Integration Panel:
Disposition of the MSR1 comment by HKIRC (Henry Chan)

Last updated: May 5, 2014

Summary

This document describes a proposed disposition by the Integration Panel of the comments made by HKIRC posted on the ICANN web site as http://forum.icann.org/lists/comments-msr-03mar14/msg00002.html. as part of the Public Comment of the Maximal Starting Repertoire (MSR1) of the Root Zone Label Generation Rule Procedure (http://www.icann.org/en/news/public-comment/msr-03mar14-en.htm). Each section shows the comments made by HKIRC along with the changes requested, and the disposition by the Integration Panel (accepted or rejected).

1 General comment about inclusion of HKSCS

1.1 HKIRC comment:

Hong Kong Internet Registration Corporation Limited (HKIRC), the registry of the .hk ccTLD and the .香港 (Chinese characters for Hong Kong) IDN ccTLD, submits the following comments on IDN Variant TLDs – LGR Procedure Implementation – Maximal Starting Repertoire Version 1 as announced on the ICANN public comment page: http://www.icann.org/en/news/public-comment/msr-03mar14-en.htm.

Inclusion of the full Hong Kong Supplementary Character Set (HKSCS) to MSR-1

HKIRC advocates the inclusion of the full Hong Kong Supplementary Character Set (HKSCS) to the MSR-1. Of the 5,009 Chinese characters from the current version of the HKSCS, there are 2,677 Chinese characters not yet included in the MSR-1. A significant number of the HKSCS characters are commonly used in Hong Kong, one of the major Chinese speaking communities in Asia.

HKIRC requests including the 2,677 missing HKSCS Chinese characters to the MSR-1 and a list of the characters in question is attached to this comment letter alongside with their respective Code Points for reference.

About HKSCS

The publication of the HKSCS is an initiative led by the Government of the Hong Kong Special Administrative Region (HKSAR), aiming at facilitating electronic communication and data exchange conducted in Chinese in Hong Kong. The current version of the 5,009 HKSCS characters includes characters used in proper names, Cantonese dialect (a lingua franca in the Guangdong Province, spoken by the majority population of Hong Kong, and by many overseas Chinese communities), and scientific terms. These characters are mainly proposed by HKSAR government departments, academic bodies, educational institutions and members of the public.

More information about the HKSCS can be found at the website of the Office of the Government Chief Information Officer of HKSAR:
http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/hkscs.

Rationale for the inclusion of the full HKSCS to MSR-1
According to the "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels" (Procedure), Integration Panel should create the maximal set of code points for the root zone, and the MSR-1 is the first deliverable to that ends. With reference to the principles in respect of individual Code Points exclusion as discussed in the document "Maximal Starting Repertoire – MSR-1-Overview and Rationale", HKIRC does not see any justification for excluding any characters of the HKSCS from the MSR-1. The inclusion of the full HKSCS also poses no risk to the DNS and implementations, as a significant number of characters from the HKSCS have been employed in electronic data exchange and Internet communications for years.

Many characters from the HKSCS are commonly used by the population of Hong Kong on a daily basis and they should be qualified as members of the set of characters used by modern living Chinese language.

With a population of more than seven millions and an Internet penetration rate of 70+%, Chinese characters commonly used in Hong Kong should not be excluded from the consideration of the ICANN root zone IDN label generation exercise.

As a result, the 2,677 missing HKSCS characters should be added to the MSR-1 for the Generation Panel's further evaluation, according to the Procedure.

1.2 Requested change

Add 2,677 missing HKSCS characters MSR1.

1.3 Disposition

This section address the generic points made in the comment. Specific points related to various repertoires are addressed in details and disposed in the following sections.

1.3.1. Clarification on HKSCS role and content

The Hong Kong Supplementary Character Set was created as a Hong Kong specific supplement to Big-5: Computer Chinese Glyph and Character Code Mapping Table [Big5]. As such, the union of both sets (HKSCS and Big-5) is to be considered as the character set used in Hong Kong.
The latest version of HKSCS was published in 2008. There has been no need to create a more recent version because further version and/or editions of the related international standards (Unicode and ISO/IEC 10646) now incorporate directly all characters required for use in Hong Kong.

