सत्यमेव जयते

Department of Information Technology,
Ministry of Communications and Information Technology,
Government of India, New Delhi

Draft Policy Document

for

INTERNATIONALIZED DOMAIN NAMES

Language: URDU

RECORD OF CHANGES

***A - ADDED M - MODIFIED D - DELETED**

VERSION NUMBER	DATE	POINTS AFFECTED	**A* M D**	TITLE OR BRIEF DESCRIPTION	COMPLIANCE VERSION OF MAIN POLICY DOCUMENT
1.0	29 September 2014	All		Final Policy document	1.0

Table of Contents

1. PERSO-ARABIC SCRIPTS: GENERAL INTRODUCTION

1.1. OVERVIEW

Three languages in India use the Perso-Arabic script. These are Urdu, Sindhi and Kashmiri[1].

Unlike Brahmi derived languages which are abugidas i.e. syllable driven, Perso-Arabic driven languages are abjads i.e. character based. The concept of the ISCII syllable has therefore no pertinence insofar as languages derived from the Perso-Arabic script are concerned. Therefore, unlike Hindi or Tamil for example, Urdu has no Augmented Backus Naur Formalism (ABNF). However Urdu does admit restriction rules as given in Section 5 below. The template for Perso-Arabic derived languages admits only the Code-chart with the pertinent characters marked in yellow, the corresponding nomenclatural table as well as the variant list.

1.2. GENERAL STRATEGY FOR URDU

Of all the Indian languages, the Perso-Arabic script represents the greatest amount of difficulties and also chances of spoofing and phishing. This is because of the intrinsic nature of the script which has a large degree of homographs and also the fact that Unicode code block (U+0600 – U+06FF) caters to a large number of languages and there is a large degree of resemblance between two or more characters.

To simplify the problem and ensure that as far as possible spoofing and phishing will be reduced to a bare minimum, the following strategy is proposed:

1.2.1. MAPPING IN CONSONANCE WITH THE POLICY LAID DOWN BY GOVT. OF INDIA

- **www** will always remain in English. It is the Middle layer and the ccTLD which will remain in Urdu.
- It is assumed that the Bidi algorithm built into the browser used should handle the directionality of English and Urdu efficiently.
- The ccTLD used will be a suitable equivalent of “.in” in Urdu. The translation of India into Urdu shall be بـهـارت
- The character set prescribed for Urdu will be IDNA 2008 compliant.

[1] Sindhi and Kashmiri are also written in the Devanagari script.

- The number of permissible characters shall not exceed 63 when converted to Punycode (inclusive of ACE Prefix).
- Script vs. Language: Unicode Code Block (U+0600 – U+06FF) caters to a large number of languages. Only the pertinent character set for Urdu shall be used.
- No mixing of two languages will be allowed with in the domain label inside the zones .
- The Latin full-stop shall be used instead of the corresponding URDU punctuation marker.
- All digits will be the International Digit Set i.e. 0,1,2,3,4,5,6,7,8,9 and not the Arabic-Indic digit set as prescribed in the Code-page for Arabic.
- Similarly English Hyphen will be used and not the corresponding Urdu Hyphen.
- ZWJ and ZWNJ shall not be permitted.
- Space (A major issue in Perso-Arabic scripts) shall not be permitted within the domain name.

1.2.2. DIRECTIVE PRINCIPLES SPECIFIC TO URDU:

PRINCIPLE I: *The permissible Character Set*

The Urdu code-set will be defined and isolated from the Arabic page i.e. only those characters which are permissible in Urdu will be retained. Since Unicode Code Block (U+0600 – U+06FF) is highly liable to spoofing, the choice of the character-set pertinent to Urdu alone will reduce spoofing and phishing.

PRINCIPLE II: *Identification of Characters liable to Spoofing.*

Characters liable to cause spoofing shall be identified and treated as variants. These will also include normalization.

