# Policy Document
# For

# INTERNATIONALIZED DOMAIN NAMES

# Language: KASHMIRI
# (Perso-Arabic)

CDAC
Gist

niXi

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

# 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 abujidas 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, Kashmiri has no Augmented Backus Naur Formalism (ABNF). However Kashmiri 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 KASHMIRI*

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+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, following principles have been laid down:

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

- HTTP://www. will always remain in English. The domain names and  ccTLD which will remain in KASHMIRI.
- It is assumed that the Bidi algorithm built into the browser used should handle the problem of English and KASHMIRI efficiently.
- Script vs. Language: Unicode Code block U+600 caters to a large number of languages. Only the pertinent character set for KASHMIRI shall be used.

---

1    Kashmiri is also written in the Devanagari script.

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

- No mixing of two languages will allowed inside the ccTLD
- Punctuation markers of Code-page KASHMIRI shall NOT be permitted. Thus, the Latin full-stop shall be used instead of the corresponding KASHMIRI punctuation marker.
- Similarly English Hyphen will be used and not the corresponding KASHMIRI hyphen.
- All digits will be from the Latin Set i.e. 0,1,2,3,4,5,6,7,8,9 and not the Indo-Arabic digit set as prescribed in the Code-page for Arabic.
- 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 KASHMIRI:

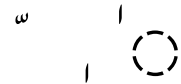1. PRINCIPLE I: *The permissible Character Set*
   The KASHMIRI code-set will be defined and isolated from the Arabic page i.e. only those characters which are permissible in KASHMIRI will be retained. Since Unicode Code block U+600 is highly liable to spoofing, the choice of the character-set pertinent to KASHMIRI alone will reduce spoofing and phishing.

2. 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.

3. 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 the shadda and the sub-script and super-script aleph.

   ﺳ     ا     ⃝
   ا

   Alif Madd and   Hamza Characters (both normal and wavy) most frequently used in KASHMIRI are as under and these will be  admitted to the permissible set.

   إ     ﻚ     ؤ     ئ     آ     ۀ     أ     أ

   Their corresponding combinations shall be treated as variant. Thus 0622 آ can also be entered as  (0627) اfollowed by  ◌(0653) in some KASHMIRI keyboards

4

and it is to resolve this alternative mode of entry that such normalization is permitted in the shape of a variant.

## 4. PRINCIPLE IV: *EZAFAT*

A serious issue will be that of the ezafat in words such as *bazm-e-adab* or *ab-e-hayat* or *bu-e-gul*. As a palliative suggestion, it is suggested that the ezafat be represented by ye baree or he-gol hamza separated by a hyphen as in the examples below:

<div dir="rtl">

ے- یاد-ے-خدا

و- آب-و-ہوا

ء- سورء-فجر

</div>

## 5. 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 achar*** which when written together would be illegible.

<div dir="rtl">

آمآچار

</div>

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

<div dir="rtl">

آم- آچار

</div>

## 6. PRINCIPLE VI: *Use of Naskh instead of Nastalique in the URL*

Nashq is more visually clear and reduces also spoofing and pharming because of clear legibility of the joining characters.

## 2. RESTRICTION RULES

The following restriction rules specific to Kashmiri shall apply. By restriction rules is meant combinations which though theoretically possible in the language, are disallowed by the norms of the language.

Kashmiri admits the following constraints:

1. Kashmiri does not permit voiced aspirated forms for /b/ /t/ /g/ i.e. bi-labial, dental and velar plosives. This means that ARABIC LETTER HEH DOACHASHMEE ( 06BE - ﻪ) shall not be followed by either of these characters -

   a. ARABIC LETTER BEH ( 0628 - ب)

   b. ARABIC LETTER DAL ( 062F - د)

   c. ARABIC LETTER GAF ( 06AF - گ)

   d. ARABIC LETTER MEEM ( 0645 - م)

   e. ARABIC LETTER NOON ( 0646 - ن)

2. Following characters shall be permitted to come only after ARABIC LETTER ALEF ( 0627 - ا)

   a. ARABIC MADDAH ABOVE ( 0653 - ٓ◌)

   b. ARABIC HAMZA BELOW ( 0655 - ◌ٕ)

   c. ARABIC WAVY HAMZA BELOW (065F - ٟ)

   d. ARABIC HAMZA ABOVE ( 0654 ◌ٔ )

3. Following characters only shall be permitted to come before ARABIC HAMZA

ABOVE ( 0654 – ‌ٔ◌ )

    a. ARABIC LETTER ALEF ( 0627 - ا)

    b. ARABIC LETTER WAW ( 0648 - و)

    c. ARABIC LETTER HEH GOAL ( 06C1 - ہ)

    d. ARABIC LETTER YEH BAREE ( 06D2 - ے)

    e. ARABIC LETTER FARSI YEH ( 06CC - ی)

# 3. LANGUAGE TABLE : KASHMIRI[2]

**0600**       **Arabic**       **06FF**



The Unicode Standard 11.0, Copyright © 1991-2018 Unicode, Inc. All rights reserved.

