



83

POLICY
FORUM



String Similarity Evaluation in the New gTLD Program: Next Round

Sarmad Hussain, Senior Director, IDN-UA Program

ICANN 83
11 June 2025



Agenda

- Overview of String Similarity evaluation
- Update on the String Similarity data collection
- Update on the String Similarity evaluation pre-screening tool
- Q&A

- Based on the [Sub Pro](#)
 - Affirmation 24.1 states that strings must not be confusingly similar to an existing top-level domain or a Reserved Name.
 - 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 the String Similarity Review Panel may decide whether and what blocked variant labels to omit when conducting a comparison.
 - 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 ASCII or IDN gTLD 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)
- Existing gTLD - The gTLD string applied-for in the previous round(s) still in the process - Existing ccTLD - Requested IDN ccTLD - Other Applied-for gTLD - Blocked Name - Any two-Character ASCII	Primary String	Yes	Yes	Yes*
	All allocatable variant string(s)	Yes	Yes	Yes*
	All blocked variant string(s)	Yes*	Yes*	No 

*String Similarity Review Panel may decide whether and what blocked variant labels to omit when conducting a comparison.

- [Initial guidelines](#) for conducting a String Similarity Review published for [Public Comment](#).
- Due to the enhanced scope of String Similarity Evaluation to cover variant strings, a two-stage process is being considered:
 - Stage 1: Data-driven pre-screening
 - Similarity data collection from script experts
 - String Similarity Evaluation pre-screening tool
 - Stage 2: Manual review conducted by an expert panel using input from the pre-screening tool

Update on String Similarity Data Collection

- Overview
 - Root Zone Label Generation Rules (RZ-LGR) Version 5 covers 26 scripts (soon 27 with Thaana script to be added in version 6)
 - Repertoires of all scripts need to be analyzed
 - Task for each script done by relevant script expert
- Scope of Work
 - ASCII, compare to ASCII characters, small and capital (a-z, A-Z)
 - Within-script, compare all elements to each other
 - Cross-script, compare to elements in repertoire of other scripts within the same family
 - Miscellaneous:
 - Capitalized elements within script
 - Conjuncts or compounded characters
 - Simple shapes (e.g., circles “o”, lines “l”, curves “c”, “s”, “n”)
 - Underlined versions

Similarity Categorization

- Categories

1. Identical or Near Identical (homograph) and defined variant

- Hebrew: מוסר vs. מוסר
- Latin: testi vs. testi
- Bangla: রাজ vs. রাজ

2. Highly Confusable (strongly similar, or near homograph)

- Ethiopic: ሀገር vs. ሀገር
- Latin: for vs. for

3. Similar

- Latin: cafe vs. café
- Thai-Lao: จบ vs. จบ

Similarity Categorization

- Categories

4. Distantly Similar (weakly similar). Confused by a minority of attentive users, when not looking at cases side-by-side. Experts are not actively looking for all of these cases, but note them on corner cases.
 - Latin: pho vs phỏ
 - Chinese: 不足 vs 不是
5. Distinct (not similar). These will be the vast majority of all pairs. There is no need to look at them.

Status	# Scripts
Preparation	Completed
Expert working on initial data	Completed
Internal review in progress	Completed
Preparing for Public Comment	27

The 27 scripts* include: 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, Thaana, Telugu, and Thai.

*Chinese, Japanese, Korean cover four scripts: Han, Hiragana, Katakana, Hangul

List of Experts

Akshat Joshi - Devanagari, Gujarati
Atiur Rahman Khan - Bengali (Bangla)
Dessaiegn Yehuala - Ethiopic
Dmitry Belyavsky - Cyrillic
Gurpreet Singh Lehal - Gurmukhi
Harsha Wijayawardhana - Sinhala
Hiro Hotta - Japanese (Hiragana, Katakana)
Katarina Gevorgyan - Armenian
Kim Kyoungsook - Korean (Han, Hangul)
Kuldeep Patnaik - Oriya
Makara Sok - Khmer
Matitiah Allouche - Hebrew
Michael Bauland - Latin

