

# Progress on IDN Related Projects for the New gTLD Program: Next Round

**Sarmad Hussain**  
IDN Programs, Senior Director

**Pitinan Kooarmornpatana**  
IDN Programs, Director

ICANN 81  
13 November 2024



# Agenda

---

- IDN Implementation in the Next Round
- Root Zone Label Generation Rules (RZ-LGR)
- String Similarity Review
- Label Generation Rules (LGR) Tool Enhancement
- Q&A

# IDN Implementation in the Next Round

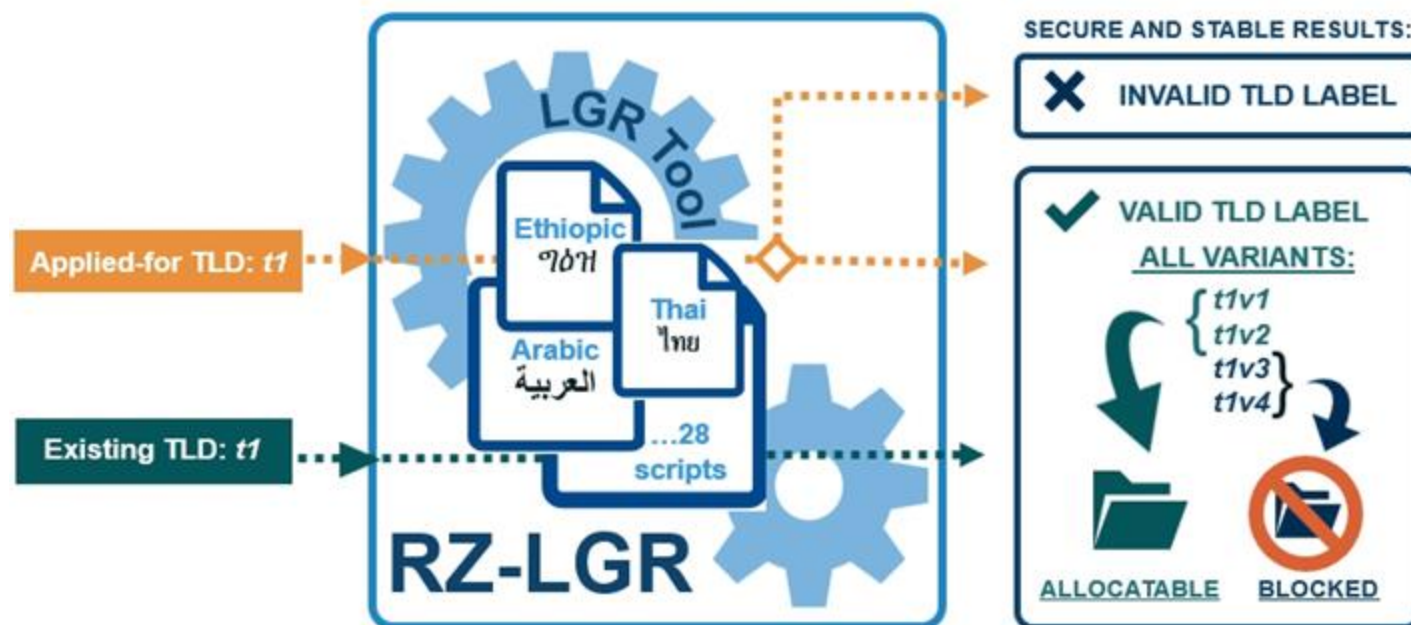
# IDN Related Policy for the New gTLD Program: Next Round

---

- Final Report on the new gTLD Subsequent Procedures Policy Development Process ([Sub Pro](#)), Topic 24: String Similarity, and Topic 25: IDNs, [adopted](#) by the ICANN Board in March 2023.
- Phase 1 Final Report on the Internationalized Domain Names (IDN) Expedited Policy Development Process ([IDN EPDP Phase 1](#))
  - ICANN Board [adopted](#) most recommendations based on the [scorecard](#) in June 2024.
  - Some pending recommendations based on the [scorecard](#) were [adopted](#) in September 2024.
- Phase 2 Final Report on the IDN EPDP ([IDN EPDP Phase 2](#)).
  - Being considered by the Generic Names Supporting Organization (GNSO) Council.

# Requirements for Top Level: Compliance with RZ-LGR

- Sub Pro Recommendation 25.2: Compliance with Root Zone Label Generation Rules ([RZ-LGR](#)) must be required for the generation of TLDs and variants labels, including the determination of whether the label is blocked or allocatable.
- IDN EPDP Phase 1 Final Recommendation 1.1: The RZ-LGR must be the sole source to calculate the variant labels and disposition values for all existing gTLDs.



