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

INTERNATIONALIZED DOMAIN NAMES

IN INDIAN LANGUAGES

A DRAFT POLICY DOCUMENT

POLICY DOCUMENT FOR CREATION OF LANGUAGE INFORMATION FOR INTERNATIONALIZED DOMAIN NAMES IN INDIAN LANGUAGE

This document is a white paper which attempts to lay down the policy for the creation of internationalized domain names in Indian languages. It is the result of discussions, email exchanges as well as document formalizations over the past months in order to arrive at a working draft which is proposed in what follows.

1. BACKGROUND

In this age of Information Technology (IT) when the entire Globe is being integrated into a web-linked village with the knowledge as the sole differentiator, development of convivial (i.e. natural, convenient, and at the same time, affordable) Access Technology has gained prime importance. Especially for India, with its diverse and multi-lingual heritage and culture, Internet is expected to play dominant integrating role for integrating all most all aspects of social and economic endeavor.

Normally, in Internet operation host name of the target Web Server is submitted to the browser who then sends a request to the Domain Name System (DNS) Resolver Service for translating it into the corresponding Internet Protocol (IP) Address for establishing a physical connection to that Web Server.

With the opening up of the multilingual net, for a number of crucial Customer-centric Applications (such as e-governance, e-learning, e-commerce) sole dependence on a single language (i.e. English) may not be sufficient to provide the requisite infrastructural support to all kinds of Internet usage in present and future. Till recently, there was no standardized method for specifying Domain Names in any non- "Simple English Latin" (say, in ASCII) Character Set. The introduction of multilingual IDN's solves this issue, and will offer many new opportunities and benefits for Internet users around the world by allowing them to establish and use domains in their native languages and scripts.

IDNs provide a convenient mechanism for users to access Web sites in local language; for example: if a person wants to give his or her system domain name in his or her local language, say Hindi, then that will look like www.भाषा.भारत

2. OBJECTIVES

The main objectives of this white paper are to clarify and explain the IDN policy by providing information regarding the broad policy and also placing at the forefront the syllable which is the nucleus of all scripts derived from Brahmi.

Since Indian scripts permit a large number of ligatures, pharming and spoofing can be a major threat.
With these two considerations in mind, the main objectives are as under:

a. To ensure that Indian languages can have their rightful place in Internationalized Domain Names and that one can have a URL in an Indian language.
b. To limit, at present the Indian language component to the Domain Name and localize the ccTLD.

3. POLICY IN BRIEF

Following are the general policy guidelines in case of Indian domain names:

1. Only letters, digits, and hyphens will be allowed in a domain name. Names cannot begin or end with hyphens.
2. Mixing of two scripts will not be allowed.
3. Use of Zero Width Joiner/Zero Width Non Joiner will not be allowed.
4. Language numerals and punctuations will not be allowed.
5. Symbols or stress markers will not be allowed.

4. BROAD POLICY

The broad policy enunciates the major guide-lines laid down for creation of IDN's in Indian languages and which will be of use to the registrars as well as all entities and organizations involved in allotment or monitoring or use of IDN. The policy guide-lines are a series of dos and don'ts which stipulate reference rules to be followed in the creation of IDN's. In addition they also handle issues such as Zero Width Joiner, Variant Tables etc. The policy guide-lines have been enunciated with the major aim of ensuring that as far as possible phishing, spoofing and pharming shall be eliminated from IDN's in Indian languages.
The broad policy guide-lines are as under:

4.1. CODE SET: UNICODE COMPLIANCY

The layouts shall be Unicode 5.1 compliant in anticipation of the same being implemented by ICANN. This will permit inclusion of Chillu Characters in Malayalam which is the latest addition to the Unicode with regard to Indian Scripts. Eventual upgradations may be visualized as and when Unicode adds new characters.

4.2. SCRIPT AND LANGUAGE: DIFFERENTIATION OF SCRIPTS AND LANGUAGES

A major decision taken is that Scripts and Languages will be differentiated at the registrar's level and that the user will be provided with keyboards which will allow him to enter an IDN in the language of his choice. Although this does not affect languages like Gujarati or Tamil where there is the relationship of One script <-> One language, it does make a difference in scripts such as Devanagari or Bangla, where one script caters to many languages e.g. Hindi, Marathi, Konkani, Nepali, Sanskrit, all use Devanagari.

4.3. BASIC STRUCTURE OR CANONICAL FORM

The basic structure of the IDN as mentioned above is determined by the notion of the Indic syllable and ensuring that the syllable is well-formed. Within this formalism, certain entities shall be permitted and others disallowed:

A. PERMISSIBLE ENTITIES

Letter-Hyphen-Digit shall be the only entities permitted. Hyphen and Digits shall belong to the Latin set. Letters will be of the language in question.
e.g. ***1 2 3 4 5 6 7 8 9 0*** and **–** will be of the Latin set.
All letters (characters) will be of the pertinent script.

