



78 | ANNUAL
GENERAL
MEETING



Planning for String Similarity Review in the Next gTLD Round



Sarmad Hussain, Sr. Director IDN and UA Programs
Asmus Freytag

ICANN78
26 October 2023

Visually Same (Variant), Similar and Distinct Characters

Category	Code Point	Script	Code Point	Script
Visually “same”	c (U+0063) с (U+0441)	Latin Cyrillic	ﻉ (U+0649) ﻉ (U+06CC)	Arabic Arabic
Visually similar	ç (U+00E7) ς (U+03C2)	Latin Greek	घ (U+0918) ঘ (U+0998)	Devanagari Bengali (Bangla)
Visually distinct	a (U+0061) f (U+0066)	Latin Latin	へ (U+30DA) ホ (U+30DB)	Japanese Japanese

- ◎ String Similarity Review recommendations:
 - Final Report on the new gTLD Subsequent Procedures Policy Development Process ([SubPro](#))
 - Phase 1 of the Internationalized Domain Names Expedited Policy Development Process ([IDN EPDP](#))
 - Public Comment version presented here
 - Will be updated based on public comments received
- ◎ Scope, Methodology and Outcomes of String Similarity Review
- ◎ String Similarity Review Guidelines

String Similarity Review Recommendations

String Similarity in SubPro

- ⦿ **Affirmation 24.1:** The Working Group affirms Recommendation 2 from the 2007 policy, which states “Strings must not be confusingly similar to an existing top-level domain or a Reserved Name.”
- ⦿ **Affirmation 24.2:** Subject to the recommendations below, the Working Group affirms the standard used in the String Similarity Review from the 2012 round to determine whether an applied-for string is “similar” to any existing TLD, any other applied-for strings, Reserved Names, and in the case of 2-character IDNs, any single character or any 2-character ASCII string. According to Section 2.2.1 of the 2012 Applicant Guidebook, “similar” means “strings so similar that they create a probability of user confusion if more than one of the strings is delegated into the root zone.” In the 2012 round, the String Similarity Panel was tasked with identifying “visual string similarities that would create a probability of user confusion.” The Working Group affirms the visual standard for determining similarity with the updates included in the recommendations below.
- ⦿ **Recommendation 24.6:** Eliminate the use of the SWORD tool in subsequent procedures.
- ⦿ **Recommendation 24.7:** The deadline for filing a String Confusion Objection must be no less than thirty (30) days after the release of the String Similarity Evaluation results.

String Similarity Review in IDN EPDP – Phase 1 – Scope of Comparisons

- ◎ **Preliminary Recommendation 4.1:** The String Similarity Review must be modified to compare 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) against the following:
 - Existing gTLDs and all of their allocatable and blocked variant labels; and
 - Existing ccTLDs and all of their allocatable and blocked variant labels; and
 - Strings requested as IDN ccTLDs and all of their allocatable and blocked variant labels; and
 - Other applied-for gTLD strings and all of their allocatable and blocked variant labels; and
 - All strings on the Reserved Names list and all of their allocatable and blocked variant labels; and
 - Any other two-character ASCII strings and all of their allocatable and blocked variant labels (if the applied-for gTLD string is a two-character string).
- ◎ In addition, the blocked variant label(s) of an applied-for primary gTLD string must also be compared against the following:
 - Existing gTLDs and all of their allocatable variant labels; and
 - Existing ccTLDs and all of their allocatable variant labels; and
 - Strings requested as IDN ccTLDs and all of their allocatable variant labels; and
 - Other applied-for gTLD strings and all of their allocatable variant labels; and
 - All strings on the Reserved Names list and all of their allocatable variant labels; and
 - Any other two-character ASCII strings and all of their allocatable variant labels (if the applied-for gTLD string is a two-character string).