---

2   Characters marked in yellow are not applicable to the language

## 4. NOMENCLATURAL DESCRIPTION TABLE OF KASHMIRI LANGUAGE TABLE

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

| Unicode Value | Character Name | Character |
|---|---|---|
| 0627 | ARABIC LETTER ALEF | ا |
| 0628 | ARABIC LETTER BEH | ب |
| 062A | ARABIC LETTER TEH | ت |
| 062B | ARABIC LETTER THEH | ث |
| 062C | ARABIC LETTER JEEM | ج |
| 062D | ARABIC LETTER HAH | ح |
| 062E | ARABIC LETTER KHAH | خ |
| 062F | ARABIC LETTER DAL | د |
| 0630 | ARABIC LETTER THAL | ذ |
| 0631 | ARABIC LETTER REH | ر |
| 0632 | ARABIC LETTER ZAIN | ز |
| 0633 | ARABIC LETTER SEEN | س |
| 0634 | ARABIC LETTER SHEEN | ش |
| 0635 | ARABIC LETTER SAD | ص |

| 0636 | ARABIC LETTER DAD | ض |
| 0637 | ARABIC LETTER TAH | ط |
| 0638 | ARABIC LETTER ZAH | ظ |
| 0639 | ARABIC LETTER AIN | ع |
| 063A | ARABIC LETTER GHAIN | غ |
| 0641 | ARABIC LETTER FEH | ف |
| 0642 | ARABIC LETTER QAF | ق |
| 0644 | ARABIC LETTER LAM | ل |
| 0645 | ARABIC LETTER MEEM | م |
| 0646 | ARABIC LETTER NOON | ن |
| 0648 | ARABIC LETTER WAW | و |
| 0679 | ARABIC LETTER TTEH | ٹ |
| 067E | ARABIC LETTER PEH | پ |
| 0688 | ARABIC LETTER DDAL | ڈ |
| 0691 | ARABIC LETTER RREH | ڑ |
| 0698 | ARABIC LETTER JEH | ژ |
| 06A9 | ARABIC LETTER KEHEH | ک |
| 06AF | ARABIC LETTER GAF | گ |

| 06BE | ARABIC LETTER HEH DOACHASHMEE | ﻫ |
| 06C1 | ARABIC LETTER HEH GOAL | ﻩ |
| 06C3 | ARABIC LETTER TEH MARBUTA GOAL | ة |
| 06C4 | ARABIC LETTER WAW WITH RING | ۄ |
| 06CC | ARABIC LETTER FARSI YEH | ى |
| 06D2 | ARABIC LETTER YEH BARREE | ے |
| 06EA | ARABIC EMPTY CENTRE LOW STOP | ۪ |

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

| Unicode Value | Character Name | Character |
|---|---|---|
| 0622 | ARABIC LETTER ALEF WITH MADDA ABOVE | آ |
| 0623 | ARABIC LETTER ALEF WITH HAMZA ABOVE | أ |
| 0624 | ARABIC LETTER WAW WITH HAMZA ABOVE | ؤ |
| 0625 | ARABIC LETTER ALEF WITH HAMZA BELOW | إ |
| 0626 | ARABIC LETTER YEH WITH HAMZA ABOVE | ئ |
| 0651 | ARABIC SHADDA | ّ |
| 0656 | ARABIC SUBSCRIPT ALEF | ٖ |
| 065F | ARABIC WAVY HAMZA BELOW | ٟ |
| 0670 | ARABIC LETTER SUPERSCRIPT ALEF | ٰ |
| 0672 | ARABIC LETTER ALEF WITH WAVY HAMZA | ٲ |

| | ABOVE | |
|------|-------------------------------------------------|----|
| 06C2 | ARABIC LETTER HEH GOAL WITH HAMZA ABOVE | ء |
| 06D3 | ARABIC LETTER YEH BARREE WITH HAMZA ABOVE | ئے |

# 5. VARIANT TABLE FOR KASHMIRI

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[3]

| VARIANTS | |
|---|---|
| ہ <br> 06C1 | ۃ <br> 06C3 |
| آ <br> 0622 | ‎ا+ٓ◌ <br> 0627 + 0653 |
| أ <br> 0623 | ا+ٔ◌ <br> 0627 + 0654 |
| ؤ <br> 0624 | و+ٔ◌ <br> 0648 + 0654 |
| إ <br> 0625 | ا+ٕ◌ <br> 0627 + 0655 |
| ئ <br> 0626 | ى+ٔ◌ <br> 06CC + 0654 |
| ۂ <br> 06C2 | ہ+ٔ◌ <br> 06C1 + 0654 |
| ۓ <br> 06D3 | ے+ٔ◌ <br> 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 KASHMIRI and absence of diacritics.

# 6. EXPERTISE/BODIES CONSULTED

Expertise provided by expert of KASHMIRI language from University of Kashmir.

# 7. COUNTRY CODE TOP LEVEL DOMAIN (ccTLD) FOR KASHMIRI

India (Bhārat) localized in Kashmiri   بارت   (xn--mgbbh1a)
–