Furthermore, as can be seen on page 2 and following of [HKSCS-2008], a large number of HKSCS characters are not Chinese ideographs:

Blocks of ISO/IEC 10646 and Unicode	**Range**	**Count**
Alphabetics (multiple blocks)	0000-1FFF	127

Blocks of ISO/IEC 10646 and Unicode	Range	Count
General Symbols (multiple blocks)	2000-2DFF	71
CJK Radicals (two blocks)	2E80-2FFF	29
CJK Symbols (multiple blocks)	3000-33FF	196
CJK Unified Ideographs Extension A	3400-4DBF	574
CJK Unified Ideographs	4E00-9FFF	2291
CJK Compatibility (multiple blocks)	F900-FFFF	8
CJK Unified Ideographs Extension B	20000-2A6DF	1701
CJK Unified Ideographs Extension C	2A700-2B73F	1
CJK Compatibility Ideographs Supplement	2F800-2FA1F	11
Total		**5009**

Many HKSCS characters are alphabetic, symbols, or compatibility characters. Many of these characters are in fact not IDNA 2008 PVALID according to [RFC5892] and cannot be used in any domain name, root or lower. Currently, the MSR1 repertoire contains 1949 CJK Unified Ideographs originating from HKSCS-2008, so by only requesting 2677 (which would make of total of 4626), HKIRC recognizes implicitly that not of all of HKSCS characters are suitable for domain use in the root nor should the full set be included in MSR1.

In addition, Root zone labels are not required to support every feature of a given writing system [Procedure, Section A.3.1, 1st paragraph]. Not all characters in common use can necessarily be represented in domain names. Further, the root is a shared resource, and one that does not a-priori have any language or script context, which further restricts the possible repertoire. Under the [Procedure] the Integration Panel is directed to review potential additions against a set of principles stated therein.

The 2677 additions can be categorized as follows:

- 29 CJK Radicals,
- 18 CJK Symbols and Strokes,
- 2618 CJK Unified Ideographs current excluded from MSR1,
- 12 CJK Compatibility Ideographs.

The following sections examine in detail these 2677 additions requested by HKIRC.

2 CJK Radicals (2E80-2FFF)

2.1 General:

These two blocks (CJK Radicals Supplement 2E80-2EFF and Kangxi Radicals 2F00-2FFF) contain 29 HKSCS code points. These radicals are typically not used in isolation but instead to describe other CJK characters. Most of them are homoglyphs to a regular CJK Unified Ideograph.

2.2 Requested change

Add 29 CJK radicals from HKSCS into MSR1.

2.3 Disposition

All these characters have the Unicode General_Category value of So. As such, they are not IDNA2008 PVALID characters and therefore not eligible for use in domain names.
The Integration Panel disposition is to reject this addition.

3 CJK Symbols (3000-33FF)

3.1 General

Of the 196 HKSCS characters located in these blocks, HKIRC only mentions 18:
U+3005 IDEOGRAPHIC ITERATION MARK,
U+31C0–U+31CF CJK STROKES T to N,
U+3231 PARENTHESIZED IDEOGRAPH STOCKS
The CJK Strokes are commonly used to describe component of CJK Ideographs in a recursive manner.

3.2 Requested change

Add 18 CJK Symbols to MSR1.

3.3 Disposition

3.3.1. U+3005 々 IDEOGRAPHIC ITERATION MARK

The character U+3005 IDEOGRAPHIC ITERATION MARK has the Unicode General_Category value of Lm. It is in essence a symbolic notation and while it is not confusable with a punctuation there is little reason to have it as part of a Chinese name in the root. It is not a CJK Ideograph although it may sometimes be used as a simplified form of U+4EDD 仝. Its main usage occurs in Japanese and it is worth noting that its presence in the HKSCS like the Hiragana and Katakana characters also located into HKSCS, is probably related to a desire to use HKSCS to reproduce Japanese text. (The Katakana and Hiragana subsets are not requested by HKIRC.)

The Integration Panel is not opposed to adding U+3005 to the MSR for consideration by Generation Panels. We would further encourage any generation panels working on these characters to consider whether U+3005 should be made a character variant of U+4EDD (仝), and to include the rationale for their decision.

However the Integration Panel has strong doubts that this character should be available for TLDs outside of a Japanese context.

The Integration Panel disposition is to accept this addition.

3.3.2. CJK Strokes

The CJK Strokes 31C0 to 31CF are part of a block CJK Strokes 31C0-31CF which currently contains code points up to 31D2, all CJK Ideograph components. The HKSCS set represents an incomplete set of the necessary repertoire. The incompleteness reflects the situation that HKSCS has not been updated to include latest development of both ISO/IEC 10646 and Unicode and that Hong Kong relies on these International Standards themselves instead of keeping creating new versions of HKSCS.

