ICANN 72

VIRTUAL ANNUAL GENERAL

PREP WEEK

Root Zone Label Generation Rules (RZ-LGR) Update

ICANN72 Prep Week
14 October 2021

Agenda

⊙ RZ-LGR Overview	Pitinan Kooarmornpatana
⊙ Latin Script RZ-LGR Proposal	Mats Dufberg
⊙ Japanese Script RZ-LGR Proposal	Hiro Hotta
⊙ Next version of RZ-LGR	Michel Suignard

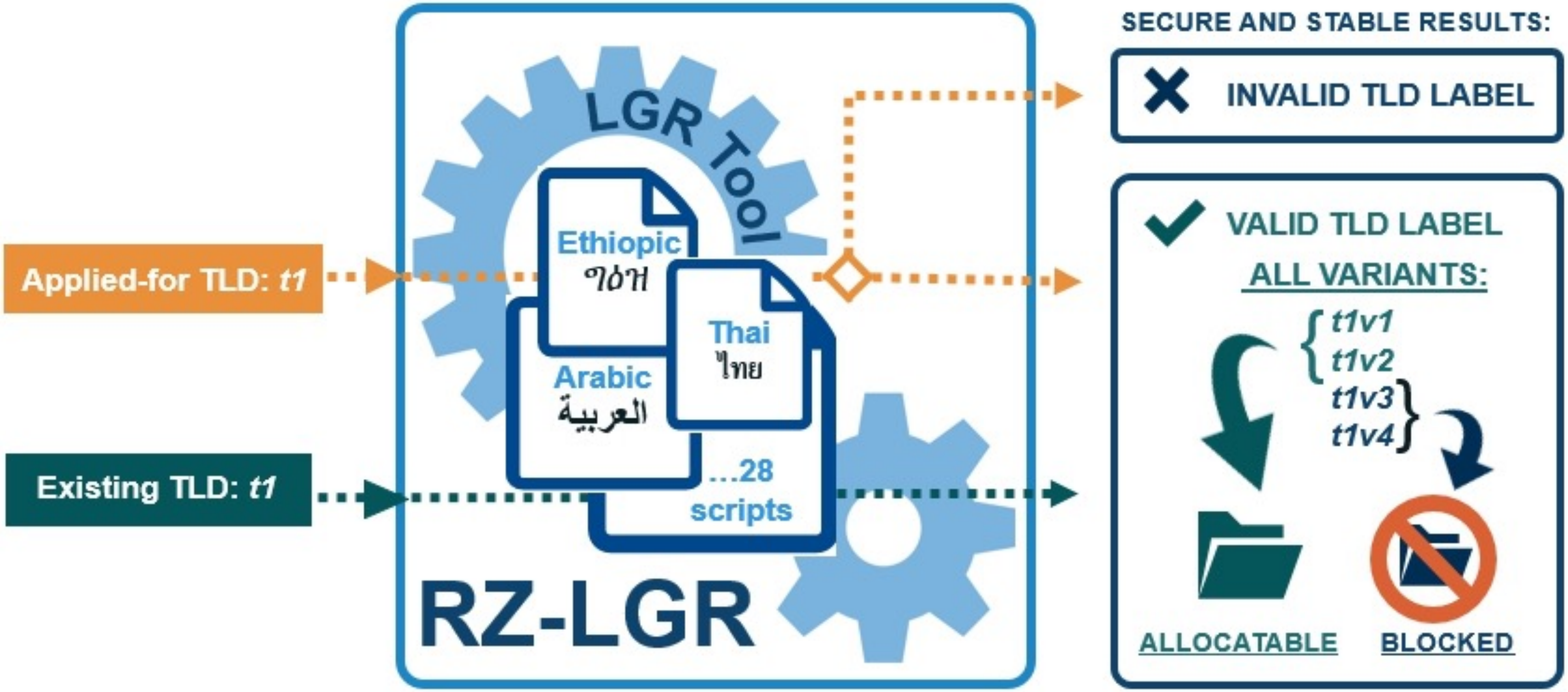
RZ-LGR Overview

Pitinan Kooarmornpatana
Senior Manager, IDN Program

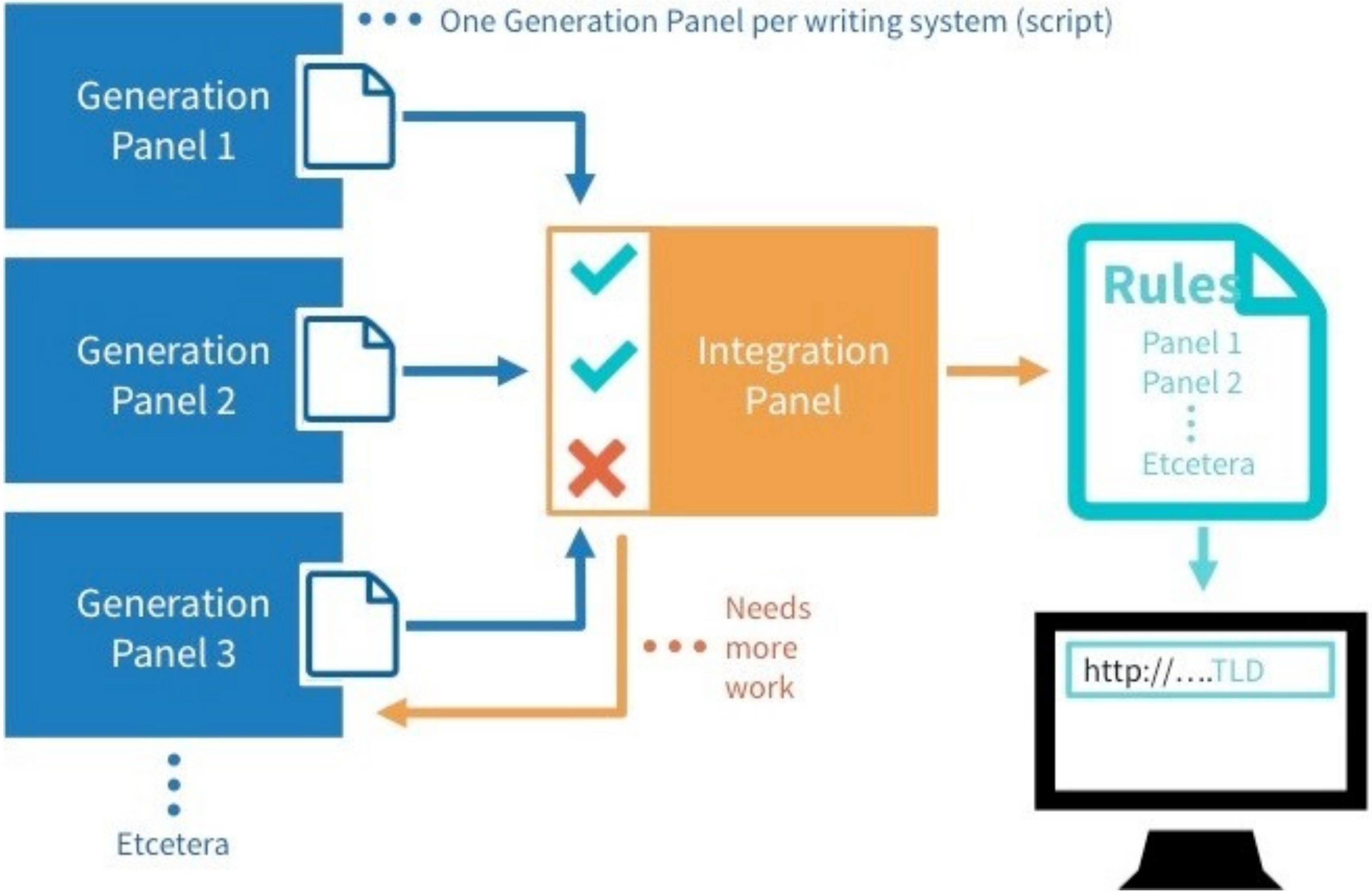
A Brief History of Root Zone Label Generation Rules (RZ-LGR)

- The ICANN community identified the need for variant top-level domains (TLDs).
- The Integrated Issues Report identified the need to define variant TLDs as a prerequisite.
- The community identified RZ-LGR as the mechanism to define variant TLDs and specified the LGR Procedure to develop RZ-LGR.
- In 2013, the LGR Procedure was approved by the ICANN Board for implementation for use with gTLDs and IDN ccTLDs.
- In 2019, the ICANN Board resolved that the GNSO and ccNSO take into account the Recommendations for Managing the IDN Variant TLDs - which integrated the use of RZ-LGR - in their policy development processes.
- In 2020, the ICANN Board resolved that the GNSO and ccNSO take into account the Recommendations for the Technical Utilization of the Root Zone Label Generation in their policy development processes.
- In 2021, the GNSO published its Report on New gTLD Subsequent Procedures which incorporates the use of RZ-LGR for the next round of new gTLDs.