B. NOT PERMISSIBLE

1. CODE-PAGE MIXING

No mixing of scripts at a given level will NOT be allowed

e.g. www.सोफ़्ट-વેર.in or www.हिन्दी-Hindi.in is Not permissible since Hindi and Gujarati are mixed together and Hindi and English are mixed together respectively.

2. DIGITS

Digits in Indian languages will NOT be allowed.

০১২৩৪৫৬৭৮৯

3. PUNCTUATION MARKERS

Punctuation markers present in Indian languages such as danda and double danda ।॥ will NOT be allowed.

4. OTHER SYMBOLS AND ABBREVIATIONS

Since IDN deals only with basic characters, abbreviations and other iconic characters like Isshar(◌ঁ), Abbreviation sign (°) etc. will NOT be allowed.

5.RARE AND OBSOLETE CHARACTERS

Characters which have been added to code-charts to accommodate rare forms especially long vocalic RR and long vocalic LL ৡ ৠ as well as their matra forms ◌ৢ ◌ৣ . In some languages such as Marathi the short vocalic L is permitted◌ৢ and used especially as a Matra. This will be permitted for Marathi.

6. STRESS MARKERS OF CLASSICAL SANSKRIT AND VEDIC

Stress markers e.g. Swarita ◌॑ and Udatta ◌॒ will NOT be allowed.

7.SINGLE DIGIT AND COMBINATION OF TWO DIGIT

Single digit (for ex. 1,2,3,4 etc.) and Combination of two digits (for ex 12, 23, 34 etc) will NOT be allowed as per the .in registry. According to .in registry ".IN domain names may be between 3 and 63 characters in length".All other rules pertaining to .in will be followed.

4.4. SAFEGUARDS

To protect as far as possible against spoofing, phishing and pharming attacks the following safeguards have been introduced. These attacks take place by substituting an address which looks visually alike but which in fact is a fake URL. It should be remembered that the browser window allows for a font size which is relatively small and hence can lead to visual spoofing.

Three such cases of visual identity are possible and safeguards have been instituted against each:

A. DISALLOWING ZERO WIDTH JOINER AND NON-JOINER

The use of Zero width Joiner / Zero width non Joiner (vide RFC 3454 Zero width non joiner (200C)/ zero width joiner (200D) shall NOT be permitted. This is done to avoid spoofing. Use of ZWJ/ZWNJ can result in the following cases, all of which look visually alike.

महाराष्ट्र

महाराष्ट्र with zero width joiner after हा

महाराष्ट्र with zero width non-joiner after म

B. VARIANT TABLE

1. General Notion of the Variant Table

Given the nature of Indian languages, priority will be accorded to graphic identity and phonetic representation will be a secondary feature. Orthographical variants of the type realise/realize will not be considered since there are no specific rules governing such variants.

The main objective of the Variant table is therefore to identify visual look-alikes or homographs and ensure that such homographs shall not be permitted. A common case of visual look-alike is the case of द्ध ध्द .

It is precisely to protect against such visual identity or homographs that a variant table has been instituted. The function of the variant table is to allow one of the homographs, debarring the other one. First use of either one of the characters shall automatically disallow the other in the case of a given word.

Thus if a user chooses समरुद्धी, the choice will automatically debar समरुध्दी, protecting against possible spoofing. The Variant tables will be determined by the Script. However for the sake of convenience and ease of reference, it is desirable that wherever more than one languages use a single script, separate variant tables be provided for each language, all the more so, since the variant table of Hindi involves normalisation whereas that of Marathi, Nepali do not need any such device.

2. Typology of the variant Table for Indian Languages

Broadly two kinds of variants can be identified:

1. Homographic Variants where the two or more variants are conjuncts which look alike in the URL, but in fact have different phonetic representation

2. Homographic Variants where owing to Unicode permitting more than one manner of inputting a given character or a string of characters, spoofing

is possible. These have different methods of inputting but phonetically have the same representation.

Details of these types are given below:

1. Homographic Variants pronounced differently but which look nearly alike

These variants occur owing to the complex nature of Indian scripts. The nature of these variants is of a set of characters when seen together become variant with another set of characters when seen together. These variants do not have same meaning but they are variants because of the similar visual representation of those particular sets of characters.