Furthermore, all these characters have the Unicode General_Category value of So. Because of that, they are not IDNA2008 PVALID characters and are not eligible for use in domain names.

The Integration Panel disposition is to reject this addition.

3.3.3. U+3231 ㈱ PARENTHESIZED IDEOGRAPH STOCK

This character has the Unicode General_Category value of So. Therefore, it is not an IDNA2008 PVALID character are not eligible for use in domain names.

The Integration Panel disposition is to reject this addition.

4 CJK Unified Ideographs

4.1 General:

HKSCS contains 4567 CJK Unified Ideographs located in 4 blocks: CJK Unified Ideograph and CJK Unified Ideograph Extensions A to C. MSR1 contains 1949 of them. HKIRC is asking to complete the set by adding the remaining 2618 CJK Unified Ideographs.

4.2 Proposed change

Add 2618 CJK Unified Ideographs from HKSCS 2008 to MSR1.

4.3 Proposed disposition

4.3.1. Ideographic code points

Concerning the proposed addition of Hong Kong specific CJK Unified Ideographs, it is of interest to first determine the context for the MSR1 HK content. In [Unicode63], ideographs from the following Chinese H (Hong Kong specific) sources are included:

Hanzi H sources:

H	Hong Kong Supplementary Character Set – 2008
HB0	Big-5: Computer Chinese Glyph and Character Code Mapping Table, Technical Report C-26, 電腦用中文字型與字碼對照表，技術通報C-26, 1984, Symbols
HB1	Big-5, Level 1
HB2	Big-5, Level 2

The following tables describe the content in terms of these sources:

Source name	Full Count	Included in MSR	Excluded from MSR	.asia Chinese set
H	4567	1949	2618	1921
HB0	10	10		10
HB1	5401	5401		5401
HB2	7650	7650		7650
TOTAL	**17628**	**15010**	**2618**	**14982**

The Integration Panel used the .asia Chinese set from the IANA Repository of IDN Practices [IANA] as the base for determining the Chinese set for MSR1. That set was developed by augmenting the original .cn tld [IANA] repertoire with ideographs needed for more Chinese communities, including Hong Kong. Many ideographic additions derive from the [IICORE] repertoire (a standardized collection of CJK Unified Ideographs characters deemed essential to all CJK Asian constituencies, except Vietnam).

As seen above, the Big-5-derived repertoire is fully included in MSR1. In addition, by definition of having all IICORE characters included in MSR1, all the Hong Kong specific IICORE ideographs (837) are also included.

Note further that 28 HKSCS ideographs are part of MSR1 by being part of another required set (Japanese .asia and .jp set), although they are not part of the base requirement (.asia Chinese set) for Chinese. Here is the list (Code point and H source value):

U+5227 H-89A6	U+6E76 H-92B3	U+7DE4 H-8EAE	U+8F19 H-904C
U+524F H-87BC	U+750E H-FEA9	U+8262 H-956B	U+8F5C H-9F78
U+5338 H-C6C7	U+754A H-FEB2	U+84AD H-9AF2	U+91D6 H-FC5E
U+616F H-A074	U+7724 H-A0C1	U+88B5 H-8FC4	U+91F6 H-8C50
U+6282 H-9FF4	U+799D H-FEF5	U+894D H-95A7	U+95A0 H-90B5
U+69F9 H-FD57	U+7ADA H-8E56	U+8B0C H-8FDE	U+984B H-90F8
U+6A0C H-986C	U+7BCF H-9F5D	U+8CCD H-9E48	U+9E95 H-91AB

4.3.2. The case for adding U+9DC0

Among the 2618 CJK Unified Ideographs requested by HKIRC the case of U+9DC0 deserves special attention. It is also part of the Chinese G1 set (GB12345-90), completing that set from 99.999% coverage to 100%. This particular code point has many other source references:

It is also part of a complex correlation between 3 code points:

9DC0: 鶵 9E5A: 鷁 and 9DBF: 鷲

Ideally, U+9DC0 should have been the traditional variant of U+9E5A, but U+9DBF was created earlier and ended up being the commonly accepted variant. While adding U+9DC0 at this stage would complicate the variant mapping significantly it would insure that the complete variant relationships between these three code points can be evaluated by Generations Panels and therefore properly addressed. This is the argument that speaks for including U+9DC0 in the MSR (even if the Generation Panels conclude in not adding it to any LGR).