Naail Abdul Rahman - Thaana
Nabil Benamar - Arabic
Panagiotis Papaspiliopoulos - Greek
Parkpoom Tripatana - Thai
Pavanaja U B - Kannada
Saysomvang Souvannavong - Lao
Shanmugam Rajabadher - Tamil
Thin Zar Phyo - Myanmar
Uma Maheshwar Rao G - Telugu
Veena Solomon - Malayalam
Wang Wei - Chinese (Han)
Zakharia Pourtskhvanidze - Georgian

Feedback Request

Overview of the SSE Data

- Data will be available in two formats: HTML, XML
- Content includes:
 - Description
 - Repertoire
 - Confusable Sets
 - Rules and Actions
 - Table of Reference

The screenshot displays the 'Armenian Script: String Similarity Evaluation Data' document. At the top, a table provides metadata: Date (2025-05-22), LGR Version (1), Language (und-Armn (Armenian Script)), and Unicode Version (Version 16.0.0). A text box on the right states: 'This document is mechanically formatted from the above XML file for the LGR. It provides additional summary data and explanatory text. The XML file remains the sole normative specification of the LGR.'

The 'Table of Contents' lists: 1 Description, 2 Repertoire, 3 Confusables, 4 Rules and Actions (with sub-items 4.1 Context Rules and 4.2 Actions), and 5 Table of References.

The '1 Description' section includes an 'Overview' paragraph: 'This document provides a number of sets of confusables mappings for the Armenian script. This data is intended to be used in String Similarity Evaluation [SSE]. Generally, if it is possible to convert a given label into another label by applying the mappings listed in this document to one or more code points or sequences, then the two labels are considered (potentially) confusable.' It continues: 'In String Similarity Evaluation, the first step will be to discover whether two labels can be mapped into each other using these mappings. In a second step, a visual distance metric is determined based on similarity category values. The higher each category is, the more it contributes to the distance. Above some threshold to be determined, the two labels would be considered too distant to be confusable.'

A yellow box contains a draft notice: 'This is a DRAFT document released for public comments and not final. For details on how to submit comments see the announcement for public comments on String Similarity Evaluation Data on the ICANN website. <https://icann.org/idn>'.

The bottom of the image shows an XML editor window titled 'sse-chinese-full-variant-script-29apr25-en.xml' in 'Free Mode'. It displays XML code with line numbers 119 to 168. The code includes a root element 'RZ-LGR' with an integration panel, a description, and a list of characters with their properties (cp, comment, type, ref, when, etc.).

Confusable Sets

- Example in Armenian SSE Data: **U+0449 (՝)** CYRILLIC SMALL LETTER SHCHA, **U+057A (ԝ)** ARMENIAN SMALL LETTER PEH, **U+1223 (ሣ)** Ethiopic ETHIOPIC SYLLABLE SZAA are in the same set.

Confusables Set 28 — 3 Members

Source	Glyph	Target	Glyph		Similarity Category	Ref	Comment
0449	՝	0449	՝	≡	out-of-repertoire-var		Not part of repertoire
0449	՝	057A	ԝ	↔	3	[101] /	Underlining (Changed due to the integration with Cyrillic. The old values were - Category:4, Reason:not_confusing)
0449	՝	1223	ሣ	↔	4	[101]	Imposed based on confusables
057A	ԝ	1223	ሣ	↔	3	/ [101]	Confusing This is a case of cross-scripts visual similarity between Ethiopic and Armenian scripts. The two glyphs (ሣ, ԝ) appear confusingly similar, especially in handwritten forms.(Added due to the integration with Ethiopic)
1223	ሣ	1223	ሣ	≡	out-of-repertoire-var		Not part of repertoire