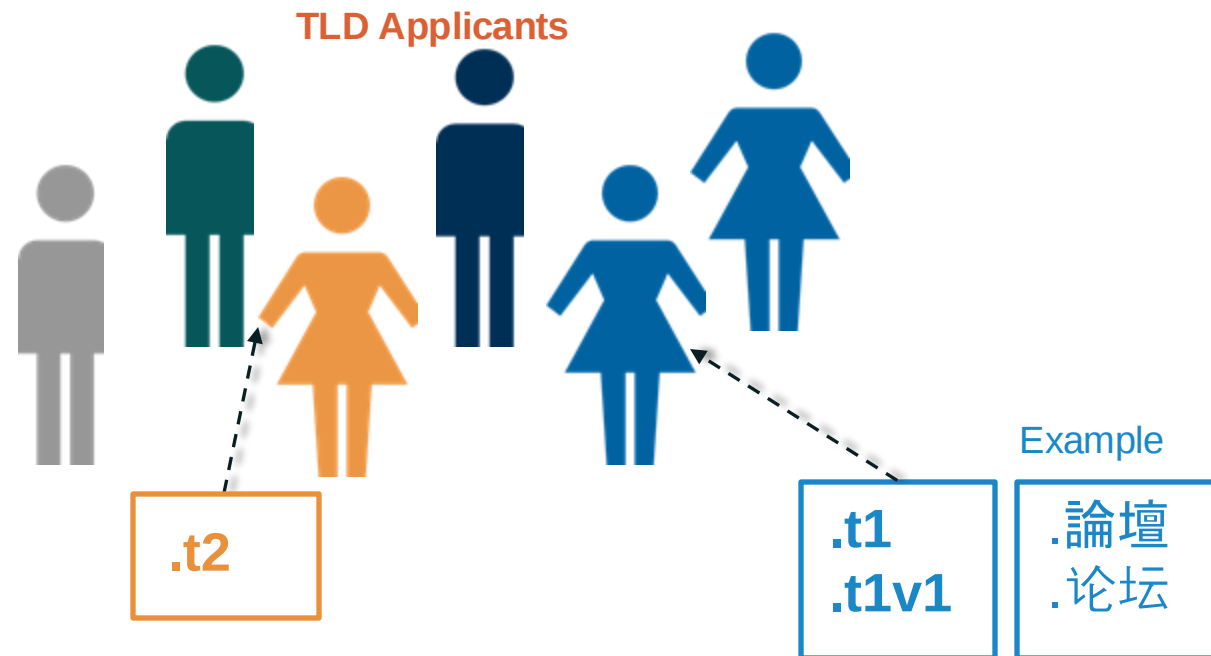
# Requirements for Top Level: Single Character TLDs

---

- Sub Pro Recommendation 25.4: Single character gTLDs may be allowed for limited script/language combinations where a character is an ideograph (or ideogram) and do not introduce confusion risks that rise above commonplace similarities, consistent with [SSAC](#) and Joint ccNSO-GNSO IDN Workgroup ([JIG](#)) reports.
- IDN EPDP Phase 1 Final Recommendation 3.17: At the time of the EPDP Team's deliberations, the only script that meets the criteria is the Han script, which is used in the Chinese, Japanese, and Korean languages. Nevertheless, applications for single-character gTLDs that are ideographs must not be accepted until relevant guidelines from the Chinese, Japanese, and Korean Generation Panels (CJK GPs) are developed, finalized after Public Comment, and implemented in the New gTLD Program...
  - The statements from the CJK GPs was published for [public comment](#).
  - ICANN org also received comments on the second set of draft Applicant Guidebook sections for [public comment](#), including the section for IDNs.
- Based on the feedback received, next steps are being discussed.