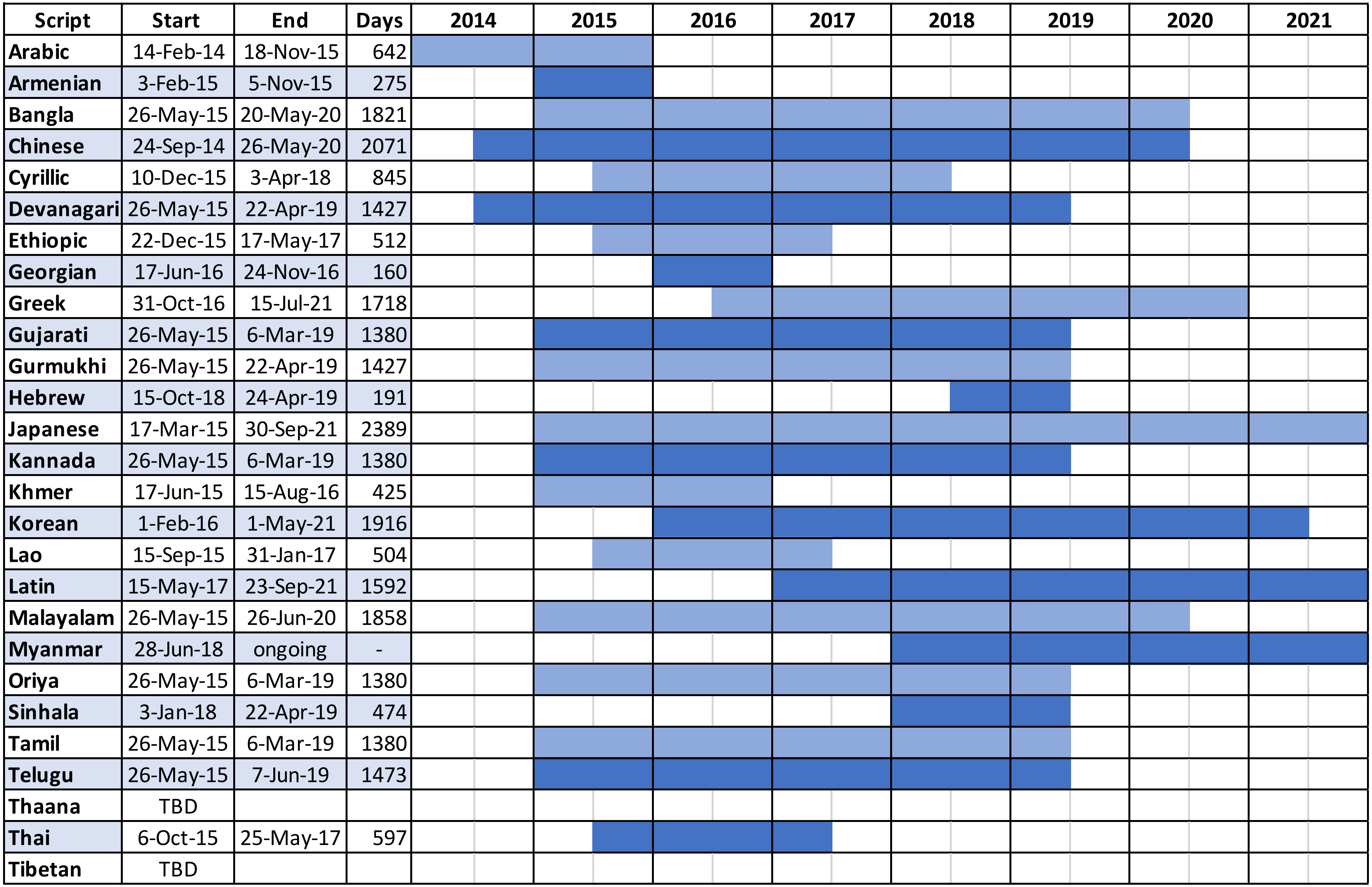
How Does RZ-LGR Work?