Reason Code

Reason Code	Description
Capitalization	The similarity is between two uppercase forms or between an uppercase and unrelated lowercase form. These cases are included because user agents will accept uppercase forms, converting them before lookup.
Confusing	A majority of users familiar with the script of either target, source or both will confuse the pair unless viewed side-by-side.
Fallback	One of the code points is a common fallback representation for the other.
Font	The confusability depends on font choice.
Handwriting	The confusability is caused or strengthened by users being familiar with a shape, even though it is normally only used in handwriting.
Other	None of the other reason code applies. In particular, it applies when the category is homoglyph (1) for a mapping that exceptionally has not been defined as variant. NOTE: This code MUST be accompanied by a comment explaining the nature of the confusability and reasoning for choosing the assigned category value.
Underlining	The presence of underlining makes these confusable. This is typically the case when the distinction between two characters rests primarily on a mark below or a descender that might be obscured when underlined.
Variant	The mapping is defined as a variant in the [RZ-LGR]. NOTE: the category value MUST be set to 1.

Reason Code - Special Values

Special values	
Not_confusing	Mappings that are marked with Reason code "not_confusing" are listed here for reference only. They will be ignored in screening or label distance calculations.
Imposed	This Reason code is automatically generated when a mapping has to be imposed to make a confusables set transitive. NOTE: After review, any imposed mapping will be assigned category 4. If that is not appropriate, for example, because the affected pair of code points is strongly similar, then a mapping needs to be added for that pair of code points stating the corrected category with an appropriate reason.

- Example of imposed case: Latin script similarity data.

Source	Glyph	Target	Glyph	Similarity Category	Comment
006D	m	006D 0327	m̲	3	Underlining
006D	m	0072 006E	rn	2	Font
006D 0327	m̲	0072 006E	rn	4	Imposed

} Defined by the expert

} Mechanically added

Feedback Request

- Identify if any similarity pairs are missed? Provide similarity level and reason along with the pair of characters.
- For the pairs identified by experts, review the degree of similarity marked, and either confirm the value, or provide an updated value with a reason.
- Any "imposed" mappings receive a category 4 by default. Review and see if this should be updated to a value between 1-3.
- Any other feedback.