The Integration Panel disposition is to accept the addition of U+9DC0.

4.3.3. The case for excluding U+4CA4

Among the 2618 CJK Unified Ideographs requested by HKIRC the case of U+4CA4 also deserves special attention:

The two sources have a different radical shape (simplified for the G source and traditional for the H source). This has been recognized as an error and will likely result in a dis-unification of the abstract character. Because the current radical classification uses the simplified annotation in the chart (recognizable by the apostrophe in the right of “195” in the chart), it is likely the HK character will be allocated to a new code point.

This means that the character U+4CA4 should not be used for HK context and should therefore be not be part of MSR1.

The Integration Panel disposition is to reject this addition.

4.3.4. The case for the other 2616 HKSCS CJK Unified Ideographs not part of MSR1

Given that Hong Kong experts were involved in the specification of the .asia Chinese set, and that MSR1 fully contains the IICORE set of essential ideographs (including HK ideographs) the case for adding the remainder of the HKSCS CJK Unified Ideographs appears not particularly strong. In particular, the fact that they are not included in the IDN table for the second level, makes their addition to the root appear premature. It is a recognized principle [RFC6912, section 2.1] that label generation rules for domain names should in general become more restricted as one move up the tree towards the root.

It is further worth noting that any addition of ideographs beyond those used in current IDN tables would require a careful examination of the entire set of variants to make sure that the additional ideographs are included in any variant pairs or sets.

Because nothing prevents future versions of the MSR to include additional CJK Unified Ideographs, as long as the variants study is correctly performed, it is deemed un-necessary to add these 2616 CJK characters at this moment.

The Integration Panel disposition is to reject this addition.

5 CJK Compatibility Ideographs

5.1 General

HKSCS contains 12 CJK Compatibility Ideographs located in 2 blocks: CJK Compatibility Ideograph and CJK Compatibility Ideograph Supplement. HKIRC is asking to add these 12 code points to MSR1.

5.2 Requested change

Add 12 CJK Compatibility Ideographs from [HKSCS-2008] to MSR1.

5.3 Proposed disposition

CJK Compatibility Ideographs are not IDNA2008 PVALID and therefore are not eligible to be part of MSR1.

The Integration Panel disposition is to reject this addition.

6 Conclusion

Based on these considerations, the Integration Panel concludes that two characters should be considered for addition to MSR1:

- U+3005 IDEOGRAPHIC ITERATION MARK but with a suggestion to restrict it to the “Japn” repertoire, this is in fact a proposed disposition for a parallel Japanese comment to MSR1.
- U+9DC0 to complete a complex variant group (along with 9E5A and 9DBF which are already in MSR1).

Among the other 2617 CJK Unified Ideographs that were requested, the Integration Panel is open to the possibility that a case could be made for adding a more limited subset to future version of the MSR.

Integration Panel Membership

This document is released by the Integration panel, composed of Asmus Freytag, Michel Suignard, Wil Tan, Nicholas Ostler, and Marc Blanchet.

References

[Big5] An overview of the Big5 character set is available at http://en.wikipedia.org/wiki/Big5.

[HKSCS-2008] Office of the Government Chief Information Officer & Official Languages Division, Civil Service Bureau, The Government of the Hong Kong Special Administrative Region, "Hong Kong Supplemental Character Set- 2008," December 2009 , available as http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/terms/doc/e_hkscs_2008.pdf, in particular section 2: http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/terms/doc/e_sect2_2008.pdf.

[IANA] *Repository of IDN Practices,* http://www.iana.org/domains/idn-tables .

[IICORE] *International Ideographs Core (IICORE),* http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/iso_10646/iicore.htm . Visited 2014-01-07.

[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

[RFC5892] Faltstrom, P., Ed., "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)", RFC 5892, August 2010. http://tools.ietf.org/html/rfc5892

[RFC6912] Sullivan, A., Thaler, D., Klensin, J., and Kolkman, O., "Principles for Unicode Code Point Inclusion in Labels in the DNS", RFC 6912, April 2013 http://tools.ietf.org/html/rfc6912

[Unicode63] The Unicode Consortium. The Unicode Standard, Version 6.3.0, defined by: "The Unicode Standard, Version 6.3.0", (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5). http://www.unicode.org/versions/Unicode6.3.0/.