Latin GP Proposal: IP Feedback

Date: 2017-03-22

Authors: Integration Panel

Re: Proposal for Generation Panel for Latin Script Label Generation Ruleset for the Root Zone

Date: 2017-03-07

Document version 1

Authors: Latin Script Generation Panel

1 Introduction

This is a response by ICANN Integration Panel (henceforth IP) to Proposal for Generation Panel for Latin Script Label Generation Ruleset for the Root, submitted by the Latin Generation Panel (henceforth GP).

For convenience, IP has constructed a Table of Contents for the Latin GP Proposal:

From this table, it is evident that the primary constituents of the Proposal are a General introduction which highlights diacritics and cross-script variant relations, the Proposed Initial Composition of the Panel, and the Work-plan.

These items will be the main topics of IP comment, with specific intent to focus the GP's activities. (By contrast, the Proposal's disquisition on complex and irregular details of the Latin script (esp. in section 1.2.2) leads to no clear conclusion, and may be better omitted.

2 Cross-script Variants

The consensus of the IP is that the scope of this class of characters should be the first concern of the Latin GP, since there is already a body of proposed equivalents in the Armenian and Cyrillic proposed LGRs, and only the Greek GP (already seated) is now addressing this issue. Cyrillic and Latin are the main scripts with a substantial overlap here.

This priority may require change in the Work-Plan and its time line.

Because of the nature of the development and adoption of the Latin script, it can be expected that such an initial subset already contains with few or no exceptions all of the candidates for cross-script variants to Cyrillic, Greek and Armenian and would allow a rather earlier investigation of this crucial relationship among scripts than the order of work sketched out in the timeline.

Overall, this is the main concern that besets the integration of Latin, Cyrillic, Greek and Armenian scripts. These cross-script issues should be agreed, in principle, before attending to other details in the repertoire, in-script variants and possible WLE rules.

The kinds of variants to be defined in the Root Zone LGR are limited to homoglyphs, which are characters with essentially identical appearance by design, instead of merely similar appearance. This explicitly excludes the consideration of more or less distant similarities in visual appearance, and so sharply reduces the number of cases needing to be considered.

The Latin script therefore promises a very limited scope for the definition of variants. (For that reason, the development of "principles" for the identification is probably best undertaken in parallel or in immediate context with specifying how code points will be included).

3 Procedure in the light of realism on Variants

A next step in the streamlined process might use the initial set as a basis to filter out any languages that do not further contribute to the set of required code points. For the remaining languages, a reasonable scheme of prioritization should be adopted, allowing progress to be made for the most widely written (not spoken) and most formally transmitted languages (those with lower EGIDS scores) first.

While it can be useful for cross-checking against inadvertent oversights, there is no requirement in the procedure that a GP analyze all available code points in the MSR; the Latin script, in particular, contains many code points added for the support of phonetic transcriptions and other notations that are not relevant to domain names. While MSR-2 attempted to identify clear-cut cases, it is not the case that it follows that all code points need to be investigated as to whether they might not be needed.

Instead, (logically speaking) once a determination is made that a language's written use of the Latin script makes it eligible for support in the Root Zone, that part of its repertoire (verified and vetted) which is not yet covered should be added to the draft repertoire of the LGR, and the investigation continue with the next language.

Any code points that were in MSR-2 but are left over at the end of such an inclusion based process, simply have the status of being not included (which is not the same as calling them "excluded", a term better reserved for the rarer case that a code point is associated with some supported language, but that association is deemed inadmissible, for example, because the code point is historical in nature, or perhaps not a letter).

By following some order of prioritization and proceeding in this kind of additive inclusion process, the GP would ensure that at each point in the development, the current status of the LGR draft maximizes the coverage both by population and by intensity of use. This minimizes the risk that any bottleneck in available resources would lead to the project having to be abandoned with an unusable LGR draft - instead, it guarantees that no matter when and for what reasons the project is deemed final, it maximizes the utility of the repertoire developed.

Proceeding in prioritization order would also discover the strongest cases for rectifying potential omissions in MSR-2 as early as possible, that is, the first language found to require an omitted code point would be the most widely and most intensively used one. This would allow the IP to maximize the utility of an extension to MSR-2; in particular, given the necessity to make any such extension ahead of the actual completion of the Latin LGR proposal, it is crucial that all (or if not all, then the most important) such omissions be identified in time to complete an MSR revision cycle so that the LGR can be based on the revised MSR.

Likewise, progression by languages in priority order can be used to verify that there are no pending (post 6.3.0) code points

a) that might be required, and
b) that might cause incompatibilities if added later.[1]

This approach would see to recommend itself over the inverse process of reviewing all the pending and omitted repertoires, whether or not a use case exists for them.

[1] However, concerning limitations to the repertoire (whether by Unicode 6.3.0 in general and MSR-2 in particular) it is worth emphasizing that the IAB has categorically ruled out making an exception for the use of U+02BC.