PRINCIPLE III: *Diacritics reduced to a bare minimum*

As far as possible, all diacritics will be eliminated from the set. Only the most important and pertinent diacritics shall be retained. These are:

(i) ARABIC MADDAH ABOVE (0653 ◌ٓ)
(ii) ARABIC HAMZA ABOVE (0654 ◌ٔ)
(iii) ARABIC HAMZA BELOW (0655 ◌ٕ)
(iv) ARABIC SHADDA (0651 ◌ّ)
(v) ARABIC SUBSCRIPT ALEF (0656 ◌ٖ)
(vi) ARABIC LETTER SUPERSCRIPT ALEF (0670 ◌ٰ)

Alif, Madd and Hamza Characters most frequently used in Urdu are as under and these will be admitted to the permissible set.

ۓ ؤ ئ ۀ أ إ آ

Their corresponding combinations shall be treated as variants. Thus (0622 آ) can also be entered as (0627 ا) followed by (0653 ٓ) in some Urdu keyboards and it is to resolve this alternative mode of entry that such as normalization is permitted in the shape of a variant.

PRINCIPLE IV: *EZAFAT*

A serious issue will be that of the ezafat in words such as *Yaad-e-Khuda* or *Aab-o-Hawa*. As a palliative suggestion, it is suggested that the ezafat be represented by:

(i) ARABIC LETTER YEH BARREE U+06D2 ے
(ii) ARABIC LETTER WAW U+0648 و
(iii) ARABIC LETTER HAMZA U+0621 ء

Separated by a hyphen as in the examples below:

ے	یاد-ے-خدا
و	آب-و-ہوا

PRINCIPLE V: *Visual Identity of the Word: The case of Space between two words within a URL.*

Since a large number of characters in Perso-Arabic can join together unless separated by a Space, Space is a cardinal issue in all Perso-Arabic driven languages. Space ensures visual identity. Since Space is not permissible within a URL, visual identity where two words constitute a URL constitutes a major issue.

A palliative to this issue would be the use of the hyphen to separate two words and thereby ensure legibility.

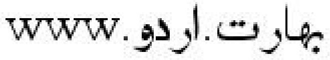
Thus in the case of a site for a mango pickle: ***aam aachaar*** which when written together would be illegible.

آمآچار

The solution would be to separate out the two words with a hyphen as shown below.

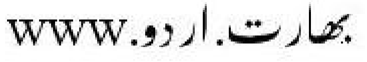
آم- آچار

PRINCIPLE VI: *Use of Naskh instead of Nastalique in the URL*
Naskh is more visually clear and reduces also spoofing and pharming because of clear legibility of the joining characters as is shown below:



www.بھارت.اردو

Naskh



www.بھارت.اردو

Nastalique

2. RESTRICTION RULES

Urdu admits following restriction rules:

1. ARABIC MADDAH ABOVE U+0653 ◌ٓ shall be allowed only after the following character.
   (a) ARABIC LETTER ALEF U+0627 ا
2. ARABIC HAMZA ABOVE U+0654 ◌ٔ shall be allowed only after the following characters.
   (a) ARABIC LETTER ALEF U+0627 ا
   (b) ARABIC LETTER WAW U+0648 و
   (c) ARABIC LETTER HEH GOAL U+06C1 ہ
   (d) ARABIC LETTER YEH BARREE U+06D2 ے
   (e) ARABIC LETTER FARSI YEH U+06CC ی