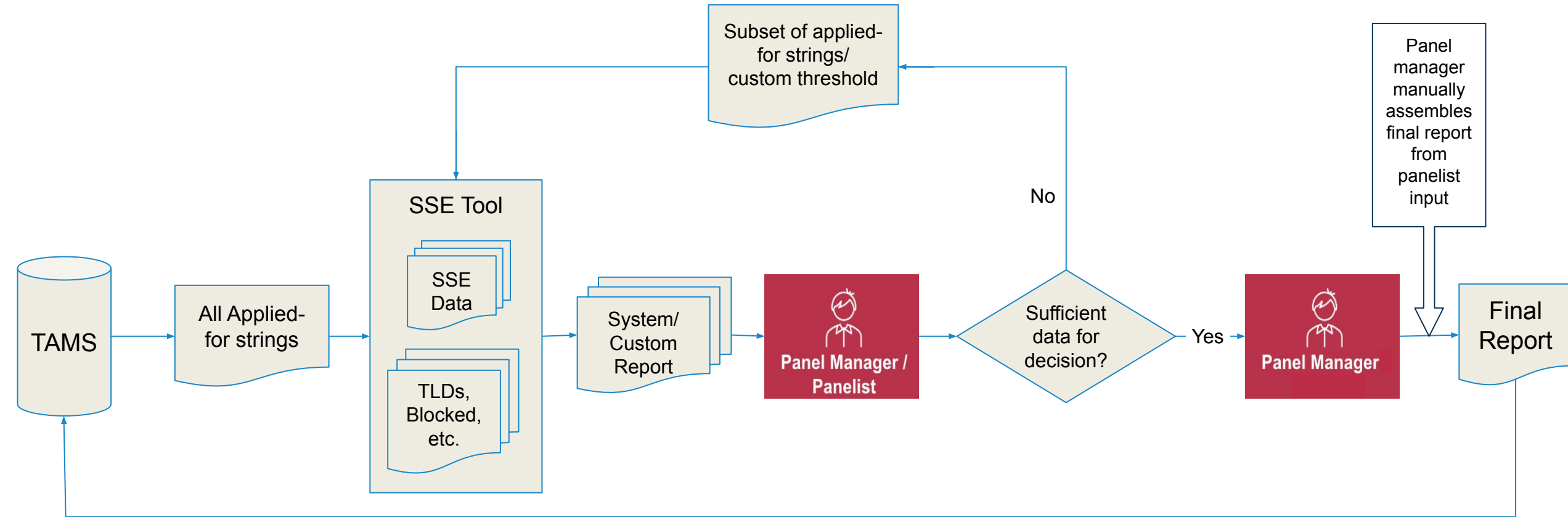
Next Steps

- - Add mechanical update to the data in preparation for Public Comment
 - Review the data and add the “imposed” cases and set as Category 4 by default
 - Add variant definitions from the RZ-LGR as Category 1
 - Review the data in consultation with experts
 - Confirm the accuracy of mechanical updates
 - Review and finalize the confusable sets as well as information in comments
 - Open for Public Comment
 - Discuss Public Comment input with relevant script experts to finalize the data
 - Publish the similarity data and incorporate in the SSE tool

Update on the String Similarity Evaluation Tool

Pitinan Kooarmornpatana, ICANN

String Similarity Evaluation Process Flow



- Panelist, not tool, determines final decision on similarity
- Panelist can create custom report to evaluate subset of applied-for strings
- Panelist can manually edit outcome before providing to Panel manager to add to Final Report
- Once Final Report is uploaded to TLD Application Management System (TAMS), TAMS will handle determination (e.g., applied-for string similar to other applied-for string is in contention, similar to Blocked name is rejected, etc.)

Overview of the Tool

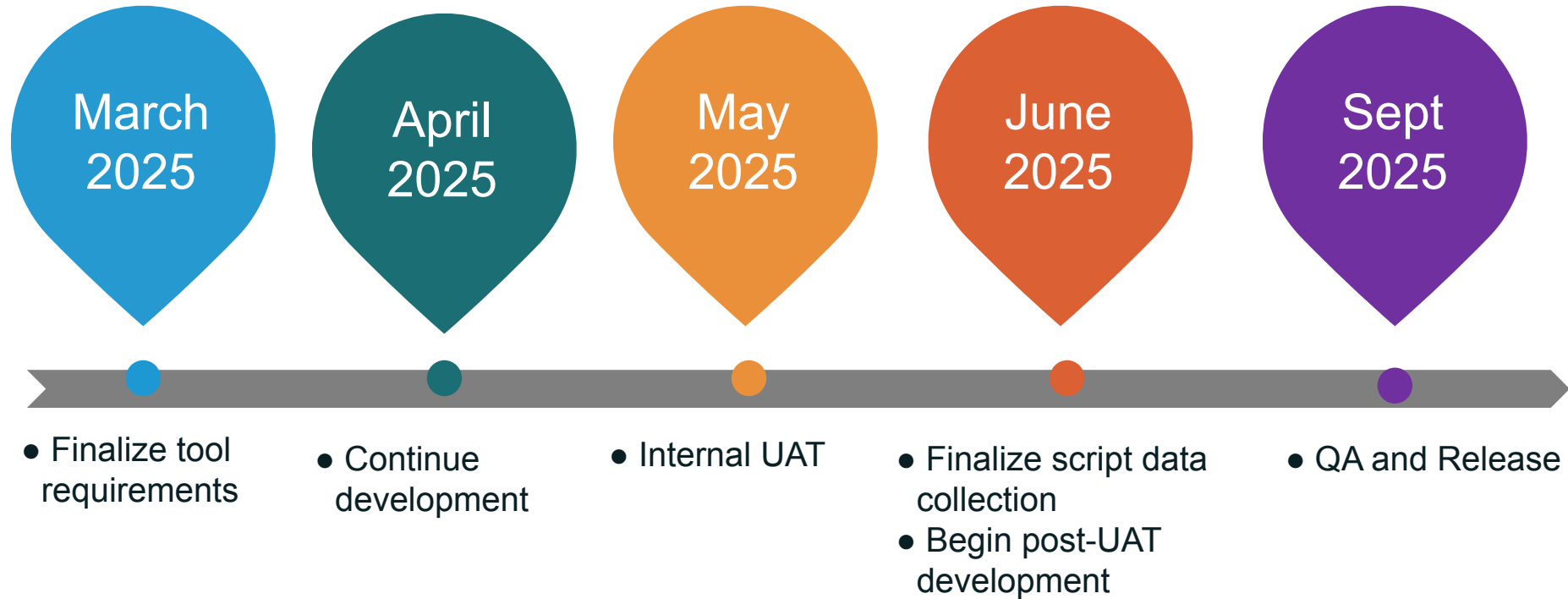
- ICANN Account with Panelist privileges can access on icann.org
- Provides automated string similarity analysis intended to optimize manual review
- Incorporates Similarity Data comparison rules as described in previous slides
- Compares applied-for strings (including tool-generated variants) to:
 - All other applied-for strings
 - Existing gTLDs & ccTLDs
 - Requested IDN ccTLDs
 - Previous round, in-process applied-for gTLDs
 - Two-character ASCII
 - Blocked Names
- Tool functions
 - Detect potentially confusable labels
 - Calculate a visual distance score
 - Labels with scores within threshold are included in the report for manual review.



