DEVANĀGARĪ VIP TEAM
ISSUES REPORT

DEVANĀGARĪ VIP GROUP

Contents

0. PRELIMINARIES

1. Background and Overview

This report targets issues pertinent to TLD's with specific reference to Devanāgarī script. However to situate these issues within a wider perspective, this general back-ground and overview are provided.
Thanks to the policy of opening up scripts other than Latin by ICANN, a flood-gate of new languages and scripts has opened up and domain-names will become truly multi-lingual in nature. Benefiting from this new policy, India has taken up the challenge of providing IDN's in Indian scripts and languages for the 22 official languages of India (A list of these languages is provided for general information in Appendix II. Official Languages using Devanāgarī are highlighted.).

The formulation of a policy document for India to provide Internationalized Domain Names in the 22 official languages has been nearly 5 years in the making. Started in 2005, the policy has been elaborated over the years to ensure that the eventual users will have as safe as an environment as possible when they register their names in an Indian language using their native script.

7 Indian languages (Hindi, Tamil, Telugu, Gujarati, Bangla, Urdu and Punjabi) have already been proposed to ICANN and IANA and the ccTLD for the country name "India" in these languages have already been approved and delegated into the DNS root zone.

Since scripts do not share the same composition rules and have their own "grammar of composition"; it was in the fitness of things, that ICANN felt that the creation of "test cases" in six scripts would allow for a better perception of the problems as well as issues involved. The scripts chosen for study (apart from Latin): Greek, Cyrillic, Arabic, Devanāgarī, Chinese reflect in fact the 4 major writing systems of the world Abugidas (Greek and Cyrillic), Abjads (Arabic), Akshar or Alphasyllabaries (Devanāgarī) and Phonetic-Semantic (Chinese).

Within this perspective a series of discussions via e-mail were initiated. A team was constituted for Devanāgarī (cf. Appendix I) which embraced not only Hindi but other major languages using the Devanāgarī script (cf. Appendix II). The discussions culminated in a meeting of all the groups at Singapore in June and another meeting of the Devanāgarī group at Pune in July.

Over a series of discussions both prior to the creation of the case-study team and after, a slow consensus building process has been evolving and a major step towards this process is a preliminary draft in which each script delineates its problems, issues especially with reference to its writing structure and the notion of variants arising there from.

It is these concerns and issues which this report addresses. The report attempts to lay down the background to writing system along with the various issues for the creation of Internationalized Domain Names in Languages using Devanāgarī. It is the result of

discussions, teleconferences, 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.

2. Structure

The report, whose basic layout was finalized at a meeting the case study team held in Pune, comprises the following sections:

Part 1 lays down the basic postulates, which in our opinion, are the corner-stone of the issues report for Devanāgarī.

Part 2 attempts to set things in perspective by providing an overview of the evolution of Devanāgarī, the languages that use Devanāgarī and also a brief sketch of the writing system of the language.

Since the aim of this document is to highlight issues pertinent to all aspects of IDN variants: linguistic, technical, societal, fiscal, and administrative, these issues are highlighted in a sequential order[1]. Part 3 is an inventory of the major issues pertinent to the topic in question and examines the problems from all angles.

There exist a certain number of areas which have no direct bearing on the variant issues, but because of their intrinsic nature, these are indirectly linked to the problematic under survey. These are listed in Part 4.

Since the Registry plays an important role in IDN, a special section, Part 5 is devoted to this area.

A certain number of Appendices which provide ancillary information complete the report.

[1] Since some of these are interesting but do not have direct relevance to the issue of Variants, they have been listed in Appendix V

1. POSTULATES

For IDN purposes, ICANN has tended to make certain assumptions about acceptance of relevant technologies these assumptions constitute the basic postulates that underlie this report on Devanāgarī variant issues. These postulates are as under:

1.1. Unicode is acceptable, if only because no other relevant global coded character set is available. Accepting Unicode includes accepting its normalization model and their stability policy with reference to normalization.

1.2. IDNA2008, including its interpretation of Unicode properties and the version evolution model, are acceptable.

1.3. The DNS, including its restrictions on exact lookup (known item search), the absence of language-specific information and language-specific or script-specific lookup or matching mechanisms, and aliases that do not carry context or that can point from anywhere in the DNS tree, is acceptable.

1.4. TLD names are limited to "letters" alone. Digits and Hyphens as well as ZWJ U+200C, ZWNJ U+200D will not be permitted within a TLD label.

