

Experience of variant handling in IDN ccTLD پاکستان . Registry

Syed Iftikhar H. Shah, Pakistan



Overview

- Background
- Pakistan's Registry - String
- Language Table
- Issues regarding IDN ccTLD Script
- Confusability of Characters
- Decision about Confusability of Characters

Background



- Pakistani people are comfortable in using their mother tongue like Urdu instead of other languages.
- There is a dire need to increase the Internet users.
- A source for the promotion of local contents development.

Background

- To achieve these objectives the first step is that the internet address should be written in Urdu and other Pakistani languages (Balochi, Pashto, Punjabi, Saraiki, Sindhi, Torwali etc) for example www.Dictionary.org.pk can be
ووو لغت ادارہ پاکستان
- Government of Pakistan (GoP) will become Internationalized Domain Names (IDNs) country code Top Level Domain (ccTLD) manager for running the registry پاکستان



Pakistan's Registry- IDN Variants



Pakistan's Registry String

- **String 1**

Primary string

ISO3166 Entry: PK (PAKISTAN)

A Label : xn--mgbai9azgqp6j

U Label: پاکستان

Unicode code points: U+067E, U+0627,
U+06A9, U+0633, U+062A, U+0627,
U+0646

String in English: Pakistan



Pakistan's Registry String

- **String 2**

This is the desired variant string of String 1.

Request to be delegated

ISO3166 Entry: PK (PAKISTAN)

A Label : xn--mgbai9a5eva00b

U Label: پاکستان

Unicode code points: U+067E, U+0627,
U+0643, U+0633, U+062A, U+0627,
U+0646

String in English: Pakistan



Pakistan's Registry String

- **String 3**

This is an undesired variant string of String 1.
Request to be blocked.

ISO3166 Entry: PK (PAKISTAN)

A Label : xn--mgba9aze4mva61a

U Label: پاکستان

Unicode code points: U+067E, U+0627,
U+06A9, U+0633, U+067A, U+0627, U+0646

String in English : Pakistan



Pakistan's Registry String

- **String 4**

This is an undesired variant string of String 1.
Request to be blocked.

ISO3166 Entry: PK (PAKISTAN)

A Label : xn--mgba9aydr42aya

U Label: پاکستان

Unicode code points: U+067E, U+0627,
U+0643, U+0633, U+067A, U+0627, U+0646

String in English: Pakistan



Language Table

- Work Started on April, 2008 and completed in March, 2010
- It is a single language table will be developed to support major languages spoken in the country.
- It is developed with the consent of from language experts

Combining Characters (Diacritics and Honorifics) Currently NOT Allowed in IDNs for Pakistani Languages
 May be allowed at a later stage after formal consideration

	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F	075	076	077
0				ذ	-		٠		پ	ت	ث	گ	ة	ي		٠	ي	و	ث
1			ء	ر	ف		١	أ	خ	ز	و	گی	ہ	ي		١	ث	و	ث
2			آ	ز	ق		٢	أ	خ	ز	و	گی	ء	ل		٢	ي	ن	خ
3			أ	س	ك		٣	أ	ج	ر	ف	گی	ت	ل		٣	ي	ن	خ
4			ؤ	ش	ل		٤	ء	ج	ر	ف	گی	و	-		٤	ي	ن	خ
5			إ	ص	م		٥	أ	خ	ر	و	ل	و	ه		٥	ي	ن	خ
6			ئ	ض	ن		٦	ؤ	ج	ر	ف	ل	ؤ		٦	٦	٦	٦	٦
7			ا	ط	ه		٧	ؤ	ج	ز	ف	ل	ؤ		٧	٧	٧	٧	٧
8			ب	ظ	و		٨	ئ	ت	ز	ف	ل	ؤ		٨	٨	٨	٨	٨
9			ة	ع	ي		٩	ئ	ت	ز	ف	ل	ؤ			٩	٩	٩	٩
A			ث	غ	ي			ئ	ب	ب	ل	ق			ث	د	د	د	د
B			ث	ن			ر	ب	ب	پ	ل	ن	ؤ			ن	ن	ن	ن
C			ج	ک			ر	ن	ت	پ	ن	ن	ی			ن	ن	ن	ن
D			ح	ی			*	ئ	د	پ	ن	ن	ی			ن	ن	ن	ن
E			خ	ت			ر	پ	ت	ص	ل	ه	ی			ن	ن	ن	ن
F			د	ث			و	ث	ت	ظ	گ	ن	و			ن	ن	ن	ن



Issues regarding IDN ccTLD Script

- Confusability of characters
- Additional composed characters
- Digits and Mixing
- Character Separator
- Label separator