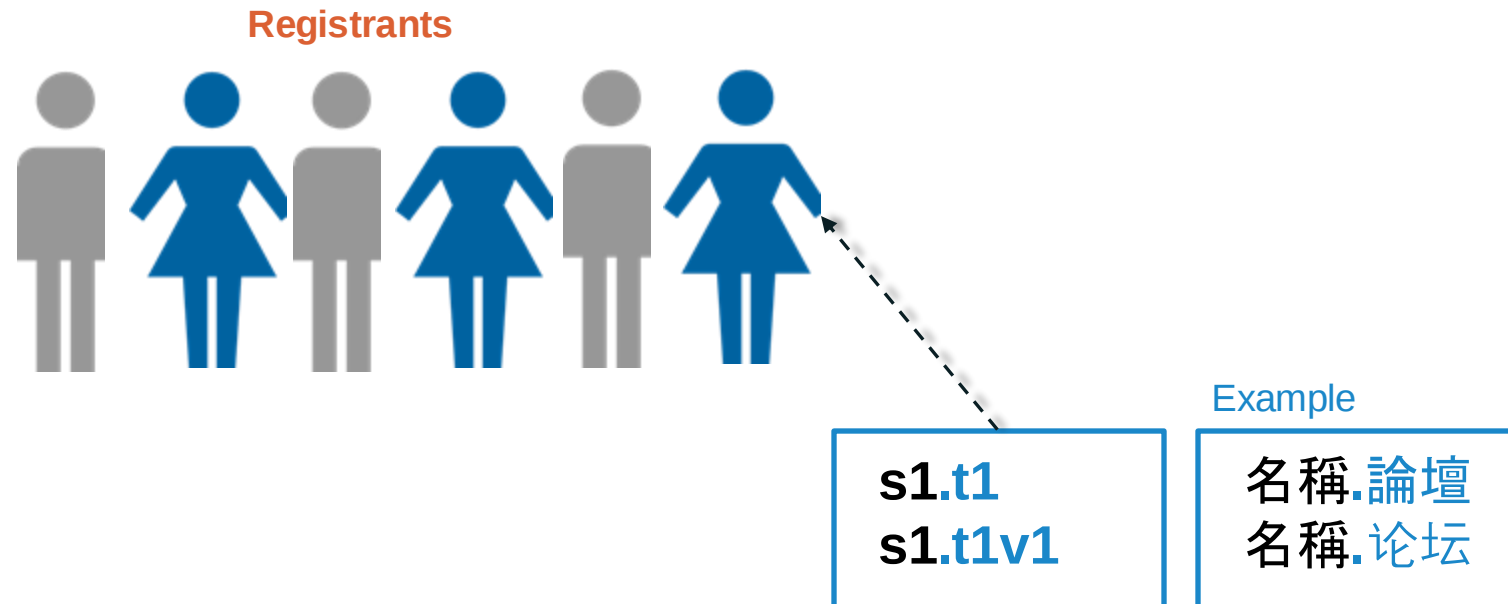
# Requirements for Top Level: Variant Management

- Sub Pro Recommendation 25.5: IDN gTLDs identified as variant TLDs of already existing or applied for gTLDs will be allowed only if labels are allocated to the same entity and, when delegated, only if they have the same back-end registry service provider.



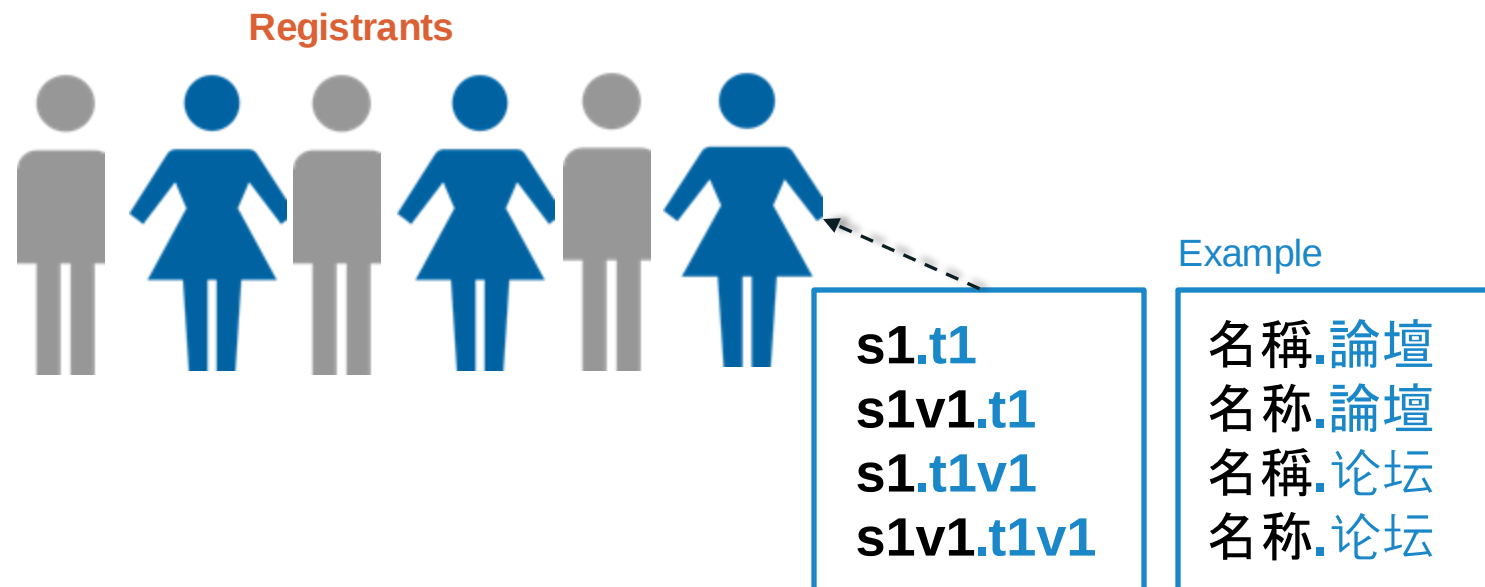
# Requirements for Second and Top Level: Variant Management

- Sub Pro Recommendation 25.6: A given second-level label under any allocated variant TLD must only be allocated to the same entity/registrant, or else withheld for possible allocation only to that entity (e.g., s1 under {t1, t1v1, ...}, e.g., s1.t1 and s1.t1v1).



