

Confusing String Similarity

Pitinan Kooarmornpatana
IDN and UA Program, Senior Manager
ICANN



APTLD83
22 February 2023

- Problem to be solved: users can be confused by two strings those are recognized as the 'same' or 'similar'
- Spectrum of 'Sameness' ranges from 'same', 'confusing similar', to 'different'
- Work done by the script community on the 'same' level
- Work in progress for the 'confusing similar' level

Problem to Solve

- Users can be confused by two strings those are recognized as the 'same' or 'similar'.

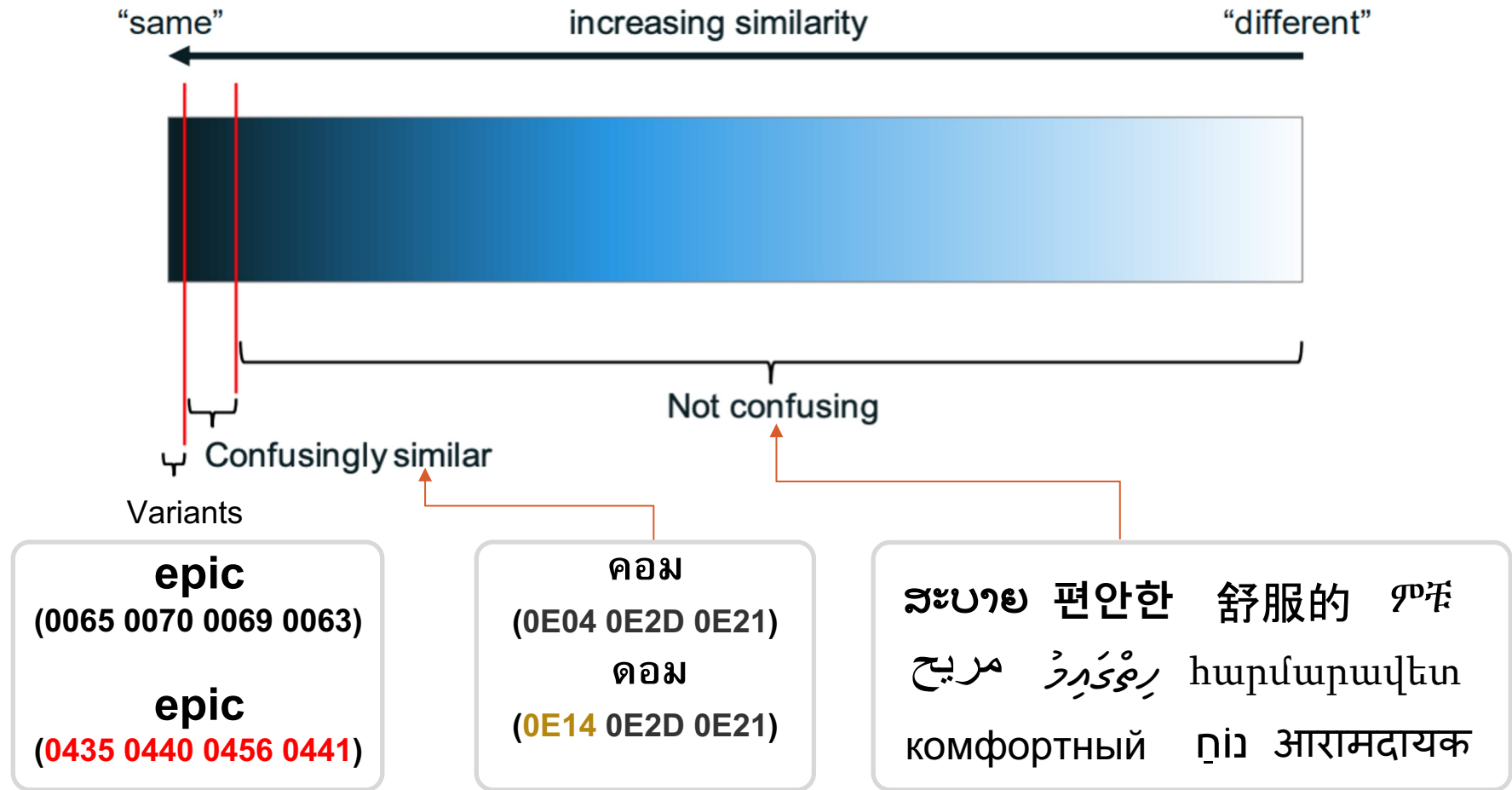
epic
(0065 0070 0069 0063)

คอม
(0E04 0E2D 0E21)

epic
(0435 0440 0456 0441)

คอม
(0E14 0E2D 0E21)

Spectrum of 'Sameness'

