L2/22-108

Proposal to Add Data for Pairs of Identical sequences

Asmus Freytag

Proposed: **7 June 2022**

This document proposes the addition of eight pairs of identical sequences to the security (UTS#39) data files in analogy to the “intentional.txt” data collection for single code points.

These same sequence to sequence mappings are also absent from the collection of confusables.

Background

ICANN recently published the Root Zone Label Generation Rules, Version 5 (see https://icann.org/idn for contents and overview). These Rules, collectively known and RZ-LGR-5, present a dedicated analysis of characters and sequences that users can be expected to freely substitute in the context of domain names. Many of them have strictly identical appearance, and these are the source for the current proposal (there are many others where the resemblance, while strong, isn’t perfect, or where the substitution rests on semantic or phonetic, instead of visual grounds).

The scripts covered in RZ-LGR-5 nearly exhaust the set of “Recommended” scripts from UA#31, while the actual repertoire is limited to characters and sequences that are in common, wide-spread and everyday use. They represent high-priority targets for security exposure but also mitigation efforts. Their absence from the Unicode data collection represents thus a serious omission that should be remedied.

RZ-LGR-5 represents the culmination of a decade of effort by local volunteer panels composed of native speakers, linguists and technologists; ICANN staff; and a panel of technical experts combining linguistic, Unicode and IDN related expertise. At multiple stages throughout the process, public comments were solicited in a manner similar to the Unicode PRI process.

The rationale for all technical decisions is extensively documented, with each script LGR referencing the underlying proposal document, and, as the case may be, supportive data files.

There are other mappings (singleton to sequence, or singleton to singleton) that also have identical appearance. These were the focus of earlier submissions. This document focuses on the need to document property information, like identical appearance for cases where it cannot be expressed as a character property, but must truly be considered a property of strings.

None of these mappings have any single code point (in the RZ-LGR-5 repertoire) that shares the appearance in question. No data file can be constructed that takes a single code point as the left-hand value. This makes them distinct from the larger set of mappings that can be captured as a property of characters (and of which many, but not all, are already in the security (UTS#39) data files).

The following set of screen captures shows the sequences and their appearance at high resolution

Pairs of Identical Sequences from RZ-LGR-5

These pairs of sequences have an identical appearance that is impossible (not just difficult) to distinguish, even when magnified. (These are the complete set of such sequence pairs from RZ-LGR-5).

Source	Glyph	Target	Glyph
0905 0901	अँ	0972 0902	अँ
0906 0901	आँ	0911 0902	आँ
090D 0902	ऍं	090F 0901	ऍं
093E 0901	ाँ	0949 0902	ाँ
0BB6 0BCD 0BB0 0BC0	ஶ்	0BB8 0BCD 0BB0 0BC0	ஶ்
1004 103A 1039	၄်	105A 103A 1039	၄်
1015 102C 103A	ဝ်	101F 103A	ဝ်
17D2 178A	្ក	17D2 178F	្ក

There's an almost imperceptible shift of the dot in this example.