Proposal on Latin script for the root zone

Compiled by the Latin Generation Panel

The proposal

The proposal is out for public comment. Comments are welcome up to 2021-11-23.

Link to proposal:

https://www.icann.org/en/announcements/details/proposal-for-latin-script-root-zone-label-generation-rules-23-9-2021-en

Content of LatinGP proposal

- "Proposal for a Latin Script Root Zone LGR" -- main description (pdf)
  - 14 Appendices to the main description (pdf)
- LGR document (xml)
- HTML version of the LGR document (html)
- Test labels (txt)

This presentation

- This presentation is a walk-through of the proposal to lower the threshold for you to read and comment.
- Focus is on the main document and its appendices.
- The LGR document is the normative document.

Introduction and chapter 2

- Chapters not commented in this presentation are just short with general information.
- Chapter 2 defines the limitation of the scripts processed by this proposal:
  - The proposal cannot include any character not included in so called Maximum Starting Repertoire (MSR).
  - MSR is a subset of IDNA valid code points, which is a subset of Unicode. MSR is defined by the Integration Panel.
  - Only the Latin script subset of MSR is available for the Latin proposal.
    - A few characters were added to MSR on the LatinGP's request.

Chapter 4

- Describes the work process of LatinGP.
  - Languages using Latin script were identified and those on level 0 ("International") to 4 ("Educational") on the EGIDS scale were selected.
    - Level 4: "The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education."
  - Those on level 5 ("Developing") with at least 1 million speakers were also selected.
    - Level 5: "The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable."

Chapter 4 – continued

- Appendix B has a complete list of all languages selected using those criteria. For each language in that list the following information is listed:
  - Language name (in some cases the name in different languages)
  - The ISO 639-3 three-letter language code
  - The EGIDS value for the language
- Characters used by selected languages were identified.
  - The character set of each language is not documented in the report, but they can be found through the reference for each language found in chapter 9.
- Candidates for in-script and cross-script variants were identified (more on that below).

Chapter 5

- The repertoire of Latin script in the proposal.
  - The repertoire is based of the notion of Unicode unit "code point".
    - In the simple case, a code point is a character, such as "a".
    - A code point can also be a modifying mark used in combination with another code point to form a character, e.g. "g" + "~" → "g̃"
      - In many cases, Unicode has a precomposed code point in which the base character is combined with an accent mark. Such precomposed code points are always used when available.
  - The principles for including or not including a character identified in a language are spelled out as an introduction in this chapter.
  - The LatinGP's proposed repertoire lists 218 characters.
    - 197 characters of the simple type of one code point
    - 21 characters formed by a sequence of two or more code points

Chapter 5 – continued

- For each character there is the following information:
  - Unicode code point code or codes if it is a sequence.
  - Language or languages that use that character for writing
  - References for the alphabets of the languages using the characters
- The list of languages for a character is not exhaustive. The languages are there to support the inclusion.
- For characters a-z no language is listed.
- The repertoire is one of the main parts of the LGR document.
- The repertoire is here sorted by code point code. The same repertoire grouped by glyph shape is found in Appendix C.

Chapter 5 – continued

- Section 4 (5.4) lists excluded characters
  - The excluded characters are characters attested in at least one selected language which cannot be included because they do not belong to MSR.
    - The LGR procedure requires that only characters included in MSR can be selected.
    - The MSR is a result of a pre-process where characters are excluded due to one or several criteria, e.g. not protocol valid or similar to a punctuation mark.
    - LatinGP cannot include any character not included in MSR.

Chapter 6

- This chapter covers the concept and proposal of variants rules for Latin code points (characters)
  - A variant set consists of two or more characters that in some sense are perceived as being "the same".
    - Same – or almost – the same shape
    - Used interchangeably for the whole or part of the script community
  - Two types of disposition for variants: block or allocatable.
    - For the Latin Script Proposal, the majority of variant rules the variant labels are blocked.
  - In-script variants sets have members from the same script
  - Cross-script variant sets have members from different scripts
  - Some sets are a combination of the two

Chapter 6 – continued

- This chapter, together with appendices D.1 to D.9, contains
  - principles for variant sets
  - data and analyses of variant sets and candidate variant sets
- With two exceptions all variant rules are blocking "other variants"
- Two variant sets are special
  - Relates to older IDNA version 2003
  - Includes rules permitting allocating "other variant"
  - The two sets relate to
    - Sharp S ("ß") and "ss"
    - Dotted I ("i") and Dotless I ("ı")
- The LatinGP proposal of variant sets is presented in section 6.7

Appendix E

- The appendix contains candidate variant sets that were rejected as variant sets but accepted as "visually confusable".
  - The appendix is not part of the formal LGR.
  - The appendix is for reference for anybody doing analysis of visual similarity between two strings (TLDs or candidate TLDs).