Integration Panel:
Proposed Disposition of the MSR1 comment made by Yoshiro YONEYA

Last updated: May 7, 2014

Summary

This document describes a proposed disposition by the Integration Panel of the comments made by Yoshiro YONEYA posted on the ICANN web site as http://forum.icann.org/lists/comments-msr-03mar14/msg00000.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 Yoshiro Yoneya along with the changes requested, and the disposition by the Integration Panel (accepted or rejected).

1 ContextO and the case of U+30FB ・KATAKANA MIDDLE DOT

1.1 Received comment (Reference 5.2 CONTEXT O Code Points):

The code point U+30FB (KATAKANA MIDDLE DOT) is a character which is often used in Japanese services, products, trademarks and so on. Therefore it should not be listed in exclusions. U+30FB is defined as True (OK as used) if all other characters in the string are Japanese characters (Hani, Hira, Kana) in IDNA2008 (RFC 5982), so it should be USABLE (part of MSR) unless it is used alone or mixed with non-Japanese scripts.

1.2 Requested change

Add U+30FB KATAKANA MIDDLE DOT to MSR1.

1.3 Disposition

There are multiple reasons to not add U+30FB to MSR1.

1) It is an IDNA2008 CONTEXTO character. The document 'Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels' [Procedure] specifically excludes that type of character in its section B.3.4.2.
2) Its General_Category value is Po (Punctuation_Other) and by the Procedure it is out of scope for the Root Zone LGR (Letter principle).
3) While not a homoglyph, it has a strong visual similarity to various middle dots (including U+00B7 MIDDLE DOT, itself another CONTEXTO character) which are also excluded from MSR1.

The Integration Panel disposition is to reject this addition.

2 Character U+3006 〆 IDEOGRAPHIC CLOSING MARK used in names

2.1 Received comment (Reference 5.8 IDNA 2008 Gaps and Side Effects):

The code point U+3006 (IDEOGRAPHIC CLOSING MARK) is a character which is used in Japanese personal names and geographic names. Therefore it should not be listed in exclusions. As well as (1) above, U+3006 should be USABLE (part of MSR) unless it is used alone or mixed with non-Japanese scripts.

2.2 Requested change

Add U+3006 IDEOGRAPHIC CLOSING MARK to MSR1.

2.3 Disposition

The character U+3006 IDEOGRAPHIC CLOSING MARK has the Unicode General_Category value of Lo. 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 Japanese name in the root. It is not a CJK Ideograph although it may sometimes be used as an abbreviation of 7DE0 締. The code point U+3006 is also used as a substitute for U+9589 (閉) (Link in Japanese: http://dictionary.goo.ne.jp/leaf/jn2/101073/m0u/).

The Integration Panel is not opposed to adding U+3006 to the MSR for consideration by Generation Panels. We would further encourage any generation panels working on these characters to consider whether U+3006 should be made a character variant of U+4E44 (メ), and to include the rationale for their decision. (See the following link for further details on definition and usage: http://en.wiktionary.org/wiki/%E4%B9%84#Japanese.)

The Integration Panel disposition is to accept this addition.

3 U+3005 々 IDEOGRAPHIC ITERATION MARK and U+3007 〇 IDEOGRAPHIC NUMBER ZERO

3.1 Received Comment (Reference 5.16 Whole Block Exclusions):

Code points U+3005 (IDEOGRAPHIC ITERATION MARK), U+3006 (IDEOGRAPHIC CLOSING MARK, see (2) above), and U+3007 (IDEOGRAPHIC NUMBER ZERO) that are included in the range of U+3000-U+303F (CJK Symbols and Punctuation) are quasi Han characters and used in Japanese personal names, geographic names, service names, product names, trademarks and so on. Especially, as U+3005 is often used for Japanese personal names and geographic names, its exclusion is extremely inexpedient in composing Japanese strings. As well as (2) above, these three characters (U+3005, U+3006 and U+3007) should be USABLE (part of MSR) unless it is used alone or mixed with non-Japanese scripts.

3.2 Requested change

Add U+3007 IDEOGRAPHIC NUMBER ZERO (along with U+3005 and U+3006 above) 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 Japanese name in the root. It is not a CJK Ideograph although it may sometimes be used as a simplified form of U+4EDD 仝.

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. (See the following link for further details on definition and usage: http://en.wiktionary.org/wiki/%E3%80%85.)

The Root zone labels are not required to support every feature of a given writing system [Procedure, Section A.3.1, 1st paragraph]. Further, the root is a shared resource, and one that does not a-priori have any language or script context. If this character, along with U+3006 IDEOGRAPHIC CLOSING MARK, was considered for the LGR, its usage would have to be restricted to Japanese context, which, because of the prohibition against script mixing would not in practice require a context rule.