1.5. The contents of the DNS are about mnemonics, not about "words" or longer statements in particular languages. The fact that something can be written in a particular language, or even looked up in its dictionary, does not imply an entitlement to have that string appear in the DNS.

1.6. Any domain name tree may have subordinate zones with separate, administratively-distinct, registration and maintenance and administrative arrangements.

1.7. This issues report is limited to IDN variant TLD's alone (with specific reference to Devanāgarī) and may not apply to registration under subordinate zones, although the issues discussed in the report could provide gainful insights into the functioning of those subordinate zones.

2. DEVANĀGARĪ: AN OVERVIEW

This over-view of Devanāgarī is a linguistic introduction to Devanāgarī. It starts off with the historical evolution of Devanāgarī and in section 2.1. studies the structure of Devanāgarī. Section 2.2. develops the notion of the underlying nucleus: the akshar and further draws attention to certain akshar structures relevant to variants. IPA as well as simple transliteration has been used as a guide to the pronunciation of the examples.

2.1. Devanāgarī: A Historical Perspective

Devanāgarī is the main script for the Indo-Aryan languages Hindi, Marathi, Maithili, Dogri, Boro, Santhali, Sanskrit and Nepali recognized as official languages of the Republic of India. The script is also shared with other countries such as Fiji (Hindi) and Nepal (Nepali). It is the only script also for the related Indo-Aryan languages Bagheli, Bhili, Bhojpuri, Himachali dialects, Magahi, Newari and Rajasthani. It is associated closely with the ancient languages Sanskrit and Prakrit. It is an alternative script for Kashmiri (by Hindu speakers), Sindhi and Santhali. It is rising in use for speakers of tribal languages of Arunachal Pradesh, Bihar and Andaman & Nicobar Islands.

It is well-known that Devanāgarī has evolved from the parent script Brāhmī, with its earliest historical form known as Aśokan Brāhmī, traced to the 4th century B.C. Brāhmī was deciphered by Sir James Prinsep in 1837. The study of Brāhmī and its development has shown that it has given rise to most of the scripts in India, as mentioned above, and some outside India, namely, Sri Lanka, Myanmar, Kampuchea, Thailand, Laos, and Tibet.

The evolution of Brāhmī into present-day Devanāgarī involved intermediate forms, common to other scripts such as Gupta and Śāradā in the north and Grantha and Kadamba in the South. Devanāgarī can be said to have developed from the Kutila script, a descendant of the Gupta script, in turn a descendent of Brāhmī. The word *kutila,* meaning 'crooked', was used as a descriptive term to characterize the curving shapes of the script, compared to the straight lines of Brāhmī. A look at the development of Devanāgarī from Brāhmī gives an insight into how the Indic scripts have come to be diversified: the handiwork of engravers and writers who used different types of strokes leading to different regional styles (cf. Singh 2006).

In what follows all mention to Brāhmī is for historical reasons and in no manner should this report be adduced as pertaining to Brāhmī, its main focus being Devanāgarī alone.

2.2. The structure of written Devanāgarī

Devanāgarī is an alphasyllabary and the heart of the writing system is the syllable or akshar. It is this unit which is instinctively recognized by users of the script. To understand the notion of akshar, a brief overview of the writing system is provided in Section 2.2. and the akshar itself will be treated in depth in Section 2.3.

2.2. The writing system of Devanāgarī could be summed up as composed of the following:

2.2.1. The Consonants

Devanāgarī consonants have an implicit schwa /ə/ included in them. As per traditional classification they are categorized according to their phonetic properties. There are 5 (Varg) groups and one non-Varg group. Each Varg contains five consonants classified as per their properties. The first four consonants are classified on the basis of Voicing and Aspiration and the last is the corresponding nasal.

Varg	Unvoiced		Voiced		Nasal
	-Asp	+Asp	-Asp	+Asp	
1 Velar	क	ख	ग	घ	ङ
2 Palatal	च	छ	ज	झ	ञ
3 Retroflex	ट	ठ	ड	ढ	ण
4 Dental	त	थ	द	ध	न
5 Bi-labial	प	फ	ब	भ	म

Non-Varg

य	र	ल	ळ	व	श	ष	स	ह

2.2.1. The Implicit Vowel Killer: Halanta[2]

All consonants have an implicit vowel sign (schwa) within them. A special sign is needed to denote that this implicit vowel is stripped off. This is known as the Halanta (्). The Halanta thus joins two consonants and creates conjuncts which can be from 2 to 3 consonant combinations (cf. 1.2. supra)

2.2.2. Vowels

Separate symbols exist for all Vowels which are pronounced independently either at the beginning or after a vowel sound. To indicate a Vowel sound other than the implicit one, a Vowel modifier (Mātrā) is attached to the consonant. Since the consonant has a built in schwa, there are equivalent Mātrās for all vowels excepting the अ.

