



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
1.0.0	1/06/2010	Whole Document	A	Language Specific Policy Document for URDU	
1.0.1.	2/06/2010	1.2.1. 1.2.2.	M	1.2.1-point 5 1.2.2- alif with hamza below missing	
1.0.2	23/09/2010	1.1	M	Point 5 has been added for restriction rules	
1.0.3	14/10/2010	Points 2,3,4 and 5	M	Point 2,3,4 and 5 have been modified in tabular form for compliance with other language policy documents.	
1.0.4	20/10/2010	4,5,6,7,9,11, 12,13	M	Modified according to testing report of first round	
1.0.5	26/10/2010	Title has been changed, and some formatting issues	M	Text in all tables have been changed to center aligned and formatting issues have been fixed	
1.0.6	01/02/11	Whole Document	M	Document has been changed for IDN Urdu font Restriction rule added	
1.0.7	05/04/11	2	A	Added Restriction Rule for Single Permissible Diacritics	

Table of Contents

1. PERSO-ARABIC SCRIPTS: GENERAL INTRODUCTION.....	4
1.1.OVERVIEW.....	4
1.2. GENERAL STRATEGY FOR URDU.....	4
1.2.1. MAPPING IN CONSONANCE WITH THE POLICY LAID DOWN BY GOVT. OF INDIA.....	4
1.2.2. DIRECTIVE PRINCIPLES SPECIFIC TO URDU:.....	5
2. RESTRICTION RULES.....	9
3. LANGUAGE TABLE : URDU.....	11
4. NOMENCLATURAL DESCRIPTION TABLE OF URDU LANGUAGE TABLE... ..	12
5. VARIANT TABLE FOR URDU.....	16
6. EXPERTISE/BODIES CONSULTED.....	18

1. PERSO-ARABIC SCRIPTS: GENERAL INTRODUCTION

1.1.OVERVIEW

Three languages in India use the Perso-Arabic script. These are Urdu, Sindhi and Kashmiri¹.

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 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 Code-page 600 caters to a large number of scripts 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, pharming 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 problem of English and Urdu efficiently.
- The CCTLD or top layer will be .in suitably transliterated to Urdu. The translation of India into Urdu shall be **بھارت**
- The set prescribed for Urdu will be Unicode 5.1 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 the header)
- Script vs. Language: Code page 600 caters to a large number of languages. Only the pertinent character set for Urdu shall be used.
- No mixing of two languages will allowed inside the ccTLD
- The Latin full-stop shall be used instead of the corresponding URDU punctuation marker.
- All digits will be the Latin Digit Set i.e. 0,1,2,3,4,5,6,7,8,9 and not the Arabic 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 URL.

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 Code-page 600 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 MADDAAH ABOVE (0653 ^ﷰ)

(ii) ARABIC HAMZA ABOVE (0654 ^ء)

(iii) ARABIC HAMZA BELOW (0655 _ء)

(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.

آ اِ اُ ءِ ؤِ وِ

Their corresponding combinations shall be treated as variants. Thus (0622آ) can also be entered as (0627 اِ) followed by (0653 ~) in some Urdu keyboards and it is to resolve this alternative mode of entry that such 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 06D2

(ii) ARABIC LETTER WAW 0648

(iii) ARABIC LETTER HAMZA 0621

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.

آماچار

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:



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

www.بھارت.اردو

Naskh

www.بھارت.اردو

Nastalique

2. RESTRICTION RULES

Urdu admits the following restriction rules:

1. ARABIC MADDAH ABOVE 0653 [~] shall be allowed only after the following character.

(a) ARABIC LETTER ALEF 0627 ا

2. ARABIC HAMZA ABOVE 0654 ^ء shall be allowed only after the following characters.

(a) ARABIC LETTER ALEF 0627 ا

(b) ARABIC LETTER WAW 0648 و

(c) ARABIC LETTER HEH GOAL 06C1 ه

(d) ARABIC LETTER YEH BARREE 06D2 ع

(e) ARABIC LETTER FARSI YEH 06CC ی

3. ARABIC HAMZA BELOW 0655 _ء shall be allowed only after the following characters.

(a) ARABIC LETTER ALEF 0627 ا

4. Apart from permissible single diacritics, only the below combinations of two diacritics are allowed-

(a) ARABIC SHADDA 0651 ˆ followed by ARABIC SUBSCRIPT ALEF 0656 ِ

(b) ARABIC SHADDA 0651 ˆ followed by ARABIC LETTER SUPERScript ALEF 0670 ٱ

5. Permissible single diacritics in the table below cannot occur in the beginning of the Internationalized Domain Name.

Unicode Value	Character	Character Name
0651	ﷰ	ARABIC SHADDA
0653	ﷲ	ARABIC MADDAAH ABOVE
0654	ﷴ	ARABIC HAMZA ABOVE
0655	ﷵ	ARABIC HAMZA BELOW
0656	ﷶ	ARABIC SUBSCRIPT ALEF
0670	ﷰ	ARABIC LETTER SUPERSCRIPT ALEF



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

3. LANGUAGE TABLE : URDU²

	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F
0	ا	آ	ب	پ	ت	ٹ	ث	ج	چ	ح	خ	گ	گھ	ن	ہ	و
1	ک	کھ	د	دھ	ر	ڑ	ز	ذ	ڈ	ڈھ	ڙ	ڙھ	ڻ	ڻھ	ڻھ	ڻھ
2	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
3	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
4	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
5	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
6	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
7	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
8	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
9	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
A	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
B	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
C	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
D	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
E	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ
F	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ	ڻھ

² 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

Unicode Value	Character	Character Name
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



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

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



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

067E	پ	ARABIC LETTER PEH
0686	چ	ARABIC LETTER TCHEH
0688	ط	ARABIC LETTER DDAL
0691	ر	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	ة	ARABIC LETTER TEH MARBUTA GOAL
06CC	ی	ARABIC LETTER FARSI YEH
06D2	آ	ARABIC LETTER YEH BARREE

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

Unicode Value	Character	Character Name
0622	آ	ARABIC LETTER ALEF WITH MADDA ABOVE

0623	اَ	ARABIC LETTER ALEF WITH HAMZA ABOVE
0624	وَ	ARABIC LETTER WAW WITH HAMZA ABOVE
0625	اِ	ARABIC LETTER ALEF WITH HAMZA BELOW
0626	يَ	ARABIC LETTER YEH WITH HAMZA ABOVE
06C2	هَ	ARABIC LETTER HEH GOAL WITH HAMZA ABOVE
06D3	عَ	ARABIC LETTER YEH BARREE WITH HAMZA ABOVE

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

Unicode Value	Character	Character Name
0651	ّ	ARABIC SHADDA
0653	َ	ARABIC MADDAAH ABOVE
0654	ءَ	ARABIC HAMZA ABOVE
0655	ءِ	ARABIC HAMZA BELOW
0656	اِ	ARABIC SUBSCRIPT ALEF
0670	اِ	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

ں 06BA	ن 0646
ہ 06C1	ۛ 06C3
آ 0622	ا + ّ 0627 + 0653
اُ 0623	ا + ُ 0627 + 0654
ۛ 0624	و + ُ 0648 + 0654
اِ 0625	ا + ِ 0627 + 0655
ئ 0626	ی + ُ 06CC + 0654



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

ہ 06C2	ہ + ء 06C1 + 0654
ے 06D3	ے + ء 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.

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