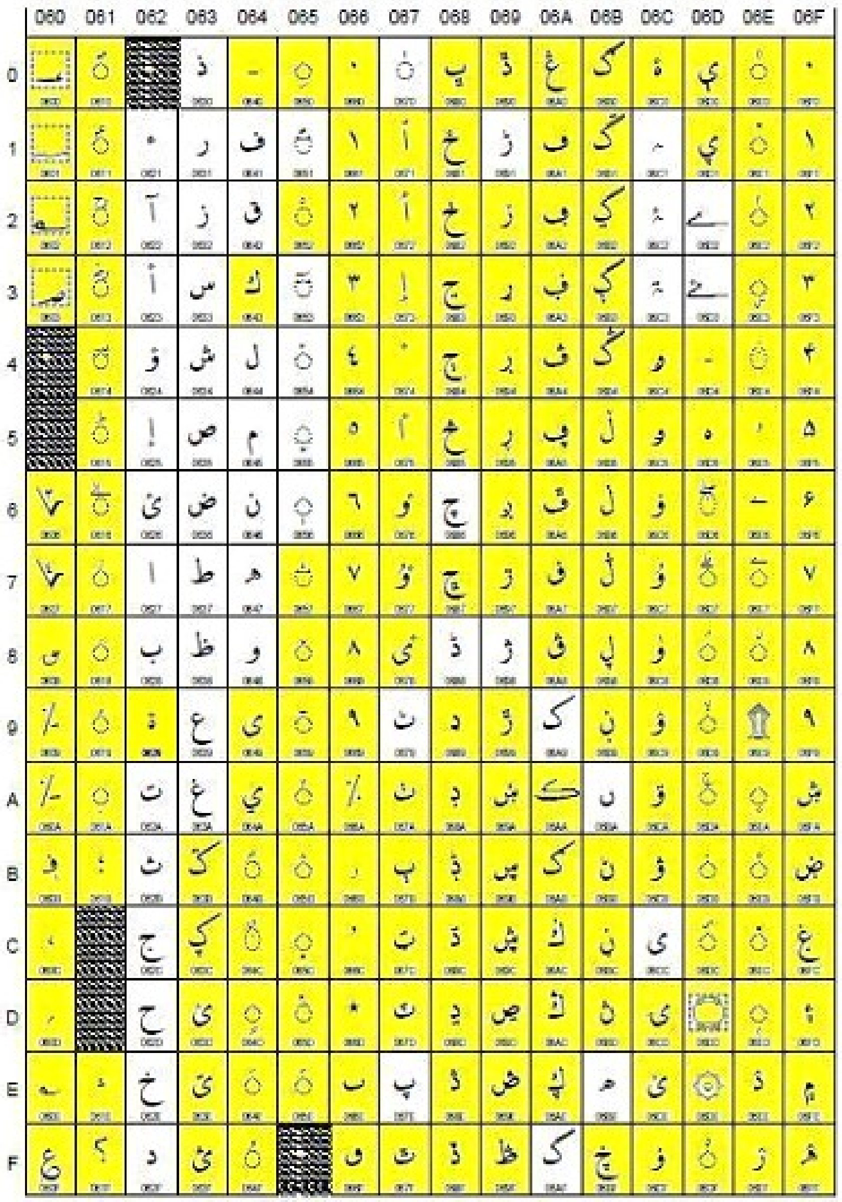
3. ARABIC HAMZA BELOW U+0655 ◌ٕ shall be allowed only after the following character.
   (a) ARABIC LETTER ALEF U+0627 ا
4. Apart from permissible single diacritics, only the below combinations of two diacritics are allowed-
   (a) ARABIC SHADDA U+0651◌ّ followed by ARABIC SUBSCRIPT ALEF U+0656◌ٖ .
   (b) ARABIC SHADDA U+0651 ◌ّ followed by ARABIC LETTER SUPERSCRIPT ALEF U+0670 ◌ٰ.

5. Consecutive Hyphens will not be permitted in a domain name.

6. A label containing more than three instances of variant character(s) will not be permitted. As an example let us consider a, b, c and d as four variants in a given label having a', b', c' and d' as variants in which case such a label will be disallowed. (E.g. of disallowed label - abcd, acdb, cdaba and so on)

Additional Note:
Wherever a variant is present in a given label, the variants shall be strictly symmetric and non-transitive. Thus given some variants ۂ (U+06C2) ⇔ ۂ (U+06C1+U+0654) and ہ(U+06C1) ⇔ ۃ (U+06C3). One of the variants of a label such as طرۂ shall be طرۂ. طرۃٔ generated by adding an extra ۃ(U+06C3) to ہ(U+06C1) shall not be permitted. This ensures that over generativity does not take place.