String Similarity Review in IDN EPDP – Phase 1 – Scope of Comparisons

- ⦿ **Preliminary Recommendation 4.2:** As an exception to the proposed modification to the String Similarity Review in accordance with Preliminary Recommendation 4.1, the String Similarity Review Panel may decide whether and what blocked variant labels to omit when conducting a comparison. Any such decision by the String Similarity Review Panel must be based on guidelines and/or criteria that justify such an omission on the basis of a manifestly low level of confusability between the scripts of labels being compared.
- ⦿ **Preliminary Recommendation 4.3:** During implementation, the guidelines and/or criteria must be developed for use by the String Similarity Review Panel to decide on the omission of blocked variant labels when conducting a comparison.

- ◎ **Preliminary Recommendation 4.4:** All labels from a variant label set, comprising the primary gTLD string and all of its allocatable and blocked variant labels, must share the same outcome out of the String Similarity Review. This means the String Similarity Review, in accordance with Preliminary Recommendation 4.1-4.3, determines that:
 - If an applied-for primary gTLD string or any of its variant label(s) is confusingly similar to an existing gTLD or ccTLD or any of its variant label(s), the entire variant label set of the applied-for primary gTLD string will be ineligible to proceed in the application process; or
 - If an applied-for primary gTLD string or any of its variant label(s) is confusingly similar to another applied-for primary gTLD string or any of its variant label(s), the entire variant label sets of the two applied-for primary gTLD strings will be placed in a contention set. Upon the resolution of the contention set, the application that prevails can proceed to the next stage of the application process.

- ⦿ **Preliminary Recommendation 6.1:** An applied-for primary gTLD string that is also a variant label of another applied-for primary gTLD string, as calculated by the RZ-LGR, must be placed in a contention set.
- ⦿ **Preliminary Recommendation 6.2:** The entire variant label set of an applied-for primary gTLD string (no matter whether it is an ASCII string or an IDN string) must be processed in the contention set.

Scope, Methodology, and Outcomes of String Similarity Review

Scope of Comparisons for the String Similarity Review

Categories for Comparison:		The applied-for gTLD string		
		Primary gTLD string	All allocatable variant string(s)	All blocked variant string(s)
• Existing gTLD • gTLD string applied-for in the previous round(s) still in process • Existing ccTLD • Requested IDN ccTLD • Other applied-for gTLD • Reserved 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

Yes: the comparisons done by the String Similarity Review Panel

Yes*: the comparisons (in gray) may be omitted by the Panel if it determines low level of confusability between the scripts of the strings, following the String Similarity Review Guidelines.

No: the comparisons not performed.

Methodology of String Similarity Review

1. Compile the lists of strings for comparison:
 1. Existing gTLDs, requested gTLDs, existing ccTLD, requested IDN ccTLDs, other applied-for gTLDs, Reserved Names, two-character ASCII strings.
 2. All allocatable variant strings of the above strings using RZ-LGR.
 3. All blocked variant strings of the above strings using RZ-LGR (in case the blocked list is large, develop an alternate feasible mechanism)
2. Decide which blocked variant strings to omit, if any, and document the rationale for the decision.
3. Identify different applications with the same variant-strings-set for contention.
4. Conduct the comparison of the strings for similarity based on the String Similarity Review Guidelines.
5. Document (along with rationale) the outcome of the String Similarity Review.

Outcomes of the String Similarity Review

- ⦿ The String Similarity Review Panel will document the following outcomes:
 - String similar to existing gTLDs
 - String similar to the gTLD strings applied-for in the previous round(s) and in the process
 - String similar to existing ccTLDs
 - String similar to requested IDN ccTLDs
 - String same or similar to other applied-for gTLDs
 - String similar to Reserved Names
 - String similar to Two-character ASCII strings
 - String not similar to any of these categories listed
- ⦿ ICANN org will publish the outcomes of the String Similarity Review on its website.