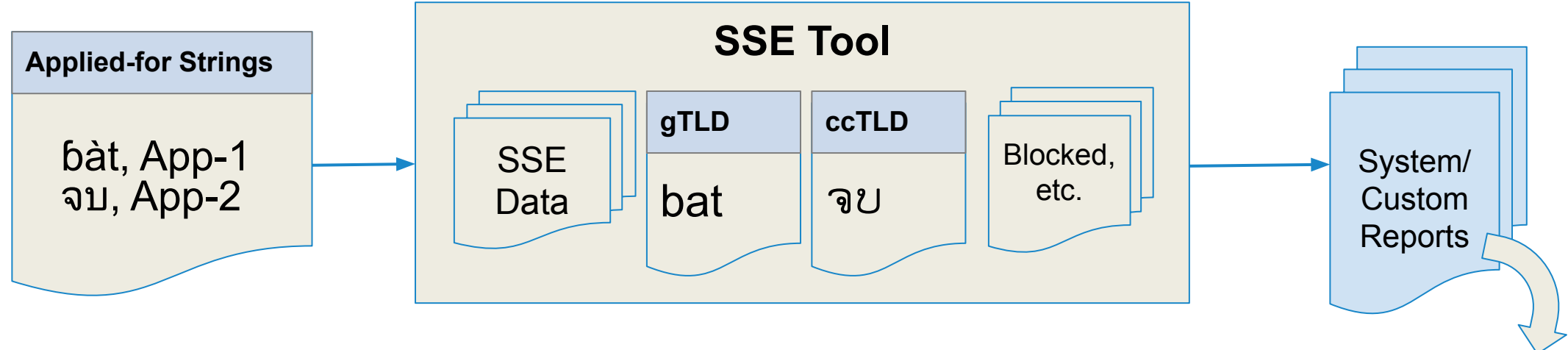
Example: Output Report

- ◉ Each contention set consists of one or more rows of comparison pairs (Label 1 & Label 2). Each row contains the following columns:
 - Set #
 - Row ID (unique row ID)
 - Label 1 & A-Label (Label 1 is always an applied-for string)
 - Label 1 App ID
 - Label 2 & A-Label
 - Label 2 App ID
 - Label 2 Source (ccTLD, gTLD, applied-for, blocked, etc.)
 - Label in Label 1's variant set with the similarity within the threshold
 - Label in Label 2's variant set with the similarity within the threshold
 - Similarity Category 3 or Less (Y/N)
 - Score
 - Rationale
 - Reviewer
 - Notes