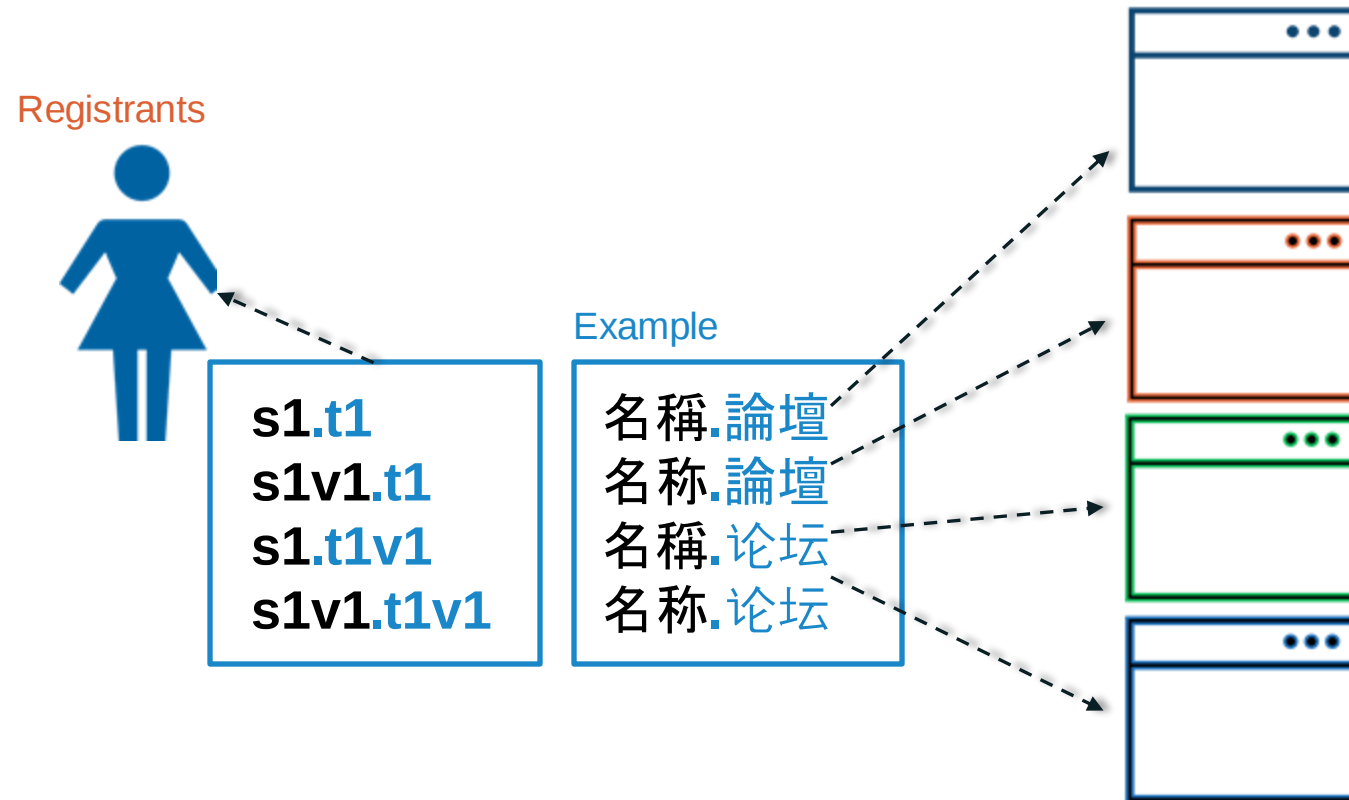
# Requirements for Second and Top Level: Variant Management

- Sub Pro Recommendation 25.7: For second-level variant labels that arise from a registration based on a second-level IDN table, all allocatable variant labels in the set must only be allocated to the same entity or withheld for possible allocation only to that entity (e.g., all allocatable second-level labels {s1, s1v1, ...} under all allocated variant TLD labels {t1, t1v1, ...}).



# Requirements for Second and Top Level: Variant Management

- Sub Pro Recommendation 25.8: Second-level labels derived from Recommendation 25.6 or Recommendation 25.7 are not required to act, behave, or be perceived as identical.