RZ-LGR Proposal Development Process

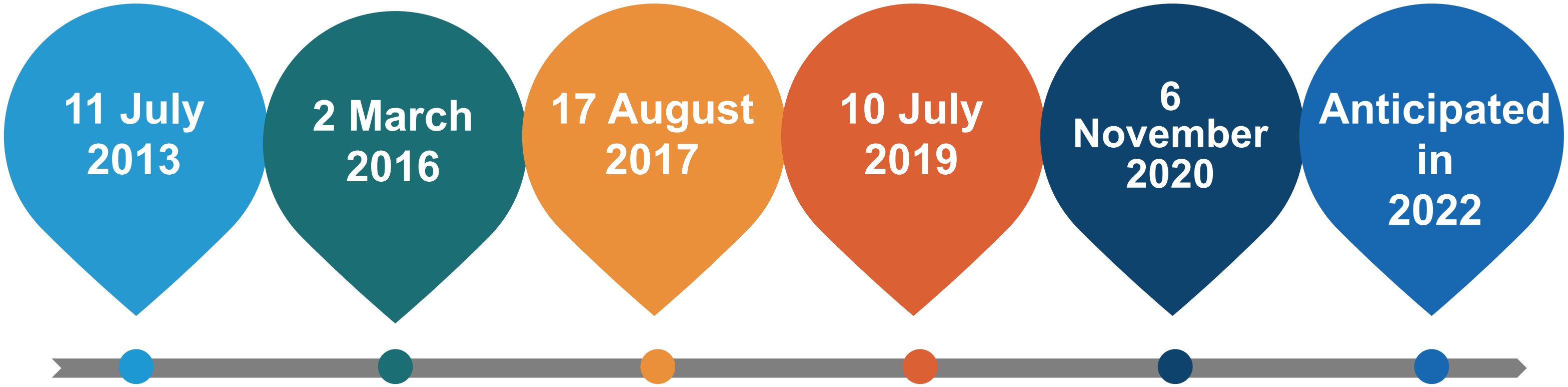


Summary of Generation Panel (GP) Work



Script	Start	End	Days	2014	2015	2016	2017	2018	2019	2020	2021
Arabic	14-Feb-14	18-Nov-15	642	■	■						
Armenian	3-Feb-15	5-Nov-15	275		■						
Bangla	26-May-15	20-May-20	1821		■	■	■	■	■	■	
Chinese	24-Sep-14	26-May-20	2071	■	■	■	■	■	■	■	
Cyrillic	10-Dec-15	3-Apr-18	845		■	■	■	■			
Devanagari	26-May-15	22-Apr-19	1427	■	■	■	■	■	■		
Ethiopic	22-Dec-15	17-May-17	512		■	■	■				
Georgian	17-Jun-16	24-Nov-16	160			■					
Greek	31-Oct-16	15-Jul-21	1718			■	■	■	■	■	
Gujarati	26-May-15	6-Mar-19	1380		■	■	■	■	■		
Gurmukhi	26-May-15	22-Apr-19	1427		■	■	■	■	■		
Hebrew	15-Oct-18	24-Apr-19	191					■	■		
Japanese	17-Mar-15	30-Sep-21	2389		■	■	■	■	■	■	■
Kannada	26-May-15	6-Mar-19	1380		■	■	■	■	■		
Khmer	17-Jun-15	15-Aug-16	425		■	■					
Korean	1-Feb-16	1-May-21	1916			■	■	■	■	■	■
Lao	15-Sep-15	31-Jan-17	504		■	■	■				
Latin	15-May-17	23-Sep-21	1592				■	■	■	■	■
Malayalam	26-May-15	26-Jun-20	1858		■	■	■	■	■	■	
Myanmar	28-Jun-18	ongoing	-					■	■	■	■
Oriya	26-May-15	6-Mar-19	1380		■	■	■	■	■		
Sinhala	3-Jan-18	22-Apr-19	474					■	■		
Tamil	26-May-15	6-Mar-19	1380		■	■	■	■	■		
Telugu	26-May-15	7-Jun-19	1473		■	■	■	■	■		
Thaana	TBD										
Thai	6-Oct-15	25-May-17	597		■	■	■				
Tibetan	TBD										

RZ-LGR Development Timeline



11 July 2013