Timeline of String Similarity Evaluation Data and Tool



Example1: Applied-for strings are similar to existing TLDs



Set	Row ID	Primary Label 1				Primary Label 2					Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source	U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
1	1	bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	App-1	bat	bat	U+0062(b) U+0061(a) U+0074(t)	N/A	gTLD	bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bát	xn--bt-mia	U+0062 (b) U+00E1 (á) U+0074 (t)	blocked	TRUE	0.118	2-2-1
	2										bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	TRUE	0.186	2-3-1
	3										bát	xn--t-y0a858s	U+0253 (b) U+1EA3 (ǎ) U+0074 (t)	blocked	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	FALSE	0.264	2-4-1
2	4	ຈປ	xn--82c0b	U+0E08 (ຈ) U+0E1A (ງ)	App-2	ຈປ	xn--w6c0b	U+0E88 (ຈ) U+0E9A (ງ)	N/A	ccTLD	ຈປ	xn--82c0b	U+0E08 (ຈ) U+0E1A (ງ)	-	ຈປ	xn--w6c0b	U+0E88 (ຈ) U+0E9A (ງ)	-	TRUE	0.177	2-2

Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 1: **bàt**, **bat** are in the same potential contention set.

- Primary Level 1 columns are details of the applied-for string.
- Primary Level 2 columns are details of the string being compared. In this case, **bat** is an example existing gTLD.

Set	Row ID	Primary Label 1				Primary Label 2				
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source
1	1	bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	App-1	bat	bat	U+0062(b) U+0061(a) U+0074(t)	N/A	gTLD
	2									
	3									

Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 1: **bat**, **bat**

- Variant labels of **bat** and variant labels of **bat** are generated
- All variant members from the two sets, except blocked variant vs blocked variant, are compared
- For each pair, the tool calculates the similarity level and similarity score using the SSE data.

bat variant set

bat (primary label)

bát (blocked variant label)

bat variant set

bat (primary label)

bát (blocked variant label)

Categories for Comparison:		The applied-for gTLD string		
		Primary gTLD string	All allocatable variant string(s)	All blocked variant string(s)
• Existing gTLD • The gTLD string applied-for in the previous round(s) still in the process • Existing ccTLD • Requested IDN ccTLD • Other Applied-for gTLD • Blocked Name • Any two-Character ASCII	Primary String	Yes	Yes	Yes*
	All allocatable variant string(s)	Yes	Yes	Yes*
	All blocked variant string(s)	Yes*	Yes*	No

Example1: Applied-for strings are similar to existing TLDs

- **Rationale** is a sequence of similarity category mappings as defined in the SSE data.
- **Level 3 Similar** is TRUE if all similarity category mapping are 3 or less, otherwise FALSE.
- **Score** is calculated based on the similarity category, result from 0 -1 where 0 is the most similar (same) and 1 is distinguishable.

Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bát	xn--bt-mia	U+0062 (b) U+00E1 (á) U+0074 (t)	blocked	TRUE	0.118	2-2-1
bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	TRUE	0.186	2-3-1
bât	xn--t-y0a858s	U+0253 (b) U+1EA3 (â) U+0074 (t)	blocked	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	FALSE	0.264	2-4-1
bât	xn--t-y0a858s	U+0253 (b) U+1EA3 (â) U+0074 (t)	blocked	bát	xn--bt-mia	U+0062 (b) U+00E1 (á) U+0074 (t)	blocked	Block vs Blocked Not in the Report		

Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 1: **bàt**, **bat**

- The most similar pair is the string 1's primary label (bàt) vs the string 2's blocked variant label (bát)

Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bát	xn--bt-mia	U+0062 (b) U+00E1 (á) U+0074 (t)	blocked	TRUE	0.118	2-2-1
bàt	xn--t-rfa62i	U+0253 (b) U+00E0 (à) U+0074 (t)	-	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	TRUE	0.186	2-3-1
bàt	xn--t-y0a858s	U+0253 (b) U+1EA3 (à) U+0074 (t)	blocked	bat	bat	U+0062 (b) U+0061 (a) U+0074 (t)	-	FALSE	0.264	2-4-1

Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 2: **၁၂**, **၁၃** are in the same potential contention set.

- Primary Level 1 columns are details of the applied-for string.
- Primary Level 2 columns are details of the string being compared. In this case, **၁၃** is an example existing ccTLD.