The correlation is shown as under:

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

In addition to show sounds borrowed from English, some languages using Devanāgarī such as Hindi, Marathi, and Konkani also admit 2 vowels and their corresponding Mātrās as in

ऍ ॅ ऑ ॉ

ऍण्ड /and/ ऑर /or/

Marathi replaces the ऍ by अॅ

2.2.3. The Anuswāra /ं/

[2] Unicode (cf. Unicode 3.0 and above) prefers the term Virama. In this report both the terms have been used to denote the character that suppresses the inherent vowel.

The Anuswāra represents a homo-organic nasal. It replaces a conjunct group of a Nasal consonant+Halanta+Consonant belonging to that particular varg. Before a Non-varg consonant the anuswāra represents a nasal sound. Modern Hindi, Marathi and Konkani prefer the anuswāra to the corresponding Half-nasal:

सन्त vs. संत /sənt/ saint चम्पा vs. चंपा /tʃəmpa/

2.2.4 Nasalization: Chandrabindu ँ

Chandrabindu/Anunasika denotes nasalization of the preceding vowel as in आँख (eye) /ãkh/ eye. Present-day Hindi users tend to replace the chandrabindu by the anuswāra

2.2.5. Nukta़

Mainly used in Hindi, the nukta sign is placed below a certain number of consonants to represent words borrowed from Perso-Arabic. It can be adjoined to क ख ग ज फ to show that words having these consonants with a nukta are to be pronounced in the Perso-Arabic style.

e.g. फ़िरोज़ /firoz/

It is also placed under ड ढ in Hindi to indicate flapped sounds

With the exception of flaps, users of modern-day Hindi hardly use the nukta characters today

2.2.6. Visarg ः and Avagrah ऽ

The Visarg is frequently used in Sanskrit and represents a sound very close to /h/. दुःख /du:kh/ sorrow, unhappiness

The Avagrah ऽ creates an extra stress on the preceding vowel and is used in Sanskrit texts. It is rarely used in other languages using Devanāgarī.

2.3. The Fundamental Unit: akshar

This classification of Devanāgarī characters can be reduced to a "compositional grammar" based on a Backus-Naur formalism (ISCII '91) which ensures the well-formedness of the akshar. The formalism describes the nodal units of the script: Consonant and Vowel and determines which elements can be conjoined to each of these Nodal Units. The fundamental properties of the akshar are defined below:

The *akshar* is the graphemic unit of Devanāgarī. The difference between the syllable and the akshar is that while the syllable includes one or more post-vocalic consonants, the akshar doesn't, as can be seen below:

Phonemic forms	Syllabic units	Akshara units
cha:rulәta:	CV. CV. CV. CV	CV. CV. CV. CV
e:k	VC.	V. C
upka:r	VC. CVC	V. C. CV. C
indira	VC. CV. CV	VC. CV. CV
әst	VCC	V. CC
әkʃәr	VC. CVC	V. CCV. C

Table 1: Syllabic and akshara divisions of spoken forms

As can be seen from Table 1, there is a marked difference between the written and spoken syllable, especially insofar as the division of consonant clusters across syllable boundaries e.g. /upka:r/ is concerned.

The only exception to the generalization about the post-vocalic consonants vis-à-vis akshar is the anuswāra, the underlying nasal consonant surfacing as homorganic with the following stop. The anuswāra is treated as a part of the grapheme. The orthographic and phonetic transcriptions of forms with the anuswāra are given below:

बिंदी	[bindi:]	'point$_N$'
कंबल	[kәmbәl]	'blanket$_N$'
डंडा	[ɖәɳɖa:]	'stick$_N$'
खंजर	[k^{h}әɲʄәr]	'knife$_N$'
कंघी	[kәŋg^{ɦ}i:]	'comb$_N$'

Table 2: Representation of anuswāra in Devanāgarī

The vowel is an independent unit of *akshar* word-initially and post-vocalically.

