

Confusing String Similarity (CSS)

.IN IDN ccTLDs measures

Presented By:
Anil Jain, CEO NIXI
83 APTLD
22 February 2023

STRING SIMILARITY

“Visually similar looking strings to a common Netizen through naked eyes”

Some Examples of such cases

1. Paypal & Paypal(Paypai)
2. G00gle & G00gle
3. .hotels & .hotels (.hoteis)

POSSIBLE SCENERIOS OF STRING SIMILARITY

1. NO MATCH
2. MIS MATCH

FURTHER COMPLICATIONS

- STRING : ASCII / IDN
- VARIANTS
- STAGES OF DELEGATION
- RESERVED / BLOCKED STRINGS

Multilingual diversity of India

- Number of constitutionally recognized Languages – 22
- Number of language families – 4
 - Indo-Aryan
 - Dravidian
 - Tibeto-Burman
 - Munda
- Number of major families of scripts - 2
 - Brahmi – Left to Right writing system
 - Perso-Arabic - Right to left writing system
- 19 languages use 11 derivations of the Brahmi script.
- While 3 languages Urdu, Sindhi and Kashmiri uses the Perso-Arabic derivations.

.IN IDN ccTLDs policy

- **Fundamental Blocks:**

- Code point repertoire
- Whole Label Evaluation rules

Devanagari									
	090	091	092	093	094	095	096	097	
0	ॐ	ऐ	ठ	र	ी	ऊ	ऋ	ॠ	
1	ॡ	ऑ	ड	र	ॢ	ॣ	।	॥	
2	०	ओ	ढ	ल	॥	॥	॥	॥	अँ
3	०	ओ	ण	ळ	०	०	०	०	अँ

किताब कितााब किि्ताब

✗ ✗

- Variants (Confusing Similar Strings /Homographs) – **Points of discussion for this session**

अद्रक
अद्रक
अद्रक

Variant Cases

- **Group 1:** Confusing due to pure visual similarity

ਬ U+0918	ਖ U+0927
ਮ U+092D	ਮ U+092E
ਭੁਲ U+0B92 U+0BB3	ਭੁਲ U+0B94

ਫ਼ U+0926 U+094D U+0917	ਫ਼ U+0926 U+094D U+0930	ਫ਼ U+0926 U+094D U+0928
ਫ਼ U+0926 U+094D U+0927	ਫ਼ U+0926 U+094D U+0918	
ਫ਼ U+0937 U+094D U+091F	ਫ਼ U+0937 U+094D U+0920	
ਫ਼ U+0926 U+094D U+ 0935	ਫ਼ U+0926 U+094D U+092C	

Variant Cases

- **Group 2:** Confusing due to deviation from normally perceived character formations by larger linguistic community

आ U+0906	आ U+0906 U+093C
औ U+0913	औ U+0913 U+093C
। U+093E	। U+093E U+093C
ो U+094B	ो U+094B U+093C

च U+0931 U+094D U+092F	-य U+002D U+092F
ीं U+09C0 U+0981	ी U+09C0

EXISTING GUIDELINES

1. Technical Evaluation at the time of delegation.
2. Review by “Special Group” on demand.

Discussion is Progress in IDN ccTLD WG / IDN gTLD (EPDP)

Three Levels of Comparison



Some New Considerations

Inter-language similarity

- What happens when there is a similarity between different scripts – but the utility is vastly different – should there be review of usage of the domain before blocking or disqualifying?

Phonetic algorithms for matching string similarity

- What about similar sounding words when the future is looking at audio inputs and not text? Are we ready with phonetic based algorithms to handle such situations?

धन्यवाद !