Set	Row ID	Primary Label 1				Primary Label 2				
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source
2	4	၁၂	xn--82c0b	U+0E08 (၁) U+0E1A (၂)	App-2	၁၃	xn--w6c0b	U+0E88 (၁) U+0E9A (၃)	N/A	ccTLD

Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 2: ခပ, ခပ

- Both labels do not have variant labels

ခပ variant set
ခပ (primary label)

Categories for Comparison:		The applied-for gTLD string		
		Primary gTLD string	All allocatable variant string(s)	All blocked variant string(s)
- Existing gTLD - The gTLD string applied-for in the previous round(s) still in the process - Existing ccTLD - Requested IDN ccTLD - Other Applied-for gTLD - Blocked Name - Any two-Character ASCII	Primary String	Yes	Yes	Yes*
	All allocatable variant string(s)	Yes	Yes	Yes*
	All blocked variant string(s)	Yes*	Yes*	No ☒

ခပ variant set
ခပ (primary label)

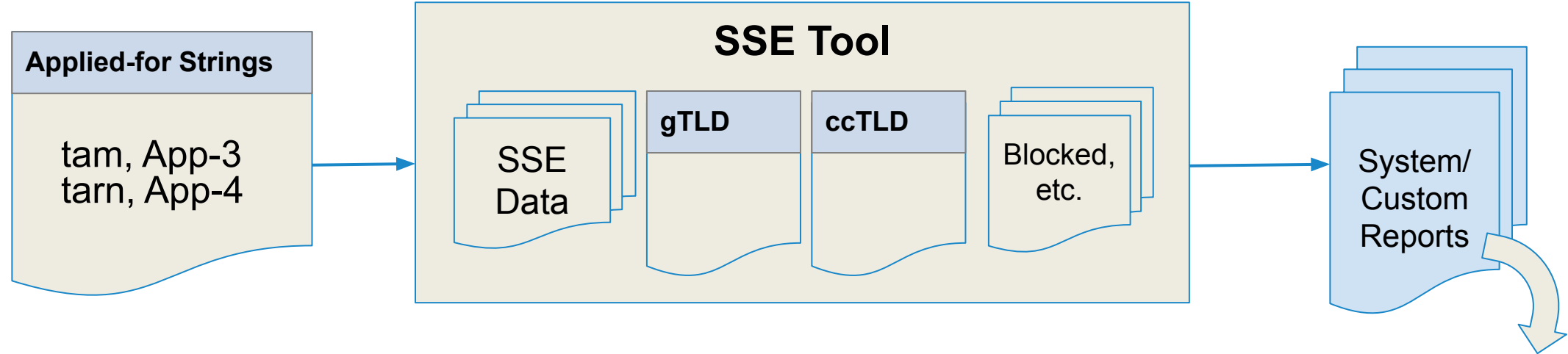
Example1: Applied-for strings are similar to existing TLDs

Zoom-in to Set 2: ခဏ, ခဏ

- The most similar pair is the string 1's primary label (ခဏ) vs the string 2's blocked variant label (ခဏ)

Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
ခဏ	xn--82c0b	U+0E08 (ခ) U+0E1A (ဏ)	-	ခဏ	xn--w6c0b	U+0E88 (ခ) U+0E9A (ဏ)	-	TRUE	0.177	2-2

Example2: An applied-for string is similar to another applied-for string



Set	Row ID	Primary Label 1				Primary Label 2					Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source	U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
1	1	tam	tam	U+0074 U+0061 U+006D	App-3	tarn	tarn	U+0074 U+0061 U+0072 U+006E	App-4	Applied For	tam	tam	U+0074 (t) U+0061 (a) U+006D (m)	-	tarn	tarn	U+0074 (t) U+0061 (a) U+0072 U+006E (rn)	-	TRUE	0.083	1-1-2
	2										tám	xn--tm-mia	U+0074 (t) U+00E1 (á) U+006D (m)	blocked	tarn	tarn	U+0074 (t) U+0061 (a) U+0072 U+006E (rn)	-	FALSE	0.264	1-4-2

Example2: An applied-for string is similar to another applied-for string

Zoom-in to Set 1: **tam**, **tarn** are in the same potential contention set.