Call for Generation Panels

2 March 2016

RZ-LGR-1

Arabic

17 August 2017

RZ-LGR-2

Ethiopic
Georgian
Khmer
Lao
Thai

10 July 2019

RZ-LGR-3

Devanagari
Gujarati
Gurmukhi
Hebrew
Kannada
Malayalam
Oriya
Sinhala
Tamil
Telugu

6 November 2020

RZ-LGR-4

Bangla
Chinese

Anticipated in 2022

RZ-LGR-5

Armenian
Cyrillic
Greek
Japanese
Korean
Latin
Myanmar

Next Steps

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

Mats Dufberg
Latin Script GP Member

Topics

1 Latin Script RZ-LGR Proposal

2 Introduction and Chapter 2

3 Chapter 4 – Development Process and Methodology

4 Chapter 5 - Repertoire

5 Chapter 6 - Variants

6 Appendix E

Latin Script RZ-LGR Proposal

- The Latin Script RZ-LGR proposal developed by the Latin Script Generation Panel (GP) is currently open for Public Comment. Comments will be accepted until 23 November 2021.
- Link to the proposal: https://www.icann.org/en/announcements/details/proposal-for-latin-script-root-zone-label-generation-rules-23-9-2021-en
- Everyone is encouraged to review the proposal and provide comments and suggestions.
    - Both minor and major comments are welcome. All input 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 the 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 the MSR at the Latin GP's request.

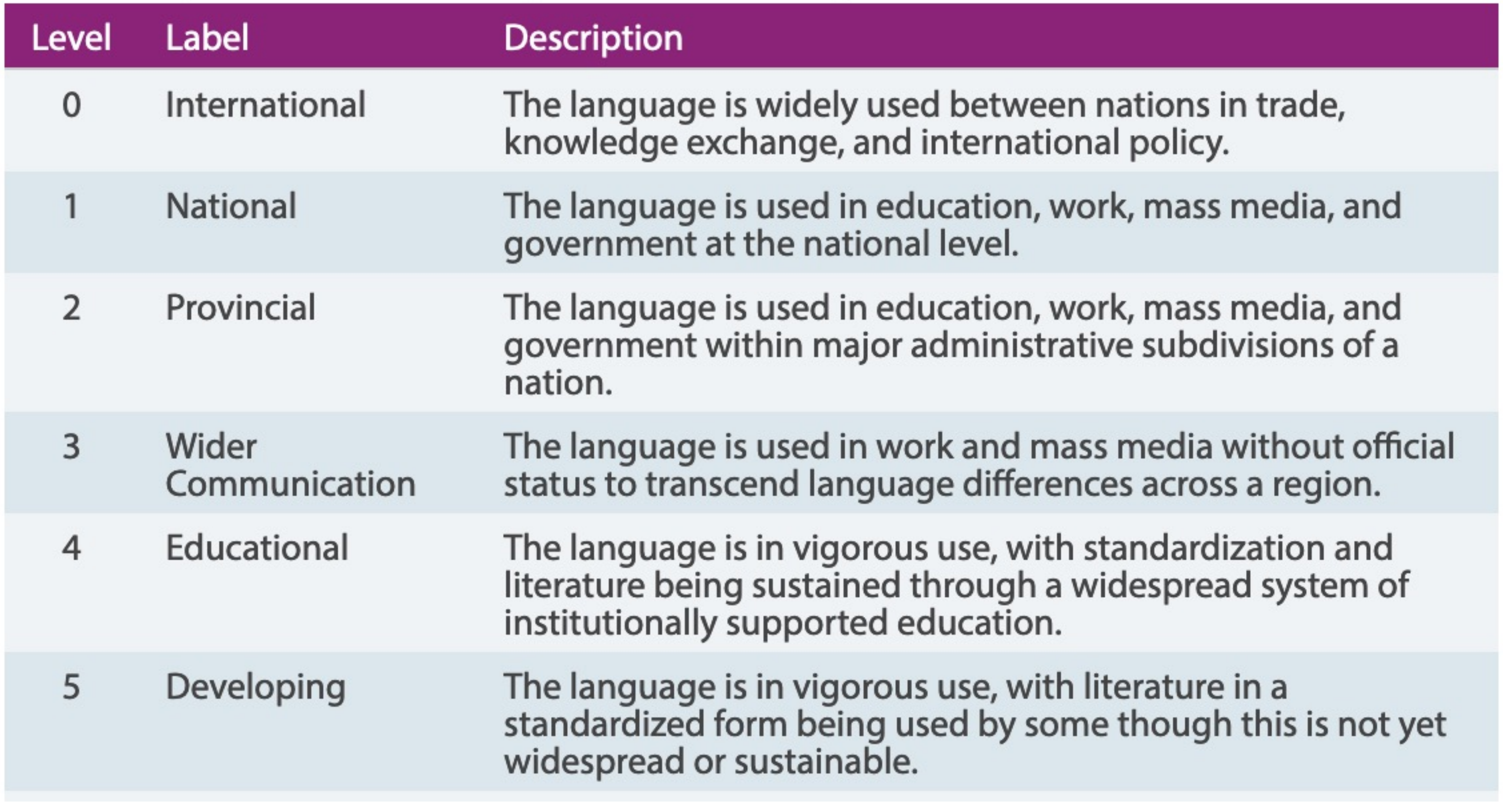
Chapter 4: Development Process and Methodology