# Some Relevant Application Guidebook Modules

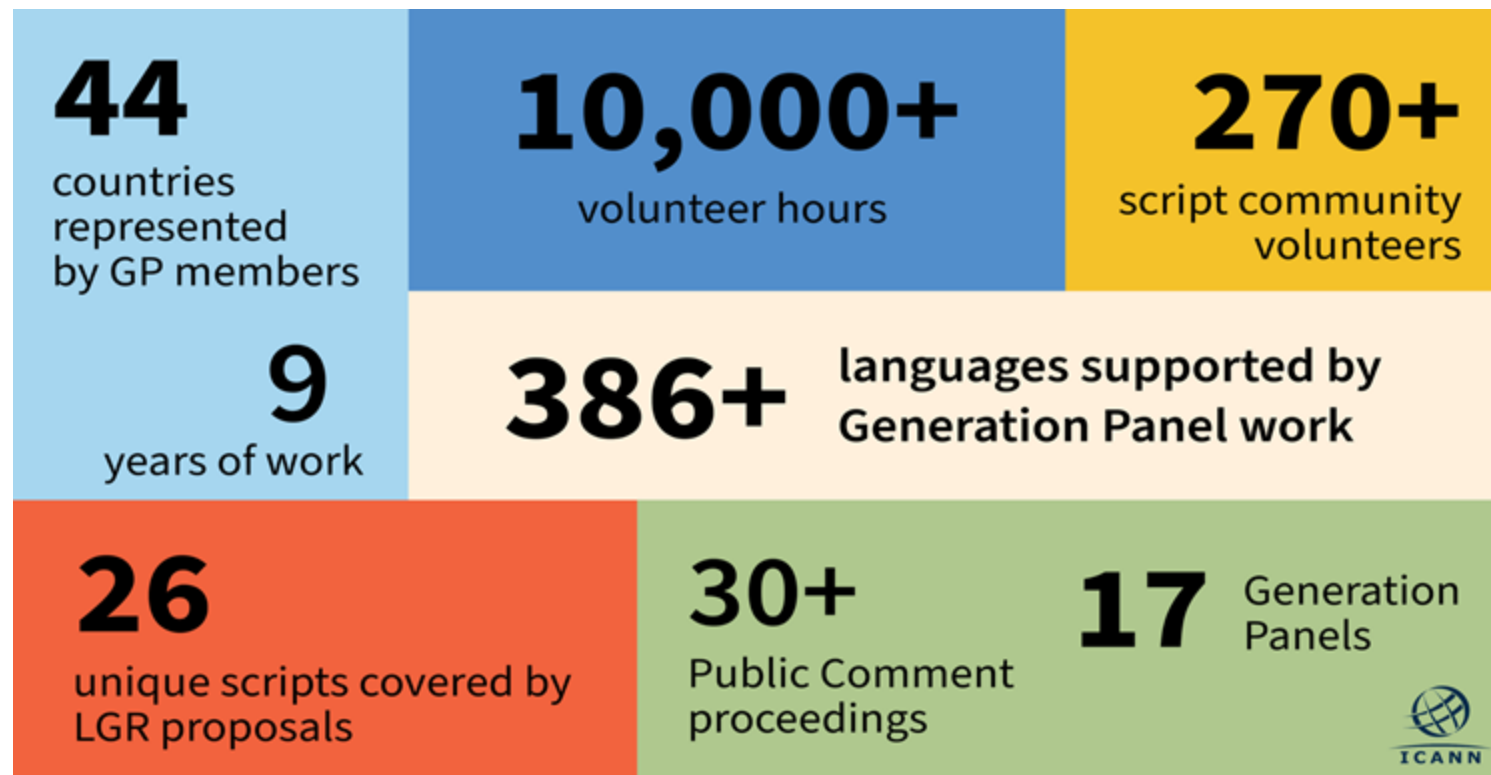
---

- Following draft modules were published for [public comment](#) review.
  - String Similarity Review (Topic 24: String Similarity Evaluations)
  - Internationalized Domain Names (Topic 25: IDNs)
  - Root Zone Label Generation Rules (Topic 25: IDNs)
- ICANN org is reviewing community inputs and addressing them.

# Root Zone Label Generation Rules (RZ-LGR)

# RZ-LGR Version 5 (Current Version)

- RZ-LGR version 5 published May 2022 include 26 scripts: Arabic, Armenian, Bangla, Chinese (Han), Cyrillic, Devanagari, Ethiopic, Georgian, Greek, Gujarati, Gurmukhi, Hebrew, Japanese (Hiragana, Katakana, and Kanji [Han]), Kannada, Khmer, Korean (Hangul and Hanja [Han]), Lao, Latin, Malayalam, Myanmar, Oriya, Sinhala, Tamil, Telugu, and Thai.