Outcomes of the String Similarity Review

	If the applied-for gTLD string or any member of its variant-strings-set is found to be		
	Same as	Variant of	Visually similar to (but not a variant of)
Existing gTLD	Application cannot proceed.	Application can proceed if existing registry operator is also the applicant.	Application cannot proceed.
The gTLD string from the previous round(s) still in the process	Application put on hold until the previous string completes evaluation. Application can proceed with evaluation if the gTLD string from the previous round is withdrawn or not successfully evaluated.	Same registry operator cannot proceed. For another registry operator, application put on hold until the previous string completes evaluation. Application can proceed with evaluation if the gTLD string from the previous round is withdrawn or not successfully evaluated.	Application put on hold until the previous string completes evaluation. Application can proceed with evaluation if the gTLD string from the previous round is withdrawn or not successfully evaluated.
Existing ccTLD	Application cannot proceed.	Application cannot proceed.	Application cannot proceed.
Requested IDN ccTLD	Application can proceed if it has successfully completed all relevant evaluation stages, or else application put on hold until ccTLD evaluation is completed and application can proceed if Requested IDN ccTLD is withdrawn or not successfully evaluated.	Application can proceed if it has successfully completed all relevant evaluation stages, or else application put on hold until ccTLD evaluation is completed and application can proceed if Requested IDN ccTLD is withdrawn or not successfully evaluated.	Application can proceed if it has successfully completed all relevant evaluation stages, or else application put on hold until ccTLD evaluation is completed and application can proceed if Requested IDN ccTLD is withdrawn or not successfully evaluated.
Other Applied-for gTLD String	Application put in contention set.	Application put in contention set.	Application put in contention set.
Reserved Name	Application cannot proceed.	Application cannot proceed.	Application cannot proceed.
Two-Character ASCII String	Application cannot proceed.	Application cannot proceed.	Application cannot proceed.

String Similarity Review Guidelines

Contents

- ⦿ Concepts
- ⦿ Relation and Contrast to Variants
- ⦿ Similarity and Scripts
- ⦿ Approaches
- ⦿ Next Steps

- ⦿ Confusability
- ⦿ Similarity
- ⦿ Target User

- ⦿ **Confusability:** the degree to which two labels can be substituted for each other without being noticed. Users are confused about which label they are selecting.
- ⦿ **Confusable:** said of two labels for which the degree of confusability exceeds an implicit and hard to define threshold.

- ⦿ **String Similarity:** similarity in appearance can contribute to the confusability between two labels.
- ⦿ **Confusable Similarity:** the degree of similarity required to make two similar labels “confusable.”

- ⦿ What looks similar and what can be confusable depends on the background of the user and expectations learned in reading the native language in the native script.
- ⦿ Evaluating two labels for confusable similarity requires definition of a target user.

Defining a Target User

- ⦿ Reasonably attentive user.
- ⦿ User natively familiar with script or language.
- ⦿ Include user being presented with label in foreign script or language when looking for a native label.
- ⦿ Exclude user deciding between two labels in a foreign script or language.
- ⦿ Assume a significant portion of users are also familiar with the ASCII subset.

String Similarity Applies to Whole Labels

- ⦿ **Unique substring:** A label with a unique substring is unique, and more so if that unique part is prominent.
- ⦿ **Recognition depends on the whole label context:** Word recognition is holistic; likewise the complete label sets up a semantic or recognition context where the reader, having seen the entire label, has an expectation for the spelling of it that can subconsciously override certain differences in presentation.
- ⦿ **Recognition does not proceed on a character-by-character basis:** Word detection relies on features. When closely spaced, readers don't always resolve which letter contains which feature.

Relation and Contrast To Variants

- ⦿ How are they different?
- ⦿ Where do they interact?