- Chapter 4 describes the work process of the Latin GP.
    - Languages using the Latin script were identified and those from level 0 (International) to level 4 (Educational) on the EGIDS scale were selected.
    - Languages in level 5 (Developing) with at least one 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 the criteria. For each language included, 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 the Latin script in the proposal is based on Unicode code point.
  - In the simplest 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 are of one code point.
  - 21 characters are 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 are characters attested in at least one selected language, but which cannot be included because they do not belong to the MSR.
    - The LGR procedure requires that only characters included in the 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.
    - The Latin GP cannot include any character not in the 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.

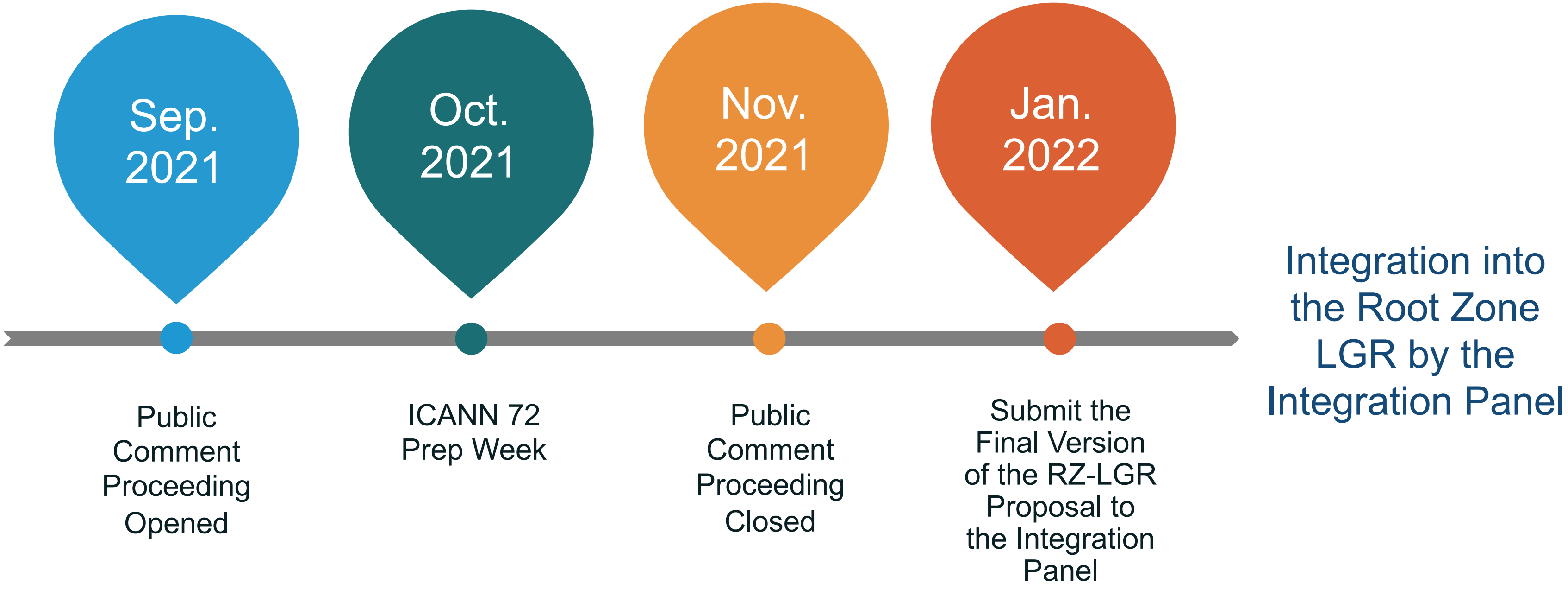
Chapter 6 – Continued

- 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



Sep. 2021

Public Comment Proceeding Opened

Oct. 2021

ICANN 72 Prep Week

Nov. 2021

Public Comment Proceeding Closed

Jan. 2022

Submit the Final Version of the RZ-LGR Proposal to the Integration Panel

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

Japanese Script Root Zone Label Generation Rules (Japanese Script RZ-LGR)

