Analysis of Variants in the Latin Script for the Root Zone

Introduction

Although the Integration Panel expects this Latin GP to produce a very narrow list of variants for the Latin script[1], this group believes it is necessary (for the benefit of the reader) to examine different cases for potential variants in the Latin script.

The guiding principle for variants in the Latin script
Given the use of the Latin script in many languages, two code points or code point sequences will be treated as variants when the relationship is sufficiently universal across the entire Latin script community. For example, orthographic conventions in one specific language may not be suitable for consideration because the rule will not be understood by the rest of the population (Least Astonishment Principle).

Scope

Only code points in MSR[2] or Latin Code Point Repertoire for the Root Zone are subject to variant analysis. For example, uppercase letters are out of scope because they are disallowed in IDNA2008, hence not part of MSR-2.

Since the GP does not expect any script-mixing within the Latin script-using community, and mixed-script repertoires are discouraged as per the Procedure[3], the GP will focus its analysis on related scripts (specifically those descended from Ancient Phoenician: i.e. Greek, Latin, Cyrillic, and Armenian) or scripts which have historically been in contact with them. The WG will give preference to previously identified conflicts, but should also look beyond what is proposed in

---

[1] "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." *Integration Panel Feedback to Latin GP Proposal, 22 March 2017*

[2] Initial work will use current version as of September 2017 or MSR-2. The Panel will adjust its work analysis on the then current version of the Maximal Starting Repertoire.

[3] "It is anticipated that [...] script mixing would be normally be restricted by the integration panel, rather than allowed to be applied widely." *Section B.3.2, Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels*

GP's proposal. There are cases (e.g. Latin è (U+00E8) and Cyrillic è (U+0450)) that are not documented in Cyrillic's Root Zone LGR proposal.

Classes of Variants

For a more detailed analysis on taxonomy of variants in the Latin script the reader can refer to [Considerations in the use of the Latin script in variant internationalized top-level domains].

1. Visually Similar: These are instances where a code point or code point sequence is visually similar to another code point or code point sequence. In the context of developing the LGR for the Latin Script for the Root Zone, this group will examine code points or code point sequences that are identical by design or which are, in the judgement of the working group, sufficiently similar.

   The group is aware of the effects related to user interfaces, rendering, fonts and typefaces that may render any two glyphs similar enough to be indistinguishable to particular users with a specific socio-linguistic background. In other instances however, code points may be conceptually the same, such as ' (U+2019) and ' (U+0027) - while these code points are excluded from IDNA and MSR-2[4], such similarity may occur also with other code points, which may be eligible for integration into the LGR. For example: 'e' in Latin (U+0065) and 'e' in Cyrillic(U+0435) have identical glyphs and would be considered variants.

2. Orthographic considerations: Visual similarity as discussed in 1 is usually the outcome of historical processes, where different renditions or glyphs for the same character have developed into different characters. However, in some cases, parts of the script using community may still consider such to be identical, such as German Eszett, which is considered a ligature of 'sz' or 'ss' by common users. Where such semantic identity occurs across a wider part of the script using community, the Generation Panel will consider such as candidates for variants, and it is expected that two such types of semantic similarities may occur, namely:
   a. Digraphs or ligatures (e.g. 'æ' (U+00E6) with 'ae' (U+0061 U+0065))
   b. Phonetics and language conventions (e.g. the case of Eszett in German: 'ss' (U+0073 U+0073) and 'ß' (U+00DF)); or the case of Umlaut in German: 'ö' (U+00F6) vs. 'oe' (U+0069 U+0065), etc.).

3. Normalization exceptions: These are the instances where normalization doesn't yield one single canonical form, therefore a careful analysis must be done for cases which yield same visual form using different sequences of code points. However, such cases may be limited by only permitting particular sequences in the repertoire.

[4] While such code points are excluded *a priori*, this is effectively based on Euro-centric conception, since essentially, all glyphs and graphemes may be used as letters. Accordingly, the Generation panel will ask for inclusion of code point to a future revision of MSR, where clear evidence is found of such code points, excluded because of their property as a non-letter, e.g. a mark.

Work Procedure

Similarly, to the construction of the Code Point Repertoire, the variant sets list will be populated using the Inclusion principle. That is, it starts with a blank table and only variant sets that meet the principles will be incorporated.

Phase 1: Cross-script analysis
In this phase, the Panel will analyze related scripts such as Cyrillic, Greek and potentially others to look for visual similar code points.

Phase 2: In-script analysis
During this phase, the Panel will analyze those code point in the Latin script (within the Scope) that may be subject to variant rules.