4 Diacritics

There is some imprecision in the document: e.g. (section 1.2.1),

> 'these multiple representations are usually eliminated by normalization, except for cases where no precomposed forms have not yet been encoded in Unicode 6.3 (on which [IDNA 2008] is based).'

Actually, Unicode 6.3 was not the crucial watershed. It was much earlier in Unicode 3.2 when NFC determined what had to be precomposed and what had to be decomposed.

Since that version, all new characters in Latin script that can be decomposed are mapped onto a NFC representation in terms of their combining sequence. You could still possibly introduce a precomposed character, but NFC would decompose it, so that the encoding of such a precomposed character would be pointless. But in fact, it is unwise to put too much emphasis on the any post-Unicode 6.3 characters, since they might be overtaken by future changes to MSR.

Specifically, the last 3 paragraphs of section 1.2.1 talking about normalization forms are confusing, rather than helpful. IP suggests removing these three paragraphs (starting with « Nonetheless » and ending with « back rounded vowel ».

5 Proposed Initial Composition of the Panel

Given the breadth of the use of the Latin script, the direct expertise of the various GP members does not adequately cover the total expected usage. It is therefore critically that the GP work with ICANN **to expand its membership,** and to ensure access to qualified advisers.

6 Proposed Work-plan

As suggested above, the estimate in the projected timeline, which reserves 38 weeks for the development of variant mappings, seems too long.

More generally, the schedule is much too long. Counting the parallel activities once, it might run to 80 weeks, with the variant part alone taking up 28 weeks. There must be scope to run activities in parallel.

7 Appendix A: World languages using Latin script

Whatever the finally adopted LGR development approach for the Latin GP will be, it needs to be cognizant of the issues posed by the very large number of languages that can be written in the Latin script and which includes so many languages which upon cursory examination appear unlikely to lead to any reasonably immediate demand for TLDs. It also needs to account for the fact that, unlike so many other scripts, the number of Latin code points in Unicode vastly exceeds those that will be

needed for the Root Zone, and that even the envelope provided by MSR-2 may well turn out to be more generous than for other scripts.

Appendix A reminds us of this, but is not a fully working model. Rather, it would be useful to treat it as an approximation. (For example, if it was taken literally, one would conclude that about a quarter of the 82m population of Germany does not speak German, especially given that some of the total 69m speakers found in Appendix A would need to be attributed to Austria, Switzerland, etc.). With discrepancies as large as this for well-documented languages, discrepancies for lesser-known languages may well be even more significant.

More important than vetting or improving the data in this appendix would be to get a clear statement of policy from the Latin GP on how they will
1) identify in the end which languages are supported by the Latin LGR
2) identify the index language that provides the crucial impetus to add a code point to the LGR
3) identify the primary languages that use a certain code point (other than a-z)

Appendix A lists 180 languages of EGIDS values 4 and below, and another 109 languages at level 5. Some of the languages listed, in either group, are shown with populations of only a few hundred. The full list of languages at all levels runs to nearly 500 and the lowest population figure listed is 0. Yet, the document states that this is only "a subset" of the available data.

Without adopting a somewhat streamlined process of dealing with such a large number of languages, it is feared that the GP will spend valuable time considering cases of no or marginal practical importance.

The expectation is that the GP will develop an approach is grounded in the "inclusion principle" demanded by the procedure (which demands a positive use case before adding a code point) and also recognizes that data on language use are often only a limited "proxy" for the more relevant information describing the written use of the language, and in particular the expected level of use in domain names.

While the LGR process has as one of its stated goals the avoidance of bias, this is only one of several goals, and it is explicitly not the task of a GP to predict the future, or, more concretely, predict the future development of, or usage patterns for any orthography and writing system. Analysis of code points either already encoded past Unicode 6.3.0 should be clearly limited to cases where the later introduction of a code point in a future LGR could lead to incompatibilities with the current LGR proposal.

Such a streamlined process does not need to be fully developed at the time a GP is seated, but it might nevertheless be useful to indicate the necessity for it and to give an outline or parameters for its development by the GP as starting point for their detailed development.

For example, it may be possible to relatively quickly "clear the table" by adding up the repertoire for languages that are already well supported in domain names, and for which IDN tables or LGRs (including reference tables for the second level) already exist. (With suitable vetting/verification to ensure code points are suitable for the root zone).

8 Conclusion

In general, there is no wish in the IP to delay the seating of this GP. It should be seated without delay, although with modifications to the plan as suggested above,

noting especially:

- early settlement of characters affected by Cyrillic/Greek/Armenian variant set (section 2)
- need for expanded membership (section 5)
- condensing the work plan (section 6).

Overall, the plan as set out by the existing GP shows a very useful knowledge of the issues that the GP will encounter. Also, Latin is an area where the IP has a lot of expertise and can help the Latin GP, if necessary, when formulating the LGR.