ICANN 72

VIRTUAL ANNUAL GENERAL

# Root Zone Label Generation Rules (RZ-LGR) Update

ICANN72 Prep Week
14 October 2021

# Agenda

| | |
|---|---|
| ⊙ Latin Script RZ-LGR | Mats Dufberg |
| ⊙ Japanese Script RZ-LGR | Hirofumi Hotta<br>(to be Confirmed) |
| ⊙ Next version of Root-Zone LGR (RZ-LGR-5) | Asmus Freytag<br>(to be confirmed) |

# Latin Script Root Zone Label Generation Rules (Latin Script RZ-LGR)

Mats Dufberg
Latin Script GP Member

# Topics [colors]

1. The Latin GP proposal of Latin script for the root zone LGR
2. Introduction and chapter 2
3. Chapter 4
4. Chapter 5
5. Chapter 6
6. Appendix E

# The Proposal for Latin Script in Root Zone LGR

- The proposal for Latin script in root zone LGR presented by the Latin Generation Panel is out for public comment. Comments are welcome up to 2021-11-23.

Link to the proposal:

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

- Anyone is encouraged to review the proposal and provide comments and suggestions.
  - Both minor and major comments are welcome. All comments will be considered by the Latin GP.

# 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 XML file is the normative document.

# Introduction

- Chapters not discussed in this presentation are just short chapters with general information.
- Chapter 2 defines the delimitation of the scripts processed by this proposal:
  - The proposal cannot include any character not included in so called Maximal Starting Repertoire (MSR).
  - MSR is a subset of IDNA protocol 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 Latin GP's request.

# Chapter 4: Development Process and Methodology

- Chapter 4 describes the work process of the Latin GP.
  - Languages using Latin script were identified and those on level 0 ("International") to 4 ("Educational") on the EGIDS scale were selected.
  - Those on level 5 ("Developing") with at least 1 million speakers were also selected.

# Chapter 4 – continued

Definitions of the EGIDS scale levels can be found at https://www.ethnologue.com/about/language-status

**Table 1. Expanded Graded Intergenerational Disruption Scale**

| Level | Label | Description |
|---|---|---|
| 0 | International | The language is widely used between nations in trade, knowledge exchange, and international policy. |
| 1 | National | The language is used in education, work, mass media, and government at the national level. |
| 2 | Provincial | The language is used in education, work, mass media, and government within major administrative subdivisions of a nation. |
| 3 | Wider Communication | The language is used in work and mass media without official status to transcend language differences across a region. |
| 4 | Educational | The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education. |
| 5 | Developing | The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable. |

Level 6 and above are excluded here for clarity.

# 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 level 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: Repertoire

- The repertoire of Latin script in the proposal is based on Unicode "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, e.g. "á". 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 in the introduction to Chapter 5.

# Chapter 5 – continued

- The Latin GP'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
- 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 character
- 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.

# Chapter 5 – continued

- The repertoire is also one of the main parts of the LGR XML file.
- The repertoire in Chapter 5 is sorted numerically by code point code.
  - The same repertoire, grouped by glyph shape, is found in Appendix C.

# Chapter 5 – continued

- Section 4 in Chapter 5 (5.4) lists excluded characters
  - The excluded characters listed there are characters attested in at least one selected language, but which cannot be included because they do not belong to MSR.
    - The LGR procedure requires that only characters included in MSR 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.
    - Latin GP cannot include any character not included in MSR.

# Chapter 6: Variants

- Chapter 6 covers the concept and proposal of variant 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: blocked or allocatable.
    - For the Latin Script Proposal, in 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, e.g. Latin, Cyrillic and Greek.
  - Some sets are a combination of the two types

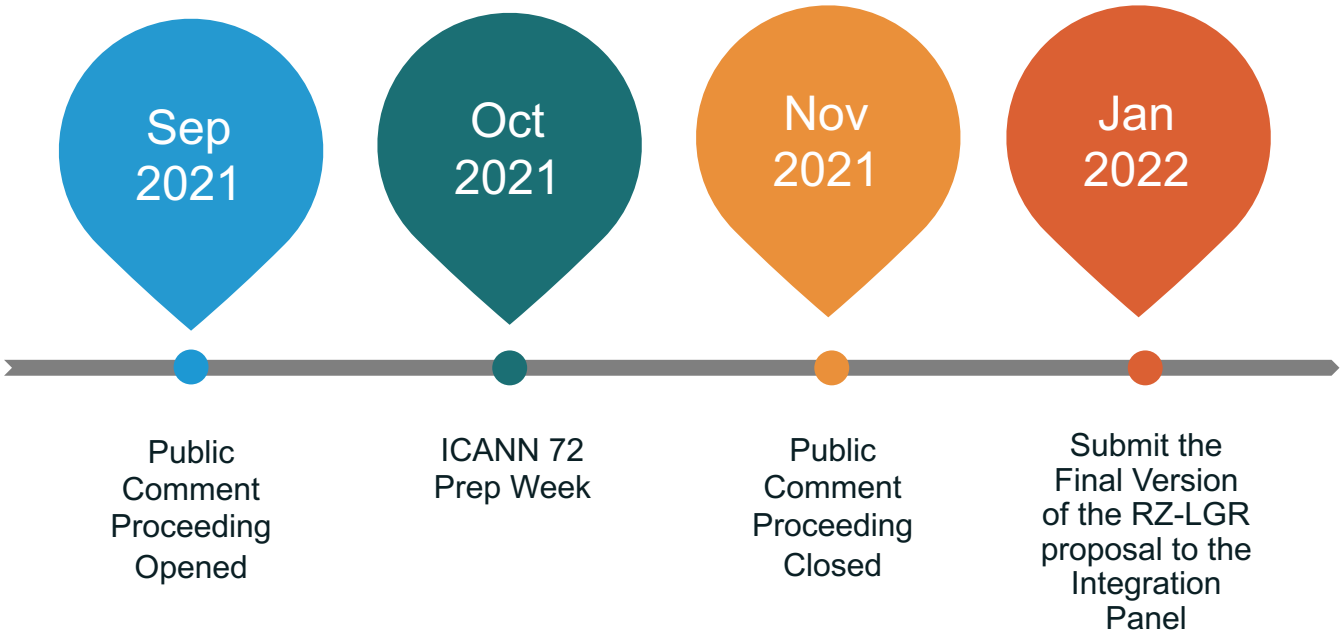
# Chapter 6 – continued

- Chapter 6, 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 Latin GP proposal of variant sets is presented in Section 6.7

# Appendix E: Confusables

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

# Current Step: Ongoing Public Comment Proceeding




Integration into the Root Zone LGR by the Integration Panel

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

Closing Date: 23 November 2021

# Engage with ICANN and IDN Program

**Thank You and Questions**

Visit us at **icann.org/idn**
Email: IDNProgram@icann.org

@icann
facebook.com/icannorg
youtube.com/icannnews
flickr.com/icann
linkedin/company/icann
slideshare/icannpresentations
soundcloud/icann