Contrast to Variants I

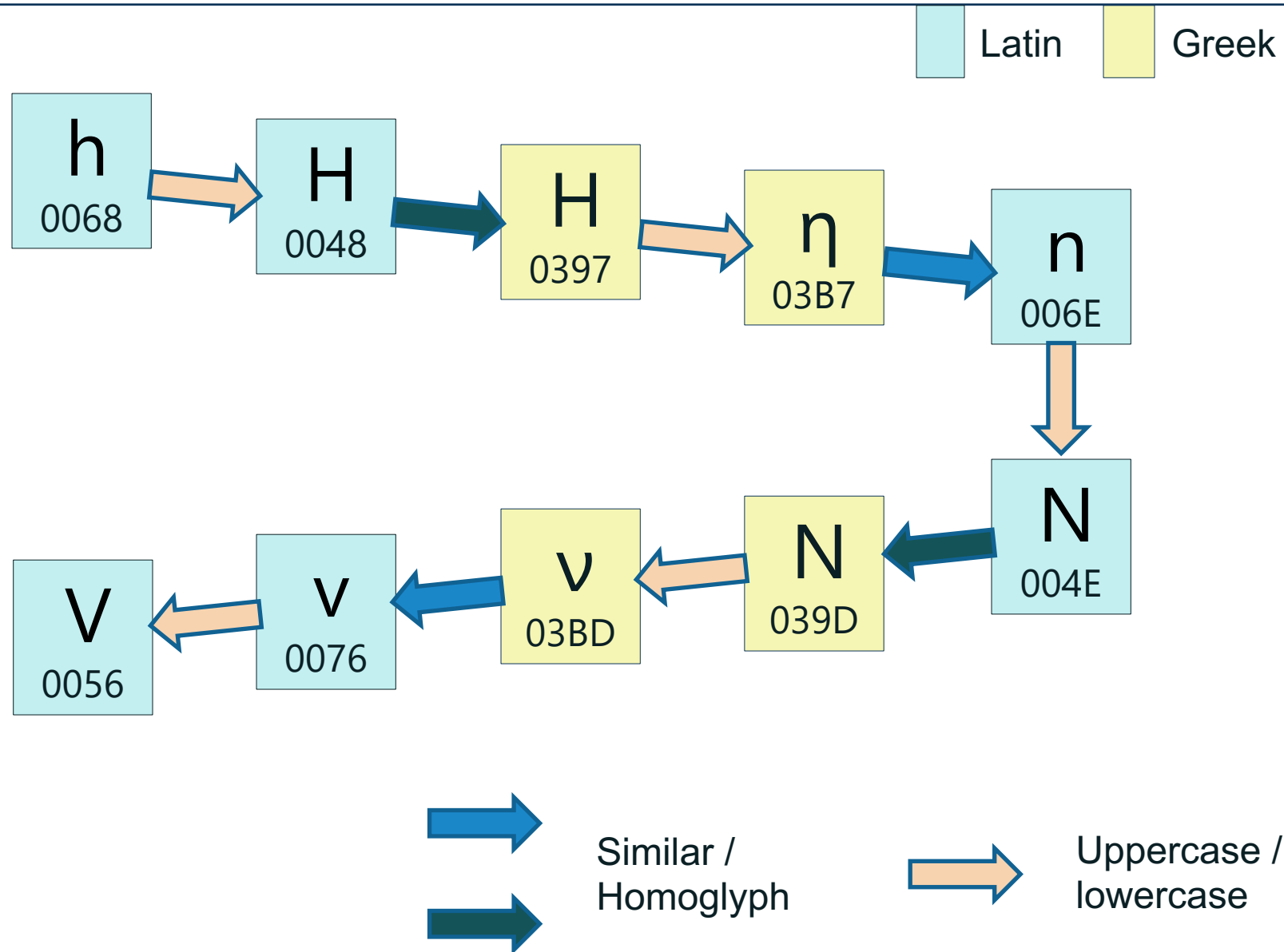
- ⊙ Variant: a label that a user will accept as “the same” as the original label.
 - Variants not limited to visual similarity.
 - Variant is an equivalence relation, and therefore transitive:
 - All variants for a label are variants of each other as well.
 - In contrast, “mere” similarity implies some distance in perception space.
 - Two labels similar to a third may not be similar to each other.

- ⊙ ⇒ Similarity and confusability are not transitive in the general case.

Contrast to Variants II

- ⦿ Variants based on visual similarity have identical or near identical appearance.
- ⦿ Variants do not consider uppercase-based similarity.
 - In scope for string similarity.