The Integration Panel disposition is to accept this addition.

3.3.2. U+3007 〇 IDEOGRAPHIC NUMBER ZERO

The character U+3007 IDEOGRAPHIC NUMBER ZERO has General _Category value of “Nl”, which makes it the representation of a number and provides the reason for exclusion from the root zone under the Letter Principle. This character is therefore ineligible for addition to MSR1.

The Integration Panel disposition is to reject this addition.

4 Addition of 16 Katakana Phonetic Extensions and 1208 Han Ideographs

4.1 Received Comment (Reference 5.13 Han Ideographs):

[IICORE] is specified as a source of Hani range, but it does not cover all of Han characters used for personal names and geographic names in recent Japanese writing. Therefore, selection of Japanese Hani characters (sc:Japn) should be in the range of JIS X 0221 annex JA that covers recent Japanese writing. (Attached list is a set of Japanese Hani characters that are not included in MSR-1).

4.2 Requested change

Add 16 Katakana Phonetic Extensions (U+31F0..U+31FF) and 1208 CJK Ideographs (listed both in comment) to MSR1.

4.3 Disposition

4.3.1. Katakana code points

The code points U+31F0..U+31FF KATAKANA LETTER SMALL KU .. KATAKANA LETTER SMALL RO are phonetic extensions for Ainu and are not in modern use. They should not be added to MSR1.

The Integration Panel disposition is to reject this addition.

4.3.2. Ideographic code points

Concerning the CJK Unified Ideographs additions, it is first interesting to determine the context for the MSR1 Japanese content. In [Unicode 6.3], ideographs from the following Japanese sources are included:

Kanji J sources:

J0	JIS X 0208-1990
J1	JIS X 0212-1990
J3	JIS X 0213:2000 level-3
J3A	JIS X 0213:2004 level-3
J4	JIS X 0213:2000 level-4
JA	Unified Japanese IT Vendors Contemporary Ideographs, 1993
JH	Hanyo-Denshi Program (汎用電子情報交換環境整備プログラム), 2002-2009
JK	Japanese KOKUJI Collection
JARIB	Association of Radio Industries and Businesses (ARIB) ARIB STD-B24 Version 5.1, March 14 2007

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

Source name	Full Count	Included in MSR	Excluded from MSR	.jp or .asia Japanese set
J0	6356	6356		6356
J1	5801	5210	591	-
J3	125	83	42	-
J3A	8	7	1	-
J4	659	266	393	-
JA	660	3	657	-
JH	107		107	-
JK	367		367	-
JARIB	3		3	-
TOTAL	**14086**	**11925**	**2161**	**6356**

The .jp and the .asia Japanese set in the IANA Repository of IDN Practices [IANA] were the bases for determining the Japanese set for MSR1. Their repertoire of Unified CJK ideographs consists solely of the J0 repertoire. By virtue of having derived the MSR1 CJK repertoire from the union of the .asia Chinese set and IICORE, the Japanese CJK repertoire covered in the MSR grows from 6356 characters to 11925 characters, which almost doubles what is present in the current IDN table for the ccTLD.

4.3.3. The 1208 proposed CJK Ideograph additions to the MSR

The next evaluation step is to dispose the proposed 1208 CJK Ideographs in the same context (note that all counts refer to CJK Unified Ideographs, except that the code points listed as unaccounted include some that are not CJK Unified Ideographs):

Source name	Full Count	Included in MSR	Excluded from MSR	.jp or .asia Japanese set	Proposed additions
J0	6356	6356		6356	
J1	5801	5210	591		591
J3	125	83	42		42
J3A	8	7	1		1
J4	659	266	393		393
JA	660	3	657		86
JH	107		107		
JK	367		367		
JARIB	3		3		
unaccounted					95
TOTAL	**14086**	**11925**	**2161**	**6356**	**1208**

Out of the proposed 1208 code point additions, 1113 code points would complete the J1, J3, J3A, and J4 sets and also add 86 code points to the JA set, but not complete it. The additional unaccounted 95 code points (1208-1113) are discussed in the next section.

4.3.4. The 95 unaccounted code points

The following table summarizes how these unaccounted 95 code points are allocated:

Source name	Count	PVALID	Comment
J3 source CJK Compatibility block	69	No	In block F900..FAFF
J4 source CJK Compatibility block	6	No	In block F900..FAFF
U source CJK Compatibility block	11	No	In block F900..FAFF
Various non-J unified CJK Main block	5	Yes	In block 4E00..9FFF
Various non-J unified CJK Compatibility block	4	Yes	In block F900..FAFF
Unaccounted (total)	**95**		

The 86 (69+6+11) CJK Compatibility ideographs with J3, J4, and U source sources are not IDNA2008 PVALID according to [RFC5892] and therefore are not eligible to be part of MSR1. (All compatibility ideographs are normalized away by NFC.)