2. Homographic Variants which look alike and are pronounced the same.

These variants occur due to the fact that Unicode has given two different ways to input a single logical character. Mostly in these cases, one of the variants is a single character while the other variant is a set of characters. This category is further divided into two sub-categories:

a. Variants subject to Normalization as per Unicode: These are the characters that have been identified by the Unicode under the Unicode normalization rules.

b. Variants due to legacy inputting: These variants occur due to the legacy way of inputting a certain character using a set of code points and later been assigned a separate code point by the Unicode.

3. Rules governing the variant Table

The following rules determine variant tables:

1. Since exclusion tables based on variants can debar a large number of words commonly used, the variant table shall be used sparingly and only when absolutely necessary.
2. Further the variant table shall apply only to ligatures or conjuncts or combination of two or more consonants and single characters that have homographic identity shall not be part of the variant table, the logic being that a native speaker can easily disambiguate single characters. It is the conjunct forms that are the potential source of spoofing, phishing and pharming.
3. Normalization variants shall be part of the variant table and shall be provided as a safety measure. Therefore characters which need normalization should automatically normalize if they are a part of a variant table.
   e.g. in case of Hindi:

क(0915) + ़(093C) = क़(0958)

This is added as a safeguard, since our analysis has shown that different browsers do not handle normalization as laid down in IDNA2008

4.5. LAYOUT

Each language policy will have the following layout :

1 The generic syllable structure shall be suitably modified to suit the respective language.

2 A Code-chart for each language based on Unicode 5.1 shall be provided. Characters which are not in consonance with the LHD policy and which are to be excluded shall be marked in yellow on the code-chart.

3 A map of the above code-chart specifying accurately the information regarding the characters present in the above code-chart shall be provided.

4 Finally to reduce the risk of spoofing a variant Table will be provided where the possible variants shall be listed. **As far as possible these variants shall not be individual characters but ligatures that are close homographs.** Normalization shall be part of the variant table and shall be provided as a safety measure, although a large number of browsers automatically normalize.

4.6. PUBLIC REVIEW

The final document so prepared for all major languages is put up on a site for comments and also circulated to obtain maximum feedback.

5. APPLICATION OF THE POLICY TO HINDI

In the following an example will be taken from Hindi and the policy as applicable to Hindi language will be succinctly provided. Each part of the policy will be taken in turn and explained with examples from Hindi wherever possible:

1. MECHANISM TO ENSURE THE WELL-FORMEDNESS OF THE SYLLABLE

Scripts used for Indian languages have evolved from the ancient Brahmi script and have a common phonetic structure, making a common character set possible with some language exceptions. The Brahmi syllable ensures the well-formedness of the written syllable of all languages derived from the script. This brahmi syllable is defined by the Backus Naur Formalism(BNF) which is based on ISCII. While delegating the TLDs or Second level domain name, this well-formedness should be ensured.

Example

किताब -> Well Formed

ि◌कताब -> Ill Formed

2. LANGUAGE TABLE FOR HINDI

Unicode Code charts are based on Scripts. Not all the characters present in the Devanagari code chart are applicable for Hindi. Thus a subset of the characters out of the code chart for Devanagari (0900 – 097F) has been identified which caters to the Hindi language. Yellowed out characters are non-applicable characters for Hindi Language.

0900 **Devanagari** **097F**

3. NOMENCLATURAL DESCRIPTION TABLE OF HINDI LANGUAGE TABLE

This table gives a detailed description of various characters permitted in the "Language Table for Hindi". Characters have been grouped as per their behavioural characteristics.

Chandrabindu(B)		
0901	ँ	DEVANAGARI SIGN CANDRABINDU = anunasika
Anusvara (D)		
0902	ं	DEVANAGARI SIGN ANUSVARA = bindu
Visarga (X)		
0903	ः	DEVANAGARI SIGN VISARGA
Independent vowels (V)		
0905	अ	DEVANAGARI LETTER A
0906	आ	DEVANAGARI LETTER AA
0907	इ	DEVANAGARI LETTER I
0908	ई	DEVANAGARI LETTER II
0909	उ	DEVANAGARI LETTER U
090A	ऊ	DEVANAGARI LETTER UU
090B	ऋ	DEVANAGARI LETTER VOCALIC R
090D	ऍ	DEVANAGARI LETTER CANDRA E
090F	ए	DEVANAGARI LETTER E
0910	ऐ	DEVANAGARI LETTER AI
0911	ऑ	DEVANAGARI LETTER CANDRA O
0913	ओ	DEVANAGARI LETTER O
0914	औ	DEVANAGARI LETTER AU
Consonants (C)		
0915	क	DEVANAGARI LETTER KA
0916	ख	DEVANAGARI LETTER KHA