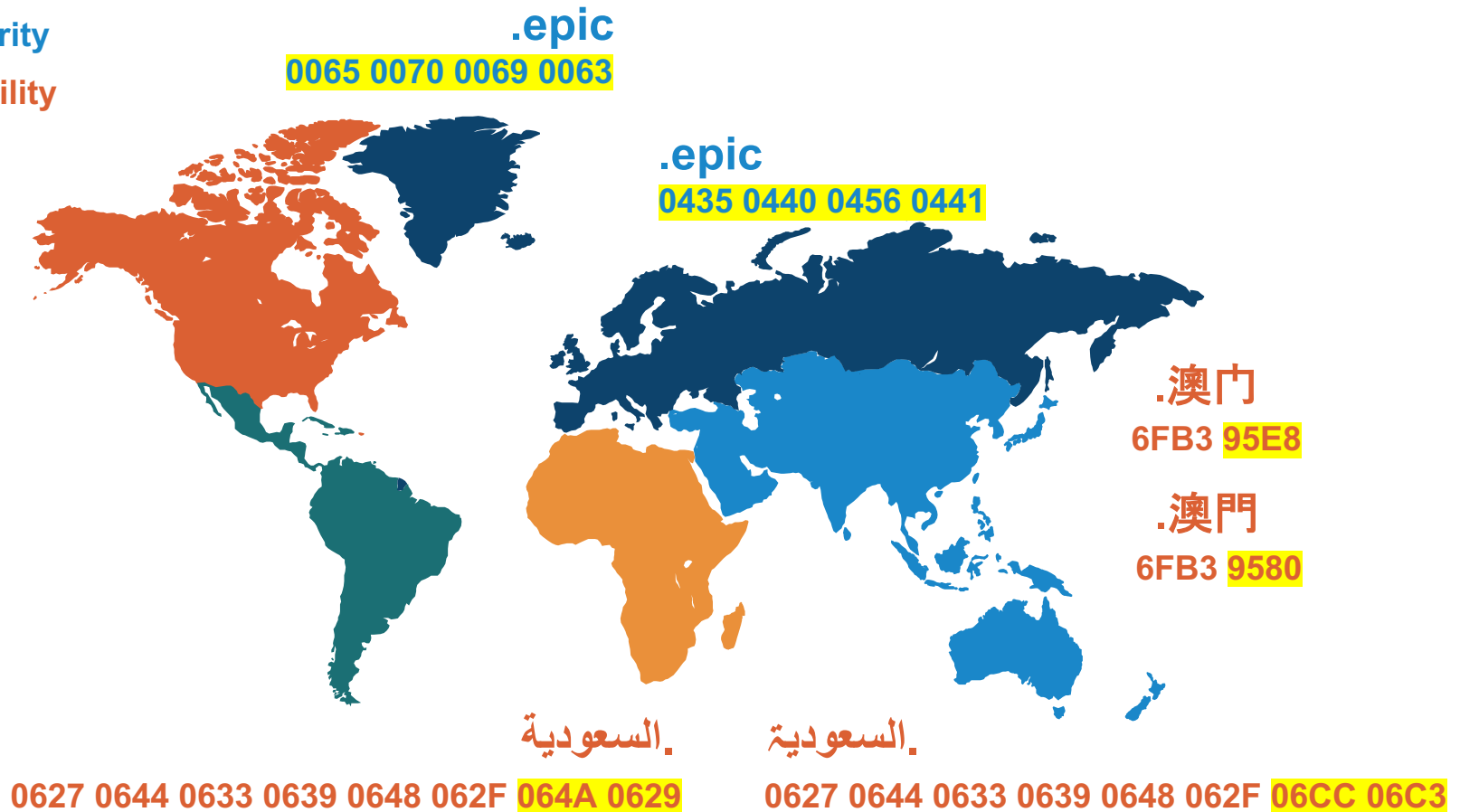


Work Done for the ‘Same’ Level

Variant labels are labels considered the “same” (*Visually or Semantically*) in some measure by a given community of Internet users.