These leaves 9 CJK Unified ideographs, 5 located in the CJK main block and 4 in the CJK Compatibility block (despite the block name, these 4 characters are Unified Ideographs, as they are part of the 12 Unified Ideographs that for historical reasons are included in that block in Unicode). Here is a more detailed description of these 9 characters:

Main block (4E00-9FFF)

Code point, Radical, Stroke count, Sources				
4EFC 人 9.4	任 GE-2151		任 T3-226E	
50F4 人 9.12	僩 GE-223E	僩 H-97C4	僩 T3-4574	僩 K2-235F
51EC 几 16.5	凪 G5-334A		凪 T3-245B	
5CF5 山 46.7	峐 G5-3B35	峐 H-8C48	峐 T3-3062	峐 K2-2E71
8807 虫 142.13	蠙 GE-3C61		蠙 T3-5B29	

CJK Compatibility block (F900-FAFF)

Code point, Radical, Stroke count, Sources		
FA0E 又 29.11	叟 UTC-00843	
FA27 金 167.7	鋿 UTC-00852	
FA28 金 167.8	錞 TF-584C	錞 UTC-00853
FA29 阜 170.10	陼 UTC-00854	

To summarize, out of the 1208 proposed ideographs, 86 characters are compatibility characters and are not PVALID, so these definitively cannot be added to MSR1. This leaves 1122 code points to be considered.

4.3.5. The set of 1122 code points corresponding to CJK Unified Ideographs

For the 1122 ideographs proposed, given that none of the other code points from the same set of J-sources are currently part of any IDN tables for Japanese (whether for .jp or .asia), the case for including these code points in MSR1 is weak. (Neither the .jp ccTLD or the .asia tables contain any code points for any sources other than J0, which is fully covered by the MSR).

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, therefore the fact that these are not in the IDN tables for .jp or Japanese table for .asia is highly relevant.

The only code point from this set of 1122 that should be considered is U+9DC0 because it completes a variant set and including it allows a more complete analysis of a complex variant relationship by the Generation Panels (see full explanation in the next section). If included in the MSR, it would become one of the code points that may not be appropriate for the LGR, despite being listed in the MSR — because its purpose for inclusion in the MSR would primarily be to allow investigation by the relevant GPs.

4.3.6. The case for adding U+9DC0

The code point U+9DC0 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: 鵧 9E5A: 鵺 and 9DBF: 鷂

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 Generation 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).

4.3.7. Disposition summary

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

5 Addition of sc:Japn to U+30FC ー KATAKANA-HIRAGANA PROLONGED SOUND MARK

5.1 Received Comment (Reference Others):

"sc:Japn" should be added for tag of U+30FC (KATAKANA-HIRAGANA PROLONGED SOUND MARK).

5.2 Requested change

See above.

5.3 Disposition

The name space for the Unicode scrip property ‘sc’ does not include the value ‘Japn’, which means that the MSR1 XML file should not use the ‘sc:Japn’ notation. It has been suggested to instead use the content of the Unicode Script_Extensions property value which includes for each code point an enumeration of script values that are valid ‘sc’ values. Because the Script_Extensions property value for

U+30FC is 'Hira Kana' (standing for Hiragana and Katakana), the MSR1 XML file tag value would be: tag="sc:Hira sc:Kana". In that model, any code point that has a tag including sc:Hira, sc:Kana, or sc:Hani may be included in a Japanese context.

The Integration Panel disposition is to use the Unicode Script Extensions property in the MSR1 XML file.

6 Conclusion

Based on these considerations, three characters could be considered for addition:

- U+3005 IDEOGRAPHIC ITERATION MARK but with a suggestion to restrict it to the "Japn" repertoire,
- U+3006 IDEOGRAPHIC CLOSING MARK , with the same restriction to the "Japn" repertoire,
- U+9DC0 to complete a complex variant group (along with 9E5A and 9DBF which are already in MSR1).

In addition, the sc: tags included in the MSR1 XML file should be modified to comprise the Unicode Script_Extensions property values. This use of the Script_Extensions property was also anticipated in the Procedure. In that context, modifying the tag for 30FC to "sc:Hira sc:Kana" makes sense.

Of the remaining code points, several are ineligible for use to the root, because they are not letters, because they are not PVALID, or for other reasons. While there are no insurmountable formal obstacles to including the remaining 1121 code points corresponding to CJK Unified Ideographs, the case for them is enough weak in light of existing IDN practice that the Integration Panel regards it as insufficient.

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

[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/.