# Upcoming Update RZ-LGR Version 6

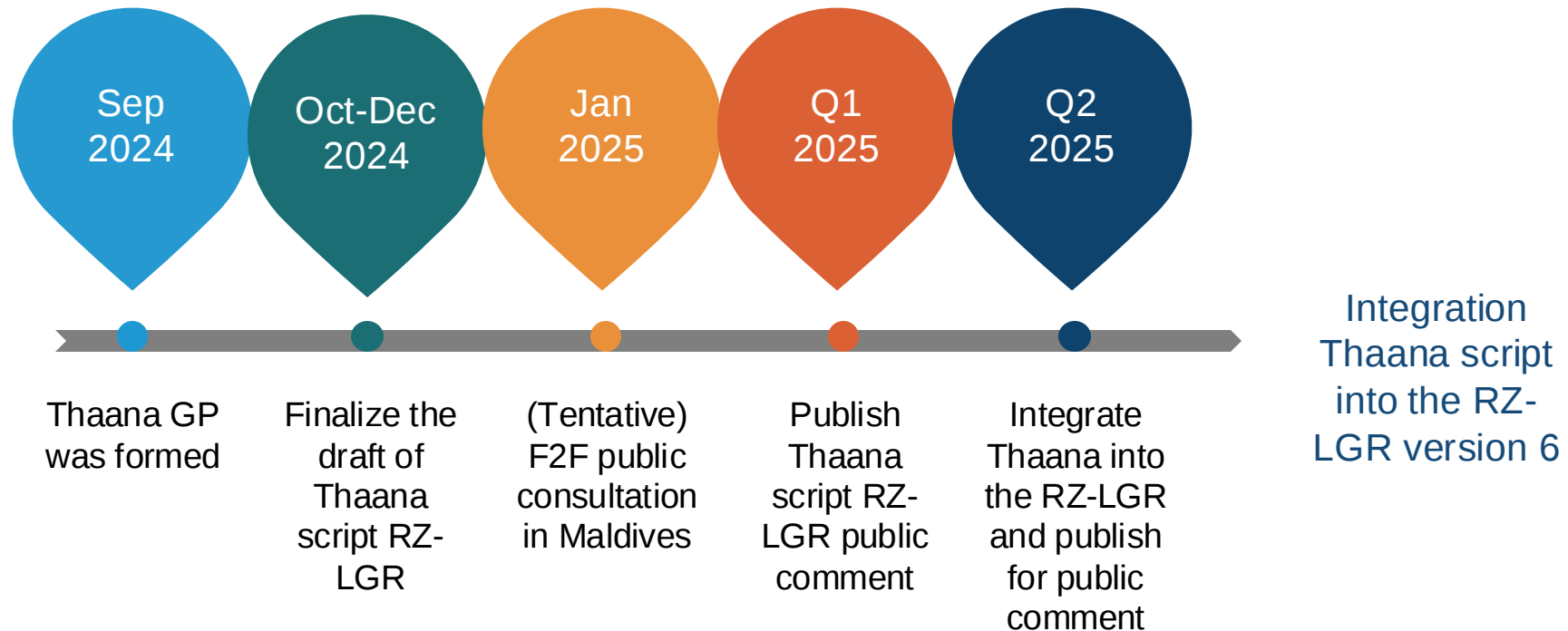
- Thaana script Generation Panel (GP) [formed](#) September 2024.
- Script information:
  - ISO 15924 Code: Thaa, ISO 15924 Key N°: 170, ISO 15924 English Name: Thaana, Latin transliteration of native script name: Thaana, Native name of the script: **سَرَوُ**,
  - Geography: South Asian group of 22 atolls, located south-west of India.
  - Thaana has 25 consonants and 11 vowels. Their inclusion in RZ-LGR is being discussed by the GP.

|        | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| U+078x | ހ | ށ | ނ | ރ | ބ | ޅ | ކ | އ | ވ | މ | ފ | ދ | ތ | ލ | ގ | ޏ |
| U+079x | ސ | ޑ | ޒ | ޛ | ޜ | ޝ | ޞ | ޟ | ޠ | ޡ | ޢ | ޣ | ޤ | ޥ | ަ | ާ |
| U+07Ax | ި | ީ | ު | ޫ | ެ | ޭ | ޮ | ޯ | ް | ޱ | ޲ | ޳ | ޴ | ޵ | ޶ | ޷ |
| U+07Bx | ޸ | ޹ |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

■ consonants
 ■ vowels
 ■ extended consonants (excluded)

# Thaana RZ-LGR Timeline

- RZ-LGR will be developed based on the LGR Procedure and the information available in the Reference LGR for the Second-Level
- The Thaana Script RZ-LGR is planned to publish for public comment in Q1 2025



# String Similarity Review

- Based on the [Sub Pro](#) Affirmation 24.1 affirms that “Strings must not be confusingly similar to an existing top-level domain or a Reserved Name” and Affirmation 24.2 affirms the visual standard for determining similarity
- Based on the [IDN EPDP Phase 1](#)
  - Final Recommendation 4.1 identifies the scope of string comparisons (details in the next slide).
  - Final Recommendation 4.2 states that decision by the String Similarity Review Panel.
  - Final Recommendation 4.3 states that the scope of omission by String Similarity Review Panel based on guidelines.