Example of Uppercase and Similarity

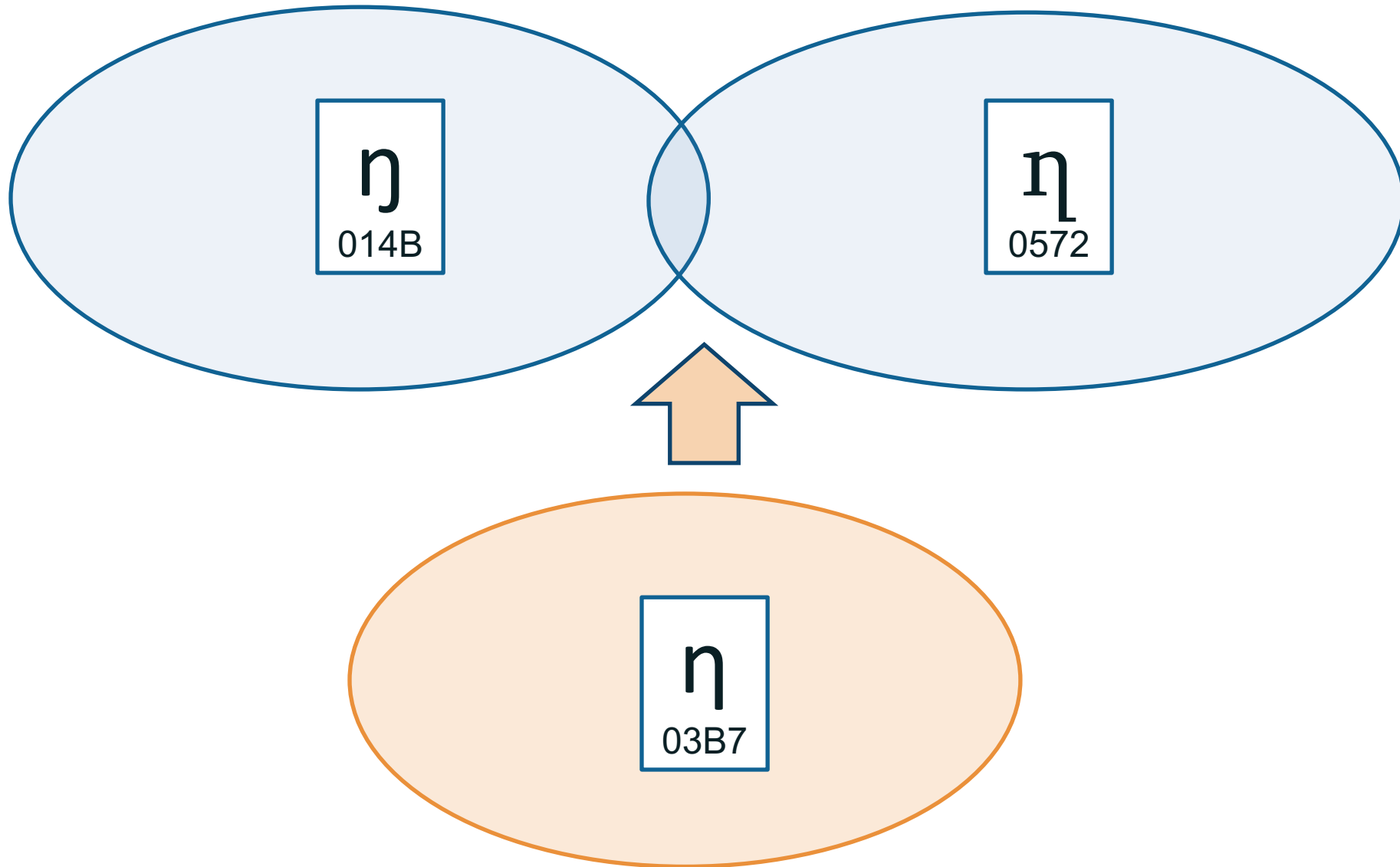


Interaction with Variants

- ⦿ All variants of the candidate are also confusable with the delegated label independent of whether they are applied for or not.
- ⦿ A candidate may be confusable with a variant of a delegated label even if the variant is not delegated.
- ⦿ Any variant label would be accepted as a “the same” of a label it is a variant of.
- ⦿ A candidate or its variant confusable with the variant of a delegated label would be accepted on the same footing as a candidate confusable with the delegated label.