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
ə	a	i	i:	u	u:	e:	æ:	o:	ɔ

Table 3: Independent vowel letters

Vowels and consonants are assumed to be different types of units and are so represented in the grapheme when the vowels follow consonants. The following akshar consist of single consonants followed by a vowel:

क	का	कि	की	कु	कू	के	कै	को	कौ
kə	ka	kɪ	ki:	kʊ	ku:	ke	kæ	ko	kɔ

Table 4: *Devanāgarī CV akshar*

As can be seen in the first grapheme in Table 3, the neutral vowel /ə/ is assumed to be inherent in a consonant. The vowel is pronounced as such word initially and medially in certain contexts, for example, in the first grapheme in पल /pəl/. The inherent neutral vowel is not pronounced word-finally or medially in certain contexts.

Two-consonant clusters

5. Generally, half the letter of the first consonant precedes the full letter of the second consonant: e.g., स्क <sk> , प्त <pt>, क्ल <kl> etc. Alternatively, the practice of specifying the diacritic for unreleased consonants, known as 'halanta', is used for the first consonant, e.g., द् भ<dbʰ> उद् भव/udbʰəʊ/

6. For a C+r cluster, as noted above, the /r/ is specified as a subscript that looks like an inscript: क्र <kr>, ख्र <kʰr>, फ्र <pʰr>.

7. For r+C clusters, the the /r/ is specified as a superscript above the grapheme, e.g., र्म <rm>, र्ज <rʤ>

8. In the case of the following two-consonant clusters, a new ligatured group is formed.

These are: त्र <tr>, क्ष <kṣ>, ज्ञ <ɟɲ>, श्र <ʃr>, क्त <kt>.

Three-consonant clusters:

9. Generally, the first two consonants are specified for half their letters, and the third is fully specified, e.g., स्प्ल <spl>. This convention is usually followed for borrowed words.

10. For C+C+r clusters, and for r+C+C clusters, which are highly restricted, the convention for two-consonant clusters applies, e.g., स्त्र <str>

3. ISSUES

From a typological point of view, the following practical considerations need to be taken into account when discussing issues and trying to identify solutions:

a. Root delegations: The root zone contains two broad categories of zones: ccTLDs and others (sometimes called "gTLDs"). These different types of zones may have different policies governing delegations from them; but the policy governing delegation of any TLD from the root zone should always be the same. Labels in the root will be limited to letters only: digits, hyphens, and other non-letters (including ZWJ and ZWNJ) are not acceptable

b. Introduction of the notion of language tables, restriction rules based on well-formedness constraints and variant-hood to reduce spoofing, pharming and phishing. Thus for Devanāgarī based languages which are akshar driven, a formalism needs to be evolved to handle well-formedness.

Potential areas where such factors apply. These are:

1. Language vs. Script.
2. Variants
3. Issues Related to Software Behavior in Relation to Display of Domains:
4. Whole-script confusables
5. The case for 02BC in Devanāgarī

These will be developed in what follows. By way of conclusion a tabular summing-up of issues has been provided.

3.1. Language vs. Script Issues

Within the ccTLD for .भारत the dichotomy of language vs. script issues can be handled (with certain issues to be tackled at the registry level). Because the root zone is necessarily shared by everyone on the Internet, it is impossible to make reasonable guesses about a user's language based on the user looking something up in the root zone. Therefore, it is the script of the characters in the label that will be relevant for the purposes of deciding on its variants.

It is hoped that with the introduction of scripts such as Devanāgarī , especially those used for representing a large number of languages, a suitable mechanism for handling languages will be evolved.

3.2. Variants in Devanāgarī Script

Two or more characters or character combinations shall be deemed as variants only if

a. They are protocol valid AND

b. Are not listed in the Unicode normalization rules (IDNA 2008) which in turn means, those characters or character combinations which are not stable under Unicode normalization forms NFC and/or NFKC.

Variants in Devanāgarī are of two types: Single characters which look visually alike within the label and ligatural shapes which are visually confusing and can be mistaken for one another.

3.2.1. Confusingly similar single characters

These are single characters which have confusingly similar shapes and tend to be confused one with the other.. This category of variants were not considered in the .भारत ccTLD policy as there was a possibility that this approach would prove to be too much restrictive.

e.g.

घ U+0918	ध U+0927
भ U+092D	म U+092E

Table 5

This table contains only a sample list. A more elaborate list is provided in Appendix IV.