# Overview

- IDN EPDP Phase 1 Final Recommendation 4.1: an applied-for primary gTLD string (no matter whether it is an ASCII string or an IDN string) and all of its allocatable variant label(s) must be compared against the scope shown in the table.

| Categories for Comparison:                                                                                                                                                                                                                                                                                                   |                                   | The applied-for gTLD string |                                   |                               |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|-----------------------------|-----------------------------------|-------------------------------|
|                                                                                                                                                                                                                                                                                                                              |                                   | Primary gTLD string         | All allocatable variant string(s) | All blocked variant string(s) |
| <ul style="list-style-type: none"><li>● Existing gTLD</li><li>● The gTLD string applied-for in the previous round(s) still in the process</li><li>● Existing ccTLD</li><li>● Requested IDN ccTLD</li><li>● Other Applied-for gTLD</li><li>● Reserved Name</li><li>● Blocked Name</li><li>● Any two-Character ASCII</li></ul> | Primary String                    | Yes                         | Yes                               | Yes*                          |
|                                                                                                                                                                                                                                                                                                                              | All allocatable variant string(s) | Yes                         | Yes                               | Yes*                          |
|                                                                                                                                                                                                                                                                                                                              | All blocked variant string(s)     | Yes*                        | Yes*                              | No                            |

- [Initial guidelines](#) for conducting a String Similarity Review published for [public comment](#).
- Due to the enhanced scope of String Similarity Review to cover variant strings, a two-stage process is being considered.
  - Data-driven pre-screening.
    - Similarity data collection from script experts
    - String Similarity Review Tool
  - Manual review conducted by an expert panel, using the input from the pre-screening.

- Overview

- RZ-LGR Version 5 covers 26 scripts
- All repertoires of all scripts need to get analysed
- Delegate tasks for each script to experts

- Scope of Work

- ASCII, compare to ASCII characters, small and capital (a-z, A-Z)
- In-Script, compare all elements to each other
- Cross-Script, compare to elements in repertoire of other scripts within the same family
- Miscellaneous, Capitalised elements within script, Conjuncts or compounded characters, Simple shapes (e.g. circles “o”, lines “l”, curves “c”, “s”, “∩”), Underlined versions

- Categories

1. Identical or Near Identical (homograph) and already defined variant
2. Highly Confusable (strongly similar, or near homograph)
3. Similar
4. Distantly Similar (weakly similar)
5. Distinct (not similar)

## Category 1: Identical or Near Identical + Variants

- Variants:  
All existing (already defined) variants automatically belong to this Category, independent of their visual appearance
- Identical:  
Cases of identical characters
  - Most should already have been defined variants
  - Uppercase forms, e.g., e (0065) vs. ε (03B5), E (0045) vs. E (0395)
- Near Identical:  
Easy to confuse even side-by-side
  - Underlining, e.g., e (0065) vs. ẹ (1EB9), e vs. ẹ (using Verdana Font)

## Category 2: Highly Confusable

- Confused by a large majority of attentive users, when not looking at cases side-by-side
  - è (00E8) vs é (00E9) → example.coupè vs. example.coupé
  - ९ (1027) vs. ९ (0D47)
  - ষ (0998) vs. ষ (09AF)

## Category 3: Similar

- Confused by a simple majority of attentive users, when not looking at cases side-by-side
  - ř (0072 0303) vs. ř (0159)
  - ѡ (1223) vs. ѡ (057A)
  - ऋ (091E) vs. ऋ (0924 094D 0930)
- Fallback Replacements (mainly Latin)
  - è (00E8) vs. e (0065)

## Category 4: Distantly Similar

- Confused by a minority of attentive users, when not looking at cases side-by-side; or similar, but not really confusable
- Experts are not actively looking for all of these cases, but note them on corner cases
  - ì (00EC) vs. ï (03CA)
  - í (0453) vs. ĩ (0491)
  - ை (0918 094D 0918) vs. ௉ (0918 094D 0927)

## Category 5: Distinct

- These will be the vast majority of all pairs. There is no need to look at them.
  - a (0061) vs. ઘઘ (0918 094D 0918)