- ⦿ How is string similarity affected by the presence of multiple scripts?
- ⦿ How does string similarity depend on a particular script?

Example of Cross-Script Visual Overlap



Script Dependence

- ⦿ Similarity is in the eye of the beholder.
- ⦿ Typographic and other features of the script influence the task of recognition of a label and discrimination from other labels.
- ⦿ Some scripts have available data on suggested confusables.

- ⦿ Some groups of scripts are closely related by history and features:
 - Latin/Greek/Cyrillic/Armenian: common origin, many common shapes.
 - Chinese, Japanese, Korean: overlapping repertoire, similar features.
 - Neo-Brahmi Scripts: common origin, similar structure, common shapes.
 - UI fonts (such as Segoe UI) show a tendency for harmonized glyphs.

Strongly Correlated Scripts

- ⊙ Some script pairs show a particularly strong overlap in letter shapes:
 - Kannada/Telugu (34).
 - Latin/Cyrillic (28).
 - Devanagari/Gurmukhi (26).
- ⊙ In contrast, Thai/Lao are strongly related, but typical font styles are distinct.
 - Latin/Greek share many uppercase forms.
 - ⇒ confusable if presented to the user in uppercase.
 - ⇒ LDH labels can look like Greek uppercase.

- ⦿ Brute force approaches don't scale.
- ⦿ For many labels it's not possible to enumerate all variants.
- ⦿ It's not possible to compare all variants of a candidate label with all delegated labels and their variants.
- ⦿ Possible approach – a two stage process:
 - Pre-screening to rule out all unlikely pairs.
 - Focused review of remaining candidates.

Pre-Screening: Goals

- ⦿ Reduce the number of labels to be verified in detail.
- ⦿ Filter out any label pairs that are dissimilar enough to not be confusable.
- ⦿ Automated, data-driven process to be scalable.

Possible Implementation

- ⊙ Leverage a modified variant mechanism to identify potentially confusable labels.
 - Score labels by suspected similarity based on letter similarities.
 - Dissimilar, Weak, Similar, Strong, Identical.
- ⊙ Treat confusables as superset of variants to cover interactions.
- ⊙ Quick and dirty (“loose variants”).
 - ⇒ May include some that shouldn’t be considered confusable.

What Kind of Data is Available?

- ⊙ Some data is available for L/G/C/A scripts (single code points):
 - Byproduct of Root Zone LGR development.
 - Additional data from registry policies.
- ⊙ Unicode confusables data:
 - Current status not consistently maintained.
 - Future direction being reviewed.
 - Need vetting and alignment with Root Zone.
- ⊙ Generally poor data for sequences:
 - ⇒ Handicap for scripts that write in “clusters” (Neo Brahmi).
 - But sequences not limited to Neo-Brahmi (.corn vs .com).

Manual Evaluation (Post-Screening Process)

- ⦿ How to get to a yes/no decision: “is it confusable”?
- ⦿ Guidelines not algorithm.
- ⦿ Final decision left to String Similarity Review Panel:
 - Panel may consult script experts.
 - Panel will not use focus groups.
 - Panel can review labels in addition to pre-screened ones.
- ⦿ String Similarity Review Panel will develop contention sets based on confusable similarity.

Next Steps

- ⦿ Finalize initial report and collect initial data set.
- ⦿ Public comment of initial report.
- ⦿ If the methodology acceptable, then
 - Trial-run of pre-screening on sample labels.
 - Refine data sets.
- ⦿ Community input on test results and finalized methodology.

Thank You!

Questions?