3.2.2. Confusingly similar Composite characters

Since Devanāgarī lends itself to conjuncts there is a possibility of conjuncts that look alike and can be easily confused in the small URL bar of the browser.

e.g.

द्र U+0926 U+094D U+O917	द्र U+0926 U+094D U+0930	द्र U+0926 U+094D U+0928
द्ध U+0926 U+094D U+0927	द्ध U+0926 U+094D U+0918	
ष्ट U+0937 U+094D U+091F	ष्ठ U+0937 U+094D U+0920	
द्व U+0926 U+094D U+ 0935	द्व U+0926 U+094D U+092C	

Table 6

This table contains only a sample list. A more elaborate list is provided in Appendix IV.

3.2.3. Variants generated because of Combining Characters

This sub-type is not relevant to the Devanāgarī script since in the present state of things no such case occurs. However, if there are future additions to code block U+0900 , and if such additions are not handled by the normalization rules for Devanāgarī , these will need to be introduced.

3.3. Issues Related to Software Behavior in Relation to Display of Domains :

The DNS is not exclusively about the web but also affects other areas such as email user agents, calendaring programs etc. However as a case study, issues pertaining to browsers and to email-clients (with specific reference to the Devanāgarī script) will be taken up. The issues highlighted here are applicable to other software behavior in relation to display of domains. It needs to be pointed out that in the case of browsers, redirection might be a feature of the protocol being used (i.e. http redirect). The same is not the case with other software (such as email-clients, calendaring, and some mobile applications) where there is underlying protocol redirect mechanism.

3.3.1. Browser Issues

The browsers for representing the domain name in the URL bar of the browser, rely on the underlying OS rendering engine. Thus the issues associated with the rendering engines of the OS are inherent in the browser. The fonts that get applied on the URL bar are chosen by the browsers as per default font for the script of the domain name provided by the underlying OS.

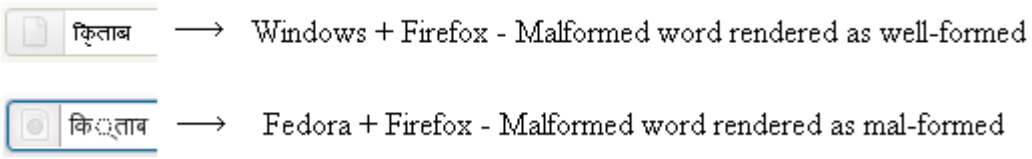
The issues related to these characteristics of the browsers belong to two broad categories as

3.3.1.1. Rendering Engine related issues

3.3.1.2. Font related issues

3.3.1.1. Rendering Engine related issues

Whenever some text is submitted to a Unicode Enabled application, the rendering engine breaks this text in the form of syllables. These syllable formation rules have not been standardized, nor has Unicode given any specific rules pertaining to the same. Thus the behavior of different rendering engines is different and depends on the understanding of the language/script of the implementing body which seldom is perfect. This is exemplified in the theoretical cases given below which show how under different environments the same browser does not display/displays a mal-formed syllable:

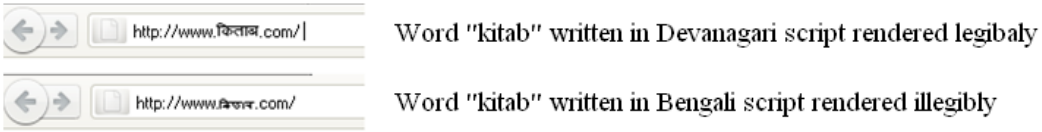


The theoretical example given above with a valid label: किताब (book) shows how the rendering engine of the operating system permits even mal-formed syllables to be rendered as well-formed. The test domain entered is किताब /kitāb/with a halanta/virama after the first syllable: कि. Firefox under Windows shows that this mal-formed syllable is not rendered as malformed. The same under Fedora shows up the malformation of the syllable. It needs to be made clear that the issue raised here addresses the problem of rendering of conjunct shapes which are either rendered in such a manner (because of the native font of the OS) as to make the variant visually not relevant or are simply rendered incorrectly. It is hoped that as software developers become more aware of such issues, these will be corrected. However the issue raised needs to be considered, but it might have a sunset clause as software brings in the right corrections[3].

3.3.1.2. Font related issues

In case of rendering of Domain Names in browsers, font that gets applied on the domain name in address bar of the browser plays major role. Each operating system has a specific font which acts as a default font for every script/language the OS supports. The browser uses default font provided by the OS for displaying the domain name in the address bar.