- Primary Level 1 columns are details of the applied-for string, **tam**
- Primary Level 2 columns are details of the string being compared. In this case, **tarn** is an example another applied-for string.

Set	Row ID	Primary Label 1				Primary Label 2				
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source
1	1	tam	tam	U+0074 U+0061 U+006D	App-3	tarn	tarn	U+0074 U+0061 U+0072 U+006E	App-4	Applied For
	2									

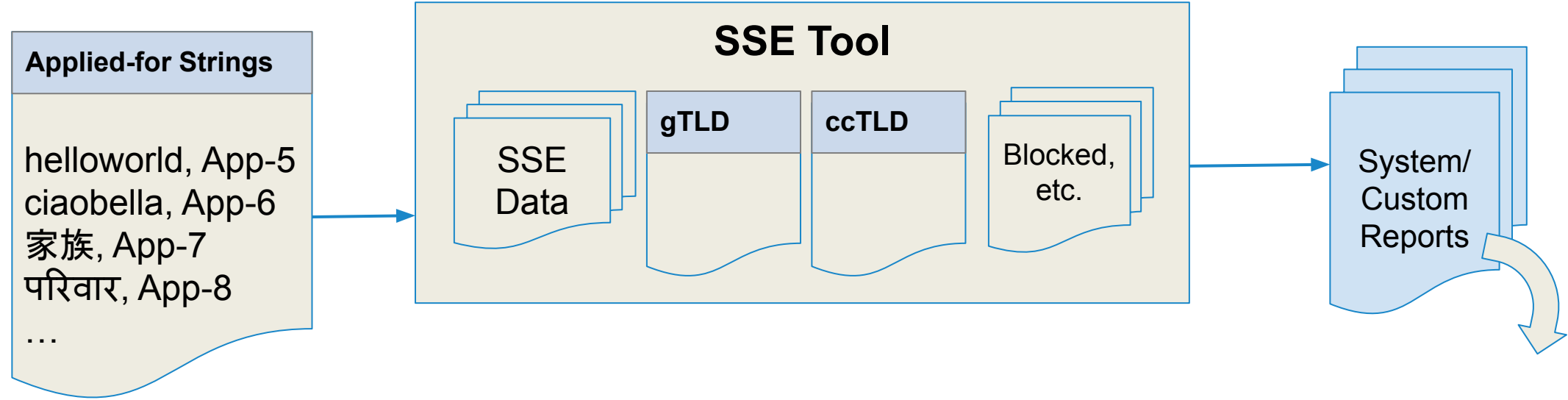
Example2: An applied-for string is similar to another applied-for string

Zoom-in to Set 1: **tam**, **tarn**

- The most similar pair is the string 1's primary label (tam) vs the string 2's primary label (tarn)

Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
tam	tam	U+0074 (t) U+0061 (a) U+006D (m)	-	tarn	tarn	U+0074 (t) U+0061 (a) U+0072 U+006E (rn)	-	TRUE	0.083	1-1-2
tám	xn--tm-mia	U+0074 (t) U+00E1 (á) U+006D (m)	blocked	tarn	tarn	U+0074 (t) U+0061 (a) U+0072 U+006E (rn)	-	FALSE	0.264	1-4-2

Example3: An applied-for string is not found similar to any other strings



Set	Row ID	Primary Label 1				Primary Label 2					Variant of Label 1 within the similarity threshold				Variant of Label 2 within the similarity threshold				Level 3 Similar	Score	Rationale
		U-label	A-label	Code Points	App ID	U-label	A-label	Code Points	App ID	Source	U-label	A-label	Code Points	Disposition	U-label	A-label	Code Points	Disposition			
No data available in table																					

Engage with ICANN – Thank You and Questions



One World, One Internet

Visit us at **icann.org**



@icann



facebook.com/icannorg



youtube.com/icannnews



flickr.com/icann



linkedin.com/company/icann



soundcloud.com/icann



instagram.com/icannorg