3. LANGUAGE TABLE[2]: URDU[3]



	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F
0	؀	◌ؐ		ذ	ـ	◌ِ	٠	◌ٰ	ڀ	ڐ	ڠ	ڰ	ۀ	ې	◌۠	۰
1	؁	◌ؑ	ء	ر	ف	◌ّ	١	ٱ	ځ	ڑ	ڡ	ڱ	ہ	ۑ	◌ۡ	۱
2	؂	◌ؒ	آ	ز	ق	◌ْ	٢	ٲ	ڂ	ڒ	ڢ	ڲ	ۂ	ے	◌ۢ	۲
3	؃	◌ؓ	أ	س	ك	◌ٓ	٣	ٳ	ڃ	ړ	ڣ	ڳ	ۃ	ۓ	◌ۣ	۳
4		◌ؔ	ؤ	ش	ل	◌ٔ	٤	ٴ	ڄ	ڔ	ڤ	ڴ	ۄ	۔	◌ۤ	۴
5		◌ؕ	إ	ص	م	◌ٕ	٥	ٵ	څ	ڕ	ڥ	ڵ	ۅ	ە	ۥ	۵
6	؆	◌ؖ	ئ	ض	ن	◌ٖ	٦	ٶ	چ	ږ	ڦ	ڶ	ۆ	◌ۖ	ۦ	۶
7	؇	◌ؗ	ا	ط	ه	◌ٗ	٧	ٷ	ڇ	ڗ	ڧ	ڷ	ۇ	◌ۗ	◌ۧ	۷
8	؈	◌ؘ	ب	ظ	و	◌٘	٨	ٸ	ڈ	ژ	ڨ	ڸ	ۈ	◌ۘ	◌ۨ	۸
9	؉	◌ؙ	ة	ع	ى	◌ٙ	٩	ٹ	ډ	ڙ	ک	ڹ	ۉ	◌ۙ	۩	۹
A	؊	◌ؚ	ت	غ	ي	◌ٚ	٪	ٺ	ڊ	ښ	ڪ	ں	ۊ	◌ۚ	◌۪	ۺ
B	؋	؛	ث	ػ	◌ً	◌ٛ	٫	ٻ	ڋ	ڛ	ګ	ڻ	ۋ	◌ۛ	◌۫	ۻ
C	،		ج	ؼ	◌ٌ	◌ٜ	٬	ټ	ڌ	ڜ	ڬ	ڼ	ی	◌ۜ	◌۬	ۼ
D	؍		ح	ؽ	◌ٍ	◌ٝ	٭	ٽ	ڍ	ڝ	ڭ	ڽ	ۍ	۝	◌ۭ	۽
E	؎	؞	خ	ؾ	◌َ	◌ٞ	ٮ	پ	ڎ	ڞ	ڮ	ھ	ێ	۞	ۮ	۾
F	؏	؟	د	ؿ	◌ُ		ٯ	ٿ	ڏ	ڟ	گ	ڿ	ۏ	◌۟	ۯ	ۿ

[2] This language table is based on Unicode Chart for Arabic script provided by the Unicode Consortium

[3] Characters marked in yellow are not applicable to the language.

4. NOMENCLATURAL DESCRIPTION TABLE OF URDU LANGUAGE TABLE

The following are basic alphabetic characters for Urdu, and will therefore be allowed.

PERMISSIBLE URDU CHARACTER SET

0621	ء	ARABIC LETTER HAMZA
0627	ا	ARABIC LETTER ALEF
0628	ب	ARABIC LETTER BEH
062A	ت	ARABIC LETTER TEH
062B	ث	ARABIC LETTER THEH
062C	ج	ARABIC LETTER JEEM
062D	ح	ARABIC LETTER HAH
062E	خ	ARABIC LETTER KHAH
062F	د	ARABIC LETTER DAL
0630	ذ	ARABIC LETTER THAL
0631	ر	ARABIC LETTER REH
0632	ز	ARABIC LETTER ZAIN
0633	س	ARABIC LETTER SEEN
0634	ش	ARABIC LETTER SHEEN
0635	ص	ARABIC LETTER SAD
0636	ض	ARABIC LETTER DAD
0637	ط	ARABIC LETTER TAH
0638	ظ	ARABIC LETTER ZAH
0639	ع	ARABIC LETTER AIN
063A	غ	ARABIC LETTER GHAIN
0641	ف	ARABIC LETTER FEH