0917	ग	DEVANAGARI LETTER GA
0918	घ	DEVANAGARI LETTER GHA
0919	ङ	DEVANAGARI LETTER NGA
091A	च	DEVANAGARI LETTER CA
091B	छ	DEVANAGARI LETTER CHA
091C	ज	DEVANAGARI LETTER JA
091D	झ	DEVANAGARI LETTER JHA
091E	ञ	DEVANAGARI LETTER NYA
091F	ट	DEVANAGARI LETTER TTA
0920	ठ	DEVANAGARI LETTER TTHA
0921	ड	DEVANAGARI LETTER DDA
0922	ढ	DEVANAGARI LETTER DDHA
0923	ण	DEVANAGARI LETTER NNA
0924	त	DEVANAGARI LETTER TA
0925	थ	DEVANAGARI LETTER THA
0926	द	DEVANAGARI LETTER DA
0927	ध	DEVANAGARI LETTER DHA
0928	न	DEVANAGARI LETTER NA
092A	प	DEVANAGARI LETTER PA
092B	फ	DEVANAGARI LETTER PHA
092C	ब	DEVANAGARI LETTER BA
092D	भ	DEVANAGARI LETTER BHA
092E	म	DEVANAGARI LETTER MA
092F	य	DEVANAGARI LETTER YA
0930	र	DEVANAGARI LETTER RA
0932	ल	DEVANAGARI LETTER LA
0935	व	DEVANAGARI LETTER VA
0936	श	DEVANAGARI LETTER SHA
0937	ष	DEVANAGARI LETTER SSA
0938	स	DEVANAGARI LETTER SA

0939	ह	DEVANAGARI LETTER HA
0958	क़	DEVANAGARI LETTER QA
0959	ख़	DEVANAGARI LETTER KHHA
095A	ग़	DEVANAGARI LETTER GHHA
095B	ज़	DEVANAGARI LETTER ZA
095C	ड़	DEVANAGARI LETTER DDDHA
095D	ढ़	DEVANAGARI LETTER RHA
095E	फ़	DEVANAGARI LETTER FA
Dependent vowel signs (Matras) (M)		
093E	ा	DEVANAGARI VOWEL SIGN AA
093F	ि	DEVANAGARI VOWEL SIGN I • stands to the left of the consonant
0940	ी	DEVANAGARI VOWEL SIGN II
0941	ु	DEVANAGARI VOWEL SIGN U
0942	ू	DEVANAGARI VOWEL SIGN UU
0943	ृ	DEVANAGARI VOWEL SIGN VOCALIC R
0945	ॅ	DEVANAGARI VOWEL SIGN CANDRA E = candra
0947	े	DEVANAGARI VOWEL SIGN E
0948	ै	DEVANAGARI VOWEL SIGN AI
0949	ॉ	DEVANAGARI VOWEL SIGN CANDRA O
094B	ो	DEVANAGARI VOWEL SIGN O
094C	ौ	DEVANAGARI VOWEL SIGN AU
Halant (H)		
094D	्	DEVANAGARI SIGN VIRAMA = halant (the preferred Hindi name) • suppresses inherent vowel
Nukta (N)		
093C	़	DEVANAGARI SIGN NUKTA
Avagraha (Y)		
093D	ऽ	DEVANAGARI SIGN AVAGRAHA

4. VARIANTS

Homographic Variants which look alike and are pronounced the same.		
क+़ 0915+093C	क़ 0958	
ख+़ 0916+093C	ख़ 0959	
ग+़ 0917+093C	ग़ 095A	
ज+़ 091C+093C	ज़ 095B	
ड+़ O921+093C	ड़ 095C	
ढ+़ 0922+093C	ढ़ 095D	
फ+़ 092B+093C	फ़ 095E	
Homographic Variants which look alike and are pronounced the same – Not under Normalization		
NOT APPLICABLE IN HINDI		
Homographic Variants pronounced differently but which look nearly alike		
द्र 0926+094D+0917	द्र 0926+094D+0930	द्र 0926+094D+0928
द्ध 0926+094D+0927	द्ध 0926+094D+0918	
ष्ट 0937+094D+091F	ष्ठ 0937+094D+0920	

श्व 0936+094D+0935	श्व 0936+094D+0930+094D+0935
श्न 0936+094D+0928	श्न 0936+094D+0930+094D+0928
श्च 0936+094D+091A	श्च 0936+094D+0930+094D+091A
श्ल 0936+094D+0932	श्ल 0936+094D+0930+094D+0932
त्त 0924+094D+0924	त 0924
ॅ 0901	ॅं 0945+0902
द्व 0926+094D+0935	द्व 0926+094D+092C