Confusability

- Visually similar character shapes create confusion
- Confusion can be due to initial, medial, final or isolated forms
- Different cases of confusability
 - ❖ Shape confusability
 - ❖ Exact shape confusion
 - ❖ Similar shape confusion
- Composition confusability



Similar Shape Confusion

- Urdu character ع (06CC) and Pashto character ع (06CD)
- Sindhi ک (06AA) and Urdu ک (06A9)
 - ک vs. ک



Exact Shape Confusion

- ک + ل = کل looks same as ک + ل = کل
- چ + ل + ی (06CC) = چلی looks same as
چ + ل + ی (0649) = چلی
- ی (06CC) + ا = یا looks same as ي
(064A) + ا = یا



Composition Confusability



- There are characters that can be typed in more than one ways
 - U+0622 (ġ) =
U+0627 (ġ) + U+0653 (̄)
- Although they look similar to the user, they translate to different ASCII codes

Contextual Shapes of Different Letters

Isolated	Initial	Medial	Final
ب	با	کبا	کب
چ	چا	کچا	کچ
و	NA	NA	کو

URDU HINDI UNIVERSITY OF TECHNOLOGIES

Composed vs. Decomposed Form

Composed Form	Decomposed Form	
U+0622 (آ)	U+0627 (ا) + U+0653	
U+0623 (إ)	U+0627 (ا) + U+0654	
U+0624 (ؤ)	U+0648 (و) + U+0654	
U+0625 (أ)	U+0627 (ا) + U+0655	
U+0626 (ئ)	U+064A (ي) + U+0654	
U+0675 (آ)	U+0627 (ا) + U+0674	



Decision about Confusability

- It has been decided that variant characters should be mapped to a single character to avoid user confusion.
- Variant characters are divided into three categories based on what makes them confusingly similar to each other.
- The following tables provide the decisions taken by all language communities regarding specific cases of confusable characters.

1. Variants Based on Shape Similarity

Character	Unicode	Character	Unicode	Mapping Decision
ي	064A	ى	06CC	Map in initial and medial positions
ى	06CD	ى	06CC	Map in final and isolated positions
ه	0647	ه	06BE	Map in all four positions
ه	0647	ه	06C1	Map in all four positions
ك	06A9	ك	06AA	Map in all four positions
ط	0679	ط	06BB	Map in initial and medial positions

1. Variants Based on Shape Similarity (Cont..)

Character	Unicode	Character	Unicode	Mapping Decision
۰	06F0	0	0030	Yes
۱	06F1	1	0031	Yes
۲	06F2	2	0032	Yes
۳	06F3	3	0033	Yes
۴	06F4	4	0034	Yes
۵	06F5	5	0035	Yes
۶	06F6	6	0036	Yes
۷	06F7	7	0037	Yes
۸	06F8	8	0038	Yes
۹	06F9	9	0039	Yes

2. Variants Based on Confusability with Arabic Language Characters

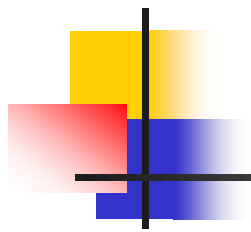
Character	Unicode	Character	Unicode	Mapping Decision
ة	0629	ة	06C3	Map during resolution
ك	06A9	ك	0643	Map during resolution
ك	06AA	ك	0643	Map during resolution
ى	06CC	ى	0649	Map during resolution
ي	064A	ى	0649	Map during resolution
ى	06CD	ى	0649	Map during resolution

2. Variants Based on Confusability with Arabic Language Characters (Cont..)

Character	Unicode	Character	Unicode	Mapping Decision
٠	06F0	٠	0660	Map during resolution
١	06F1	١	0661	Map during resolution
٢	06F2	٢	0662	Map during resolution
٣	06F3	٣	0663	Map during resolution
٤	06F4	٤	0664	Map during resolution
٥	06F5	٥	0665	Map during resolution
٦	06F6	٦	0666	Map during resolution
٧	06F7	٧	0667	Map during resolution
٨	06F8	٨	0668	Map during resolution
٩	06F9	٩	0669	Map during resolution

3. Variants Based on Orientation of Dots

Character	Unicode	Character	Unicode	Mapping Decision
چ	0683	چ	0684	No. But reserve label with confusable character until further analysis
ت	062A	ت	067A	No. But reserve label with confusable character until further analysis
ث	062B	ث	067D	No. But reserve label with confusable character until further analysis
ي	064A	ي	06D0	No. But reserve label with confusable character until further analysis



Thanks