Similar to the rendering engine, the font implementation also varies from vendor to vendor. And thus the same Domain Name can be seen differently depending on the font properties, orthography adopted by the font, hinting, weight, kerning etc as can be seen in the example below where Hindi and Bengali in the same point size have different visual display: Hindi being more readable than Bengali.



As there is no central authority that can ensure consensus implementations, it is hoped that a user-facing applications software that claims to support Devanāgarī should have a listed set of capabilities that would go a long way toward improving and rationalizing the user experience.

3.3.2. Email Addresses resolution

The queries raised here are pertinent to .भारत but could also apply in certain instances to other registries.

[3] A list of such observations both under Windows and Linux is provided as an attachment since the data is too voluminous to figure in an appendix.

The problems raised pertinent to variants (cf. 3.2. Supra) have a marked resolution for resolution of email addresses. The example taken will be for .भारत but could equally apply to any other TLD label:

Given an email such as
वित्त-मंत्रालय.भारत: Ministry of Finance
Will the owner of the address also inherit the variant: वित-मंत्रालय.भारत
given that त्त U+0924 U+094D U+0924 generates out त U+0924 as a variant

वित्त	U+0935 U+093F **U+0924 U+094D U+0924**
वित	U+0935 U+093F **U+0924**

Some questions that arise are:
- In case both emails are valid, will there be an aliasing mechanism ?

The issue is also closely tied with that of the mail-server resolving the email.

3.4. The case for 02BC in Devanāgarī Script:

The character U+02BC *Modifier Letter Apostrophe* / ’/ which acts as a tone mark or length mark is used very frequently in languages like Boro, Dogri, Maithili which are Devanāgarī script based and Bangla which is Bengali script based. An example from Dogri where 02BC is used as a syncopation marker will clarify the issue:

करा’रदा । (means : got done)

U+02BC *Modifier Letter Apostrophe* character comes from the block U+02B0-U+02FF,whereas all the characters which belong to Devanāgarī script come from the block U+0900-U+097F. If as a policy decision, script mixing is not allowed in gTLDs, this character still be allowed as an exception because without this character the language representation will not be complete[4]. It may be noted that the keyboards devised for languages (Boro, Dogri, Maithili) using this character provide the means of entering the character which has a relatively high frequency of usage in these languages. In the hope that 02BC will be made acceptable in Devanāgarī Code set, Appendix VI provides a list of recommended Devanāgarī code points.

3.5. Whole-Script Confusables

This refers to a scenario where there exist two strings, S1 and S2. The script property of every character in S1 is P1. The script property of every character in S2 is P2. So, each string is in a single script. But the two strings are confusable to any competent speaker of some language[5]. The Unicode example is that of a Latin string containing characters only from that Unicode set and which can have a whole-script confusable in Cyrillic (lowercase-only).

[4] The case needs to be debated since expert opinions are divided on the same.
[5] Cf. the Unicode report on this, http://www.unicode.org/reports/tr39/, particularly section 4.1.

In the case of Devanāgarī, a case may be visualized where a complete string in Devanāgarī can correspond to a compete string in another script, such that it could lead to spoofing. Thus within the URL of a browser the following two may look alike.

मोरी (Devanāgarī script) મોરી (Gujarati script) [morī]

The two scripts have a large number of characters in common, the only difference being the absence of the "shirorekha" or the head line above the characters in Devanāgarī which is missing in Gujarati. However in the small point size of the browser, the two look alike. This can lead to spoofing. Although these labels have not been submitted to the Unicode confusability test (cf. http://www.unicode.org/reports/tr39/, section 4.2), a manual check of the 4 code points in each case shows that these are not listed in the Unicode confusables list: http://unicode.org/Public/security/revision-02/confusables.txt

The Unicode confusables list although quite exhaustive has only 43 confusables for "Gujarati letter" of which none match the Devanāgarī set. The list would need thorough checking to validate whether all confusables are listed.

Given the close resemblance of characters in Indic scripts, and given that it is extremely difficult to identify all such characters across scripts, **this practice if permitted can lead to spoofing and should not be permitted.**

Moreover in genuine cases where the two labels are distinct in the two scripts, it could lead to dispute resolution issues. Thus if morī is a valid label in both scripts and it is disallowed in one (say Gujarati) because of Whole-script confusable, a dispute resolution case could arise.

4. EXTRANEOUS CONSIDERATIONS

These are considerations which have no direct bearing on the variant issues, but because of their intrinsic nature, these are indirectly linked to the problematic under survey and are hence termed as such.

