Maximal Starting Repertoire version 3 (MSR-3) for Root Zone Label Generation Rules (RZ-LGR)

Publication Date:	14 March 2018
Prepared By:	Integration Panel for RZ-LGR

Public Comment Proceeding	
Open Date:	17 January 2018
Close Date:	26 February 2018
Staff Report Due Date:	13 March 2018

Important Information Links
Announcement
Public Comment Proceeding
View Comments Submitted

Staff Contact:	Sarmad Hussain	**Email:**	sarmad.hussain@icann.org

Section I: General Overview and Next Steps

ICANN has released version 3 of the Maximal Starting Repertoire (MSR-3: HTML, XML) for public comment. This version is upwardly compatible with MSR-2 and adds three code points each to the repertoires of Han and Latin scripts. Under the Procedure to Develop and Maintain Label Generation Rules for the Root Zone with Respect to IDN Labels, the MSR is the starting point for the work by community based Generation Panels which are developing the proposals for relevant scripts for the Root Zone Label Generation Rules (RZ-LGR). The contents of MSR-3 and the detailed rationale behind its development are described in MSR-3-Overview and Rationale. MSR-3 will cover the same scripts. The Integration Panel will finalize the code point repertoire for MSR-3 based on the feedback received by the community. After the release of MSR-3, Generation Panels which are developing their RZ-LGR proposals will be able to use the updated contents as a starting point for their analysis.

Section II: Contributors

At the time this report was prepared, a total of three (3) community submissions had been posted to the forum. The contributors, both individuals and organizations/groups, are listed below in chronological order by posting date with initials noted. To the extent that quotations are used in the foregoing narrative (Section III), such citations will reference the contributor's initials.

Organizations and Groups:

Name	Submitted by	Initials

Individuals:

Name	Affiliation (if provided)	Initials
Bill JOURIS		BJ
Yoshiro YONEYA		YY

Section III: Summary of Comments

General Disclaimer: *This section intends to summarize broadly and comprehensively the comments submitted to this public comment proceeding but does not address every specific position stated by each contributor. The preparer recommends that readers interested in specific aspects of any of the summarized comments, or the full context of others, refer directly to the specific contributions at the link referenced above (View Comments Submitted).*

BJ provides a comment and a follow-up regarding section 5.7.5, objecting to the exclusion of six code points, among them U+01C0, U+01C1, and U+01C3.

YY provides three comments. Comment 1 welcomes the addition of Joyo (daily use) Kanji U+2098F. Comments 2 and 3 discuss the appropriateness of a variant definition for U+3005/U+4EDD and U+3006/(U+7DE0,U+9589, and U+4E44). These code points are mentioned in Section 5.14.1 of MSR-3.

Section IV: Analysis of Comments

General Disclaimer: This section intends to provide an analysis and evaluation of the comments submitted along with explanations regarding the basis for any recommendations provided within the analysis.

In his comments, BJ argues that six Latin script code points while closely resembling punctuation marks, are used today in a number of writing systems as letters and should be included in MSR-3 under the Letter Principle. The Integration Panel is charged by the [Procedure] to short-list the set of code points that are PVALID under IDNA 2008 to those that are appropriate for the Root Zone. The Letter Principle, applied in this context, means that the Root Zone must not support symbols or punctuation, nor digits, even if they are PVALID.

However, the letter principle does not simply mean that every letter must be allowed in the Root Zone, nor does it constitute the only principle that the Integration Panel is charged to follow. There are a number of other principles, many of which address the special requirements for the Root Zone to be stable and secure when shared by users of all writing system, including those not familiar with the Latin script.

Arguably the most important principle among these is the Conservatism Principle, which demands that code points for which there is some doubt, for example about their security implications, not be included in the Root Zone. The principles themselves originate with the Internet Architecture Board and are listed in RFC 6912.

This RFC explicitly discusses the example of one code points that is, like the characters being commented on, a duplicate encoding of a punctuation mark that has formally been given the letter property.

It is stated in RFC 6912:

"Public zones are, by definition, zones that are shared by different groups of people. Therefore, any decision to permit a code point in a public zone (including the root) should be as conservative as practicable. Doubts should always be resolved in favor of rejecting a code point for inclusion rather than in favor of including it, in order to minimize risk."

And:

"It is not clear that all code points permitted under IDNA2008 that have a General_Category of Lo or Lm are appropriate for a zone such as the root zone. ... not every code point with a General_Category of Ll, Lo, or Lm is consistent with the type of conservatism principle discussed..."

In evaluating the code points in question, the Integration Panel is bound by the set of Principles in general, but also by the specific stance on U+02BC in RFC 6912, which in all respects exemplifies the issues exhibited by the code points in this comment. While the Integration Panel is well aware of other writing systems it still consider that the listed code points, no matter how well-established they are in important writing systems, may be doubtful in their security implication for users with other backgrounds, and therefore subject to the Conservatism Principle. In doing so, it is effectively implementing an extension of a policy from a single example given explicitly in RFC 6912 to a class of like code points.

In his follow-up comment, BJ points to the case of U+1E37 ḷ LATIN LETTER SMALL L WITH DOT BELOW and U+013C ļ LATIN SMALL LETTER L WITH CEDILLA. In a sans-serif font, the “l” would be a straight line, and both of these code points would not only look similar to each other, but also have some similarity with an exclamation mark (!).

The Integration Panel (IP) is cognizant of the fact that, especially at small sizes and in sans-serif fonts, a number of diacritics below the character are particularly difficult to distinguish. The MSR mentions (in Section 4.5) the example of *comma below* and *cedilla below* (but could have also included dot below); it recommends to Generation Panels to investigate whether these present risks that need mitigation, for example by defining them as variants.

The IP feels that an exclamation mark, unlike the two letters in question, is sufficiently differentiated enough not to require pre-emptive removal of these code points: it does not extend below the baseline, while also having a shorter stroke than the letters. This is unlike the case of U+02BC, where fonts would use an outline *identical* to U+2019, or U+01C3 (!) LATIN LETTER RETROFLEX CLICK, which also looks identical to U+0021 (!) EXCLAMATION MARK — making it impossible to tell any difference, even for the most observant user.

In this context, it bears noting that the MSR states:

“[The IP] recognizes that the MSR is merely an interim step in the development on the LGR, and that any code points included in it, are not automatically added to the RZ-LGR; the MSR is only one of several constraints on the final LGR.

The expectation is that the Generation Panel will give these code points the benefit of very careful review and that they will be accompanied by a detailed rationale, should they be included in the LGR proposal. In turn the IP will use those Principles when reviewing LGR proposals for integration.”

In the case of the small letters “l” with diacritics below, their inclusion in the MSR does not guarantee that they will automatically be eligible to be included in the Root Zone. Instead, the IP will carefully review any justification provided by the respective Generation Panel, which is expected to address not only the desire to support a given set of writing systems, but also a careful weighing of any security risks introduced as well as a proposal how to mitigate these as far as possible.

YY presents three comments. The first comment supports the addition of the “daily use (joyo)” Kanji which extends the MSR to cover the complete set of these.

The two subsequent comments address the special code points discussed in items (1) and (2) of section 5.14.1. After some detailed arguments, YY comes to the conclusion that these should not be variants.

The IP notes that the MSR specifies *repertoire* only, and does not contain any specification of variants. However, it contains several instances, including items (1) and (2) of section 5.14.1, where the IP recommends to the Generation Panels to carefully evaluate certain code points for possible variant relations. Accordingly, the IP encourages YY to work with the Japanese Generation Panel to include the analysis into the Japanese LGR.