Hiro Hotta
Japanese Script GP Chair

Topics

1

Introduction to Japanese Script Generation Panel

2

Overview of the Japanese Script and Language

3

Overview of the Japanese RZ-LGR

4

Special Attentions

5

Current Step: Ongoing Public Comment Proceeding

Introduction to Japanese Script Generation Panel

- Hiro Hotta (chair)
  - Policy/business aspects of registry/registrar
- Akinori Maemura (vice chair)
  - Internet governance and domain name in general
- Shigeki Goto
  - Internet in general
- Kazunori Konishi
  - Internet in general
- Tsugizo Kubo
  - Trademarks and domain names
- Yoshitaka Murakami
  - Trademarks and gTLD markets from registry/registrar perspective
- Shuichi Tashiro
  - Character codes
- Yoshiro Yoneya
  - Technical aspects of IDN, LGR
- Yuri Takamatsu (secretary)
  - Policy/business aspects of registry/registrar

Overview of the Japanese Script and Language

- Script and Language
  - 3 scripts : Kanji, Hiragana, Katakana
    - Characters can be mingled in any order in a word
    - Characters defined in JIS (Japanese Industrial Standard) level-1 and level-2 are mostly used in daily life
    - 6,000+ characters
  - Kanji is also used in the Chinese and Korean languages.

- Variants
  - Basically, all Japanese characters are regarded as independent.
  - In Chinese and Korean languages, some sets of Kanji characters are regarded as variants when two or more characters have the same meaning and pronunciation.
  - Some people in the Japanese language community believe that some Chinese/Korean variants are regarded as variants in the Japanese language as well.

Overview of the Japanese RZ-LGR

- Repertoire
  - 6000+ characters in JIS level-1 and level-2
- Variants:
  - No variants stemming from the same meaning and pronunciation.
  - 10+ sets of variants stemming from visual identicalness.
  - Accommodating Kanji variants defined in Chinese LGR and Korean LGR.
  - No variant labels (other than original label) can be allocated.
- WLE: one Japanese specific rule
  - Any small kana, iteration mark, or prolonged mark must not start a label.
    - The same rule for ordinary Japanese words.

Special Attention 1 - Visual Identicalness

- Visual identicalness in Japanese scripts
  - Between one-stroke mark character and Kanji
    - ー and 一, ヽ and 丶
  - UNICODE Consortium lists confusable characters between different scripts in http://www.unicode.org/Public/security/latest/confusables.txt
    - Based on the above list, the following 10 pairs are picked up as candidates to be visually identical:

Hiragana	Katakana	Kanji
へ	ヘ	
べ	ベ	
ぺ	ペ	
	ニ	二
	ハ	八
	カ	力
	ト	卜
	ロ	口
	タ	夕
	エ	工

Words containing these characters

ヘリコプター
シャーペン
シャンハイ
インターネット
コンピューター

ショベル
コミュニケーション
ホッカイドウ
プロジェクト
ダイエット

Special Attention 1 - Visual Identicalness - Continued

- Confirmation of visual identicalness of UNICODE-based 10 pairs.
  - Field research to see if they all are visually identical.
    - a. Pairs of single confusable characters + pairs of confusable words.
    - b. 9 popular fonts with 3 font sizes.
    - c. 40 examined – among them, 20 read Japanese well, while 20 don't.
    - d. Every experiment (=every combination of a. b. c. d.) gives a rating from 1-5.
      - 1 (very similar), 2 (similar), 3 (neutral), 4 (distinct), 5 (very distinct)
    - Results
      - All pairs are rated less than 3.2.
        - » All pairs are visually identical enough to be confused.
  - Field survey to see if there are additional visually identical pairs.
    - Survey: "Were there any character pairs (other than those 10 pairs) that confused you because of visual identicalness?"
    - 73 responded to the survey out of 176 diverse recipients.
    - Results
      - No pairs confused more than 3% of the respondents.
        - » No pairs other than the 10 pairs are not confusingly similar.

Special Attention 2 – Allocatable Variants

- Basically, any combination of characters is allowed in Japanese labels as is the case of Japanese words used in daily life.

- The above may make the number of variant strings large and considering that the definitions of many variants are imported from the Chinese and Korean LGRs.
  - 慶応大学 has 3 variant strings – 慶應大学/慶応大學/慶應大學
  - Keio University registers and uses all 4 variant SLDs under .jp - 慶応大学.jp 慶應大学.jp 慶応大學.jp 慶應大學.jp
  - If Keio University is allowed to use all of them, 4 TLDs are allowed to be in the root zone simultaneously – such a rule may explode the size of the root zone when longer labels are considered.

