire (including Katakana and Hiragana) but not for the totality of the KGP repertoire (one would assume that the Hangul syllables would be included in such repertoire. This is in conflict with the figure 7 name which says ‘JGP Kanji set’.

The repertoire coordination aspect should be restricted to the CJK Unified Ideographic part of the three sets (CGP, JGP, and KGP). This would show a much more complete coverage of the JGP LGR Kanji coverage by the CLGR6 (6306 Kanji out of 6356 Kanji part of the JGP LGR).

**Note**: The coverage assumption is only tentative, because the Integration Panel has not received any formal draft from either the Japanese LGR or the Korean LGR. It is however assumed that the Japanese repertoire will be limited to their J0 set.

**Integration Panel Recommendation**: *Fix the number for JGP from 6532 to 6356 which is the actual number of code points in the Kanji J0 set.*

# Variant definition in CGP (section 4.1 in CGP)

As noted in the Summary, the evaluation of the variants in the CGP proposal is preliminary and not complete.

For comparison of variants to the CPG LGR proposal there is available another XML-format Chinese LGR which is publicly open for public comment at <https://www.icann.org/sites/default/files/packages/lgr/lgr-second-level-chinese-15may16-en.xml>

This LGR was developed as part of reference for 2nd level domain. It is a transcription of the .asia (ZH) and as such shares many features with the proposed root Chinese LGR. The notable difference is that in the CGP LGR no reflexive variant is defined when a character is neither mapped to traditional nor simplified such as:

<char cp="4E04" tag="sc:Hani" ref="0 100 104">

 <var cp="4E0A" type="both" />

 <var cp="4EE9" type="blocked" />

</char>

Ideally, in such cases it is preferable to introduce an additional reflexive mapping that describes that the character is not mapped to itself in any case, like ‘r-neither’ as follows:

 <char cp="4E04" tag="sc:Hani" ref="0 100 104">

 <var cp="4E04" type="r-neither" comment="identity" />

 <var cp="4E0A" type="both" />

 <var cp="4EE9" type="blocked" />

 </char>

Concerning the 2 additional code points part of HKIRC, the documentation provided by HK experts shows them part of variant sets as follows (loosely based on RFC3743 format[[1]](#footnote-1), with 86 and 886 identifying simplified and traditional variants respectively):

㩜(0);㩜(86),㩜(886);㩜(0),揽(0),擥(0),攬(0);

墵(0);墵(86),墵(886);坛(0),埮(0),墰(0),墵(0),壇(0),壜(0),罈(0),罎(0);

The same information with code points revealed:

3A5C(0);3A5C(86),3A5C(886);3A5C(0),63FD(0),64E5(0),652C(0);

58B5(0);58B5(86),58B5(886);575B(0),57EE(0),58B0(0),58B5(0),58C7(0),58DC(0),7F48(0),7F4E(0);

Therefore, these two code points would have to be added in two existing variant tables. For U+3A5C, the members are: U+63FD, U+64E5, U+652C and U+39DB (last not present in the list above but part of the CLGR6) and for U+58B5 all code points mentioned in the list above (already defined in CLGR6, except of course U+58B5). Our understanding is that they would have ‘r-both’ type in their own entries with all other variants blocked. We are not sure what should be done for the other variant set member entries for mapping values for U+3A5C and U+58B5, but we would assume ‘blocked’.

**Integration Panel Recommendation**: *Implement the ‘r-neither’ type to indicate reflexive mapping when the character is not mapped to either traditional or simplified. Update variant tables for U+3A5C and U+58B5 as appropriate (if these code points are needed in the CGP repertoire).*

# Coordination between CJK GP (section 4.2)

While this section is useful reading, it is not relevant for the purpose of the CGP LGR documentation. It is not a foregone conclusion that the process described in that section will be followed. Many aspects are related to LGR integration, which is the responsibility of the Integration Panel. Coordination is useful between GPs to minimize the issues at the time of integration.

Any description of integration procedure and related process should be removed from this section and the amended section should be moved to an appendix.

**Integration Panel Recommendation**: *Delete the section or move it to an appendix with all aspects covering LGR integration deleted.*

# Whole Label Evaluation Rules (section 5)

Along with the introduction of the ‘r-neither’ type, the <rules> element needs to be modified to take that new type

<rules>

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

<action disp="invalid" match="leading-combining-mark" comment="labels with leading combining
 marks are invalid" />

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

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

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

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

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

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

</rules>

(The above also restores the default action for leading combining marks from the MSR, retaining which is a requirement for all LGRs).

While letting the r-neither type “fall through” to the “catch-all” rule would give the same result and would not even require that r-neither be mentioned explicitly, relying on an implicit fall through has the disadvantage is that it makes it harder for a human reader to discern the intent. At this stage of LGR development, the focus should probably be on making the design more explicit rather than trying to optimize the LGRs from a processing perspective.

**Integration Panel Recommendation**: *Implement the changes to the <rules> element.*

1. following the notation of figure 8 on p. 13 of the CGP draft. [↑](#footnote-ref-1)