0642	ق	ARABIC LETTER QAF
0644	ل	ARABIC LETTER LAM
0645	م	ARABIC LETTER MEEM
0646	ن	ARABIC LETTER NOON
0647	ه	ARABIC LETTER HEH
0648	و	ARABIC LETTER WAW
0679	ٹ	ARABIC LETTER TTEH
067E	پ	ARABIC LETTER PEH
0686	چ	ARABIC LETTER TCHEH
0688	ڈ	ARABIC LETTER DDAL
0691	ڑ	ARABIC LETTER RREH
0698	ژ	ARABIC LETTER JEH
06A9	ک	ARABIC LETTER KEHEH
06AF	گ	ARABIC LETTER GAF
06BA	ں	ARABIC LETTER NOON GHUNNA
06BE	ھ	ARABIC LETTER HEH DOACHASHMEE
06C1	ہ	ARABIC LETTER HEH GOAL
06C3	ۃ	ARABIC LETTER TEH MARBUTA GOAL
06CC	ی	ARABIC LETTER FARSI YEH
06D2	ے	ARABIC LETTER YEH BARREE

The following combinations of base character and diacritic will also be allowed:

0622	آ	ARABIC LETTER ALEF WITH MADDA ABOVE
0623	أ	ARABIC LETTER ALEF WITH HAMZA ABOVE
0624	ؤ	ARABIC LETTER WAW WITH HAMZA ABOVE

0625	إ	ARABIC LETTER ALEF WITH HAMZA BELOW
0626	ئ	ARABIC LETTER YEH WITH HAMZA ABOVE
06C2	ۂ	ARABIC LETTER HEH GOAL WITH HAMZA ABOVE
06D3	ۓ	ARABIC LETTER YEH BARREE WITH HAMZA ABOVE

Apart from above set of characters, the following diacritics are also allowed:

0651	◌ّ	ARABIC SHADDA
0653	◌ٓ	ARABIC MADDAH ABOVE
0654	◌ٔ	ARABIC HAMZA ABOVE
0655	◌ٕ	ARABIC HAMZA BELOW
0656	◌ٖ	ARABIC SUBSCRIPT ALEF
0670	◌ٰ	ARABIC LETTER SUPERSCRIPT ALEF

5. VARIANT TABLE FOR URDU

The following variants are based on a single character combination which can be also entered as a combination of two characters. It should be noted that these variants have been admitted to accommodate keyboards where a single character representing a combination such as *alif madd* آ is not available and the user has to enter alif and madd separately.

VARIANTS	
ں 06BA	ن 0646
ہ 06C1	ۃ 06C3
آ 0622	آ 0627 + 0653
أ 0623	أ 0627 + 0654
ؤ 0624	ؤ 0648 + 0654
إ 0625	إ 0627 + 0655
ئ 0626	یٔ 06CC + 0654
ۂ 06C2	ۂ 06C1 + 0654
ۓ 06D3	ۓ 06D2 + 0654

Caveats

- Other characters distinguished by a single Nukta such as suad ~ zuad have not been included, since this would have made the attribution of URL's too restrictive.

- All other cases are handled by the exclusive character set for Urdu and absence of diacritics.

6. EXPERTISE/BODIES CONSULTED

Expertise provided by experts of Urdu language and Urdu computational Linguistics of Osmania University and Maulana Azad National Urdu University.

7. PROPOSED ccTLD FOR URDU

India (Bhārat) localized in Urdu - بھارت

Note: You can send your feedbacks to idn-feedback@cdac.in