- The reduction of the number of allocatable variant labels was required to prevent the explosion of root zone size.

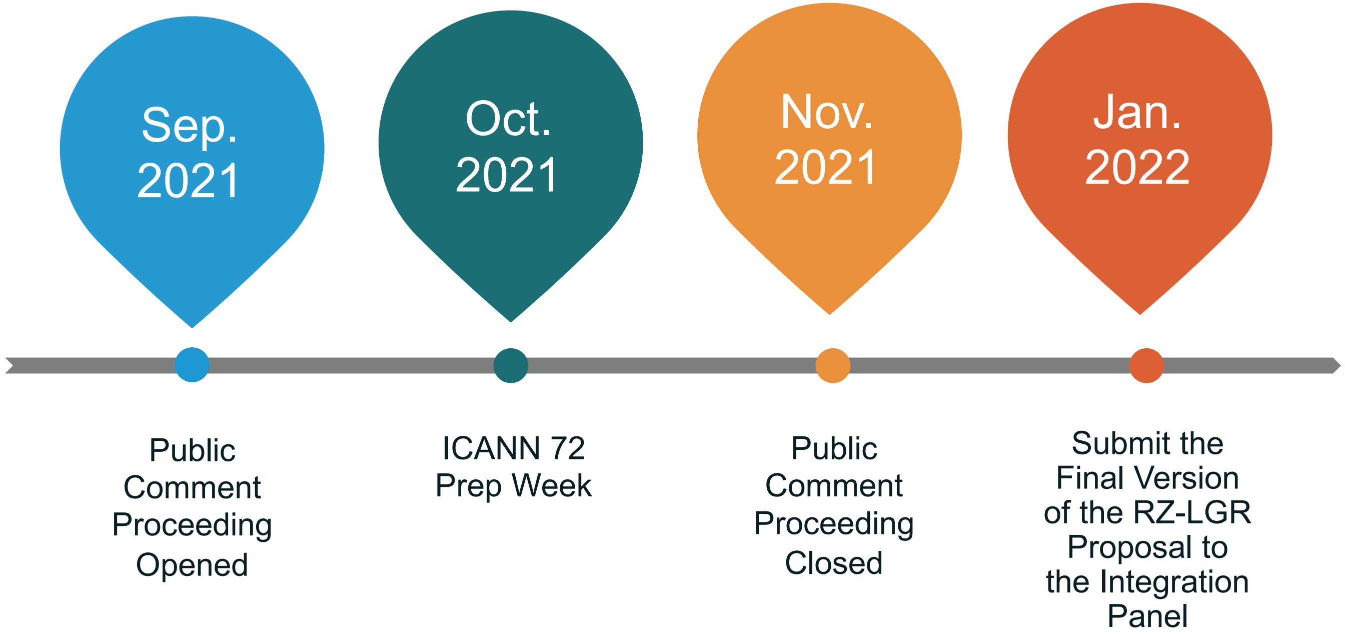
Special Attention 2 – Allocatable Variants - Continued

- Japanese GP devised various methods to reduce the number of allocatable labels.
  1. All variant labels can be allocated.
  2. Making variant labels containing only variants that are Joyo-Kanji (about 2,600 Kanji characters for everyday use) allocatable.
  3. In addition to the above, making variant labels containing only 3 or less characters that have Joyo-Kanji variants allocatable.
  4. Only allowing the applied-for label and blocking all variant labels.

  The number of allocatable labels reduces in the order of 1 – 4. Even by method-3, the theoretically possible number of labels that can be allocated can be as large as 27 – which is still regarded as a large number.

- Finally, the Japanese GP decided to allow valid applied-for label only and make all variant labels blocked.

Current Step: Ongoing Public Comment Proceeding


Sep. 2021
Oct. 2021
Nov. 2021
Jan. 2022
Public Comment Proceeding Opened
ICANN 72 Prep Week
Public Comment Proceeding Closed
Submit the Final Version of the RZ-LGR Proposal to the Integration Panel

Integration into the Root Zone LGR by the Integration Panel

Public Comment Proceeding:

https://www.icann.org/en/public-comment/proceeding/proposal-for-japanese-script-root-zone-label-generation-rules-30-09-2021

Closing Date: 16 November 2021

Root Zone LGR Version 5 (RZ-LGR-5)

Michel Suignard
Integration Panel

Topics

- Chinese, Japanese, and Korean Integration
- Steps for Processing a Label
- Armenian, Cyrillic, Greek, and Latin Scripts
- Scope of RZ-LGR-5