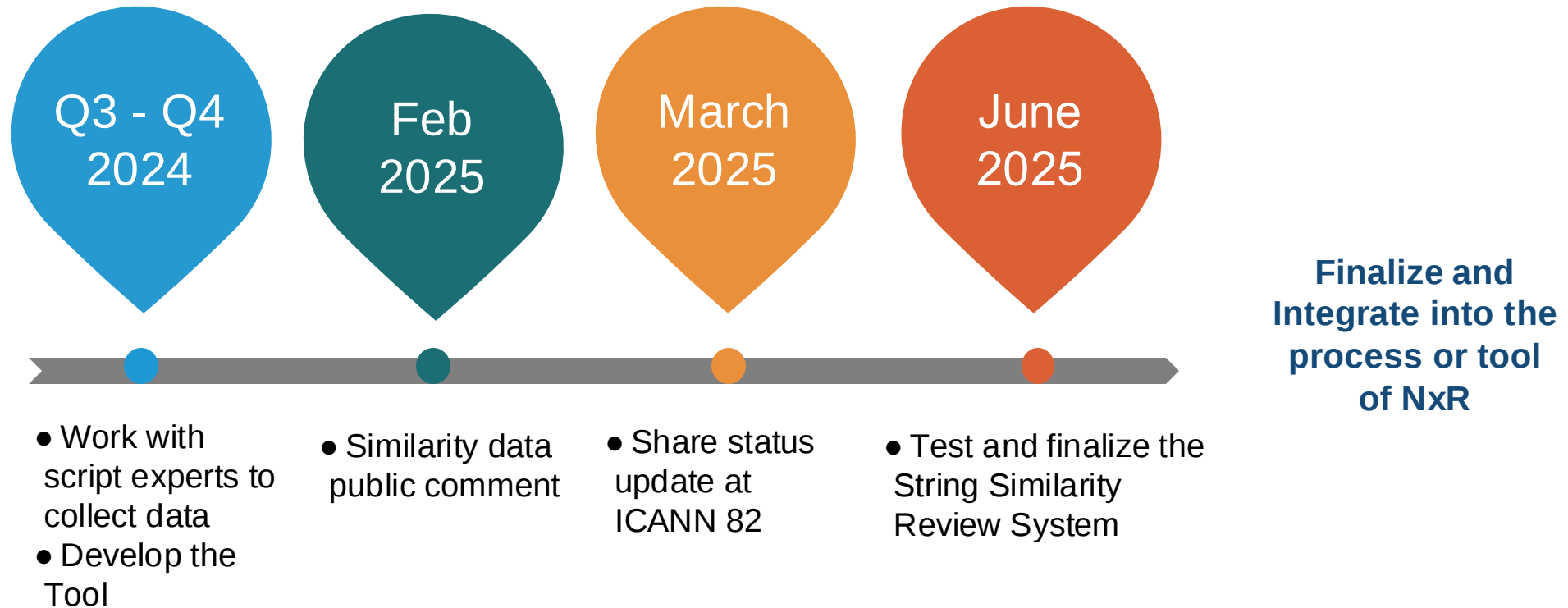
- Expert is working on initial data 8 Scripts
- Expert submitted data; internal review in progress 16 Scripts
- Data ready 2 Scripts

# String Similarity Review Tool

---

- Input
  - List of existing TLDs, Reserved Names, and Blocked Names
  - List of applied-for strings
  - Similarity data collection
- Process
  - Based on the similarity data collection, calculate if any applied-for strings, existing TLDs, Reserved Names, and Blocked Names, are same or similar to each other or their variant labels.
- Output
  - Report of the potential contention set for String Similarity Review Panel

# Timeline of String Similarity Review Data and Tool



# Label Generation Rules (LGR) Tool Enhancement

# Background

---

- The online LGR Tool is available in three modes
  - **Basic mode:** for validating label(s) against a [reference LGR](#) or a script [Root Zone LGR](#)
  - **Review IDN Table mode:** for reviewing an IDN Table with [reference LGR](#), a script [Root Zone LGR](#), or a customized uploaded LGR
  - **Advance mode:** for script expert to develop, modify and test an LGR
- Online deployment at: <https://lgrtool.icann.org/> accessible for all users with an ICANN account
- The Label Generation Rules Tool is an open source application. Source code on github: [lgr-core](#), [lgr-django](#), [munidata](#), [picu](#).
- User guides are available at <https://www.icann.org/resources/pages/lgr-toolset-2015-06-21-en>

- Following functions will be added to support the New gTLD Program: Next Round
  - **Verify a primary label**
    - Does the ASCII version (ASCII Label or a-label) of the string complies with protocol requirements?
    - Is the string as per IDNA2008?
    - Is the string as per RZ-LGR?
    - Does the ASCII version of the string comply with the label length requirement? ( $2 < \text{string} < 64$ )
    - If u-label is in Han script, does it meet length requirement.
    - If u-label is not in Han script, does it meet length requirement.
    - Does the label collide with any existing TLD variant sets?
    - Does the label collide with any Reserved Names
    - Does the label collide with any variant of a Reserved Names
    - Does the label collide with any Blocked Names variant set?
  - **Verify a variant label**
    - Perform the checks as listed above
    - Is the string an allocatable variant label of the primary label as per RZ-LGR?

# Thank You

---

Q&A