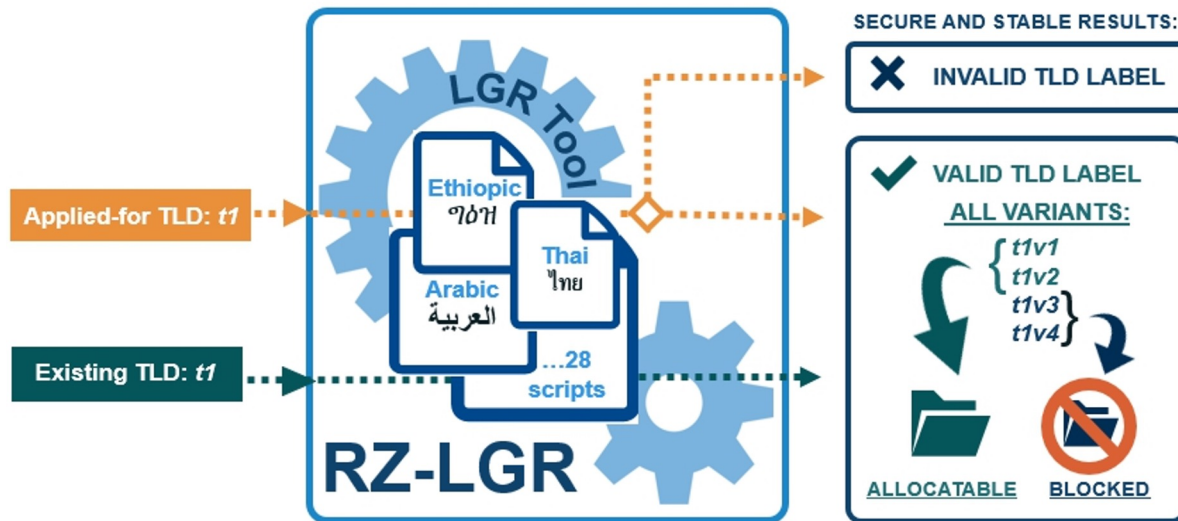
Security

Usability



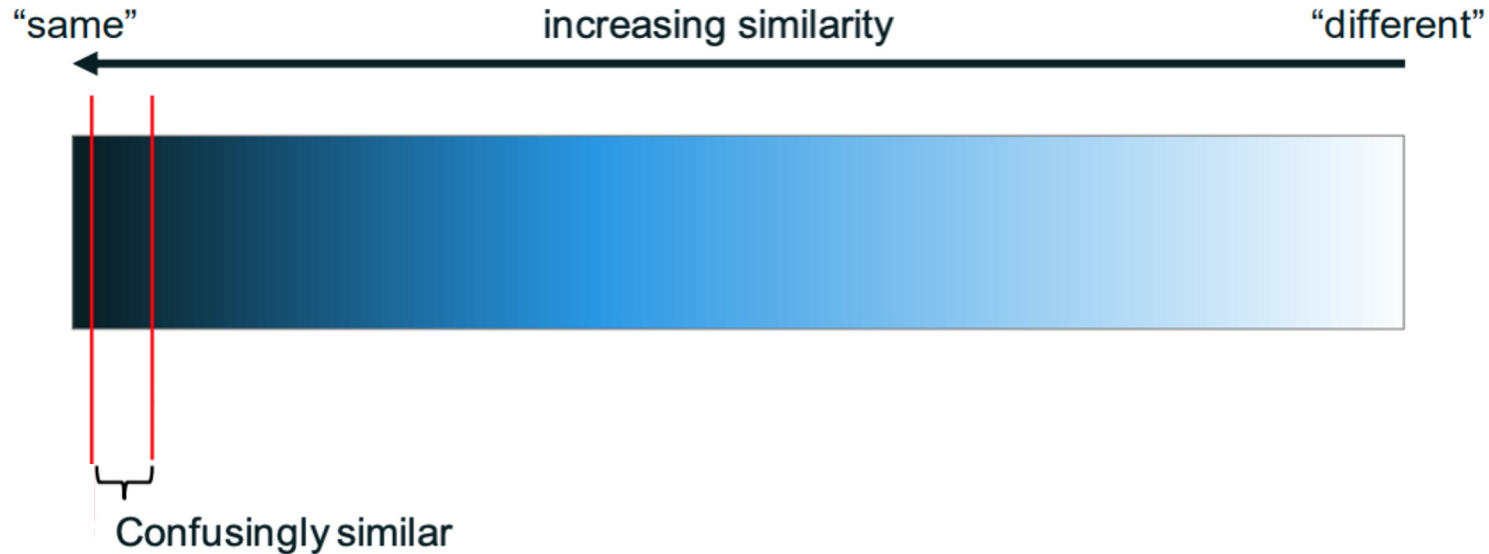
Variant Definitions for Top-Level and Second-Level

- Script communities define the variant within script and cross script in the Root Zone Label Generation Rules ([RZ-LGR](#)), 26 scripts are supported:
 - 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.



- Solutions in the RZ-LGR, including the variant definitions, are used to develop the Reference LGRs for the second-level.

Work in Progress for the ‘Confusing Similar’ Level



- For the top-level
 - Based on the 2012 new [gTLD Applicant Guidebook](#), the confusing similar strings are determined by the String Similarity Review process.
 - Based on the [IDN ccTLD Fast Track Process](#), the confusing similar strings are determined by the DNS Stability Panel.
- For the second-level
 - The [IDN Implementation Guidelines v.4.1](#) “TLD registries are encouraged to consider IDN policies to minimize confusion of IDN labels with other labels”

Work in Progress for the ‘Confusing Similar’ Level

- Additional Input
 - More work has been done by the community on analysis of visual similarity as part of the process for developing [RZ-LGR proposals and supporting documents](#). In many supporting documents for RZ-LGR, the community has suggested similar code points for consideration of the string similarity review process, in addition to the variant code points.
 - Additional work is also available through Unicode standard ([UTR#36](#) and [UTS#39](#)) and data (e.g., [confusables.txt](#)).
- Need to develop guidelines with the support of the community.

Q&A

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/company/icann



soundcloud/icann



instagram.com/icannorg