Integrating Chinese, Japanese, and Korean Scripts

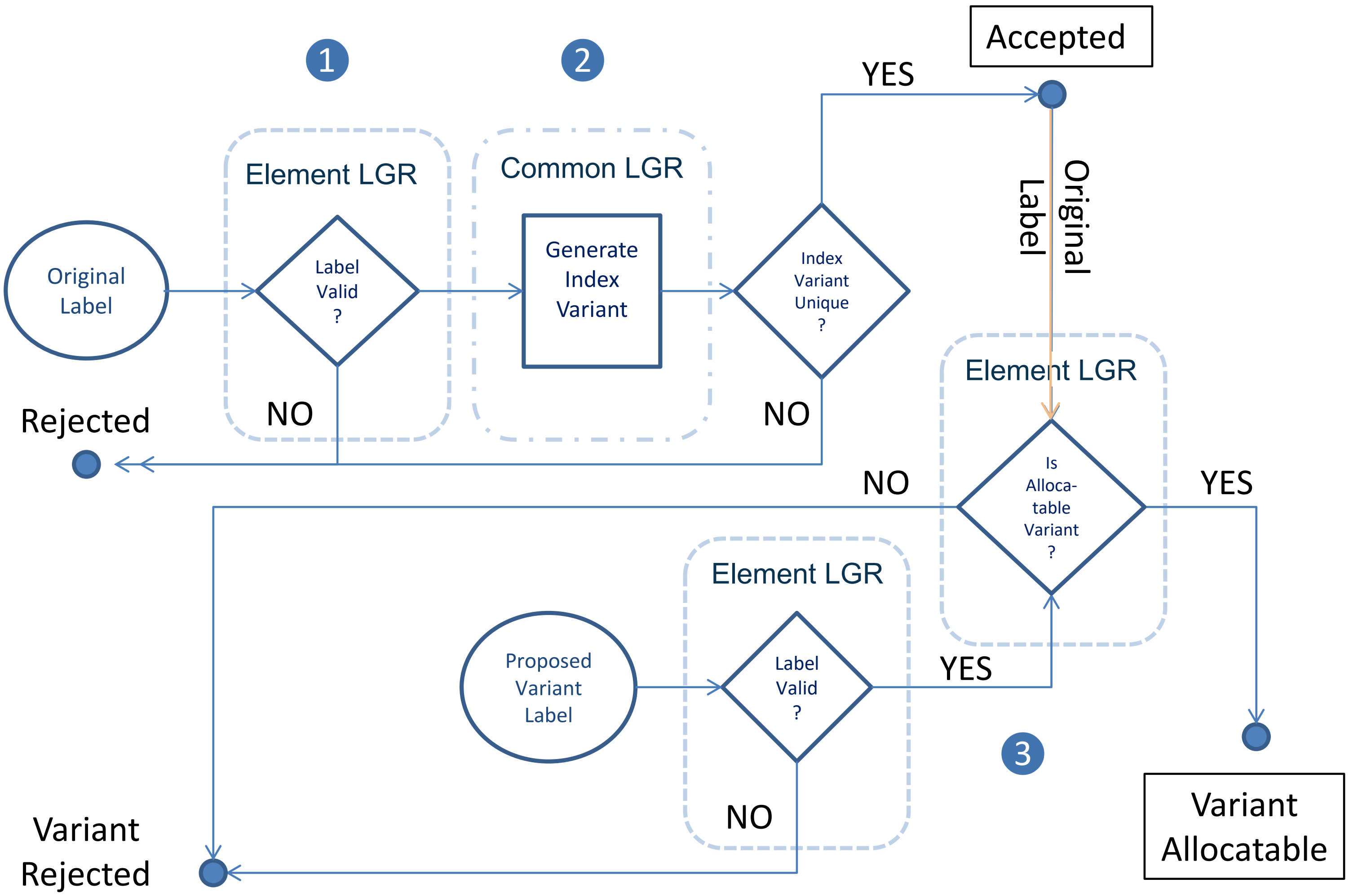
- Chinese, Japanese, and Korean scripts are now complete or open for Public Comment. Thank you, GPs!

- The Integration Panel has begun working on draft integration.
  - Chinese has some unlisted variants inherited from the Japanese and Korean LGRs.
  - Korean has many unlisted variants inherited from Chinese LGR.
  - Japanese has some unlisted variants inherited from Chinese LGR.

- Results expected:
  - Label collision will be determined via the merged LGR file, which has the complete set.
  - Allocatable labels (Chinese LGR only) will be determined via Chinese LGR.

- Waiting for Japanese LGR to complete Public Comment.

Steps for Processing a Label



Armenian, Cyrillic, Greek, and Latin Scripts

- Armenian, Cyrillic, Greek, and Latin scripts are now complete or open for Public Comment. Thank you, GPs!

- Armenian, Cyrillic, Greek, and Latin are a strongly interdependent system of scripts.
  - Variants inside that system fully listed in each of the four LGRs.
  - Other variants ("generic shapes“) only listed in Latin LGR.
  - Also inherited by the others; suppressed to cut on noise.

- Determining label collision requires use of merged (Common) file.

- Two LGRs were deferred to prevent incompatibilities from integration.
  - Deferred LGRs should not need to have updated proposals.
  - Any required mappings will be added as part of integration.

- The full integrated set will see public review as RZ-LGR-5.

RZ-LGR-5 Contents

- Eighteen (18) existing scripts from RZ LGR-4
- Completed or open for Public Comment:
  - Two (2) new C/J/K scripts: Japanese and Korean
  - Two (2) new alphabetic scripts: Latin and Greek
- Two (2) previously deferred alphabetic scripts: Cyrillic and Armenian
- One (1) additional script in progress: Myanmar
- **RZ-LGR-5 anticipated to contain 25 scripts.**
- Two (2) future scripts: Thaana and Tibetan

Engage with ICANN and IDN Program

ICANN

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