4.1. Cross-script character mixing

In TLDs, there is no possibility of allowing mixing of scripts within a label.
If script-mixing had been allowed, our opnion is, this could have resulted in large amount of spoofing, phishing and scamming because within scripts, there are many cases of characters being confusable with another. This has also been considered at Unicode level[6]. The policy for .भारत ccTLD does not allow code block mixing either.

However a list of cross-lingual visual similarities is provided below. It should be noted that such similarities are restricted to single characters and not to conjuncts. Spoofing can be possible by mixing characters from these different code blocks. The list is in no way exhaustive but suffices to point out the inherent danger.

DEVANĀGARĪ SCRIPT	**COGNATE SCRIPT**	**CODE POINT IN COGNATE SCRIPT**
VOWELS		
उ U+0909	Bangla	ঙ U+0993
उ U+0909	Gurmukhi	ਤ U+0A24
ऋ U+090B	Gujarati	ઋ U+0AE0
CONSONANTS		
क U+0915	Bangla	ক U+0995

[6] http://www.unicode.org/reports/tr39/ , particularly section 5.

ग U+0917	Gujarati	ગ U+0A97
ग U+0917	Gurmukhi	ਗ U+0A17
घ U+0918	Gurmukhi	ਘ U+0A2C
घ U+0918	Gujarati	ઘ U+0A98
ङ U+0919	Gujarati	ઙ U+0A99
छ U+091B	Gujarati	છ U+0A9B
ञ U+091E	Gujarati	ઞ U+0A9E
ट U+091F	Gurmukhi	ਟ U+0A17
ठ U+0920	Gujarati	ઠ U+0AA0
ठ U+0920	Gurmukhi	ਠ U+0A20

ड U+0921	Gujarati	ડ U+0AA1
ढ U+0922	Gurmukhi	ਢ U+0A2B
त U+0924	Gujarati	ત U+0AA4
ध U+0927	Gujarati	ધ U+0AA7
न U+0928	Gujarati	ન U+0AA8
न U+0928	Bangla	ন U+09A8
न U+0928	Bangla	ণ U+09A3
प U+092A	Gujarati	પ U+0AAA
प U+092A	Gurmukhi	ਧ U+0A17
प U+092A	Gurmukhi	ਪ U+0A2A

प U+092A	Gurmukhi	ਪ U+0A6B
म U+092E	Gurmukhi	ਸ U+0A38
म U+092E	Gujarati	મ U+0AAE
य U+092F	Gujarati	ચ U+0A9A
र U+0930	Gujarati	ર U+0AAE
र U+0930	Gurmukhi	ਰ U+0A15
ल U+0932	Bangla	ল U+09B2
व U+0935	Gujarati	વ U+0AB5
श U+0936	Gujarati	શ U+0AB6
श् U+0936 U+094D	Bangla	ঽ U+09BD

ष U+0937	Gujarati	ષ U+0AB7
स U+0938	Gujarati	સ U+0AB8
ह U+0939	Gujarati	હ U+0AB9
Nukta characters		
ग़ U+095A or U+0917 U+094D	Gurmukhi	ਗ਼ U+0A5A
ढ़ U+095D Or U+ 0922 U+094D	Gurmukhi	ਢ਼ U+0A5E

Table 7

4.2. Homophones generated though Spellings

In Devanāgarī based languages, homophonic variants which admit two homophones (spelling variants as in English *color-colour*) e.g. हिंदी and हिन्दी /hĩdi:/[7] do occur but the rules for such variants are ill-defined and could increase the chances of malfeasance. Within the ambit of the ccTLD policy for .भारत such variants have not been considered .

4.3. Zero Width Joiner (ZWJ) and Zero Width Non-Joiner (ZWNJ) :

[7] cf. Part 2 supra

ZWJ (U+0200D) and ZWNJ (U+0200C) are code points that have been provided by the Unicode standard to instruct the rendering of a string where the script has the option between joining and non-joining characters. Without the use of these control codes, the string may be rendered in an alternate form from what is intended. Technically TLD's do not allow either of these and this may seem to be a non-issue. However two cases need to be considered:

4.3.1. The case of the Eye-lash ra[8]

र्‍ U+0930 U+094D U+200D	र्‍ U+0931 U+094D

Unicode 2.0 prescribes the use of RA+VIRAMA+ZWJ to represent the eyelash-ra. This is captured in what was then rule R5 of Section 9 (which is now rule R5a). Unicode 3.0/4.0 reflected the ISCII choice, in what is now rule R5: "*In conformance with the ISCII standard, the half-consonant form rrah is represented as eyelash-ra. This form of ra is commonly used in writing Marathi…*" (Unicode 3.0)

So, the word दर्‍या/ darya/ "valleys can be written with the Unicode values U+0926 **U+0930 U+094D U+200D** U+092F U+093E (दर्‍या) as well as U+0926 **U+0931 U+094D** U+092F U+093E (दऱ्या)

Users of Nepali in Nepal do not recognize the character RRA U+0931: ऱ, nor does it figure in their font inventory. To generate out the eye-lash ra, the practice has been to use the combination: Ra(U+0930)+ VIRAMA(U+094D)+ ZWJ(U+200D). However since ZWJ is not permitted in the TLD label, the Nepali users would need to find alternate methods to display the eye-lash ra within a TLD Label.

4.3.2. The case of Noun Paradigms:

In languages such as Nepali, the use of ZWNJ permits the correct generation of certain Noun paradigms, as illustrated in the following example:

श्रीमान् + को =श्रीमान्को

श्रीमान् + Non-Joining Character + को =श्रीमान्‌को

[8] The eyelash ra is used in Konkanai, Nepali and Marathi. Denoted as र् it is treated as different from the र् (repha) by certain linguists. While the former is treated as a flap, the latter is a continuant trill (*cf.*, Kalyan Kale and Anjali Soman. 1986). There are cases in Marathi of minimal pairs such as: आचार्यास "to the teacher" vs. आचाऱ्यास "to the cook or दर्या /darya/ "ocean" vs. दऱ्या /darya/ "valleys. Similar cases may exist in Konkani and Nepali.

The word श्रीमान् [shrīmān] ends in a Virama. Adjoining to it the suffix को [ko] generates an incorrect form where the suffix and the root form a conjunct श्रीमान्को. This would be inacceptable to the user community. To ensure that the root form and the suffix are clearly indicated, ZWNJ is inserted as shown in the example above.
Constraining rules cannot be applied in this case since the number of such paradigms is very large and a dictionary look-up would not be feasible.

4.4. Administrative Issues [9]

These issues are pertinent to the policy to be adopted by the Government of India in the domain of opening up domain names, reserved names, conflict resolution and also the fee structure.
Certain issues arise here, quite a few of which are in the nature of legalities and economic policies.

4.4.1. RESERVED NAMES LIST

Reserved names Lists are deployed for sensitive names which need to be protected by a given country. The following issues could arise, especially with regard to gTLD's.

1. Would gTLD's need a reserved list? Will the Government send a list of reserved names of political sensitivity? If so are payment issues involved? (in which case specific processes could be needed for variants).
2. Should all variants of a given reserved gTLD be also requested for blocking ?

4.4.2. DISPUTE RESOLUTION

This is an area of legal policies and mechanisms need to be evolved for handling the same, especially given the introduction of multi-lingual labels. While areas such as "bad faith" and cyber-squatting" already have legal redress mechanisms, multi-lingualism brings in its own issues:
Multi-lingual are bundles containing labels in different languages. The following major issues can be identified here:

1. How does a complainant claim rights to a whole label ?

2. Can a complaint be filed if a complainant comes to know that a party has filed for a domain name in which the complainant has valid claims

3. Decision-making mechanisms
   Are precedents allowed ? And if so what mechanism will be evolved for precedents ? Would a separate set of mechanisms need to be involved in multi-lingual ownership? An important issue is that of expertise in resolving a dispute. Simply put who will deem

[9] Although not truly within the purview of the Variant Issues Project, the issues presented here could widen the debate and are hence retained.

a complaint as valid in the area of a multi-lingual dispute. Will the matter be referred to the State Government or to a competent language authority

4. International Trademark resolution:
   Which procedure would be followed when a trademark or domain name is claimed by two countries ?
   e.g. Tamil is shared both by Sri Lanka and India as an official language. What would happen if a trademark in Tamil for a corporate in one country closely resembles a similar one belonging to a corporate in the other country ? Will the label be frozen and treated subjudice during the period of litigation ?

5. Government vs. an Individual or a Corporate body:
   Will priority be given to Government over Individual claim in case of such a litigation ?

4.4.3. PAYMENT ISSUES

With the creation of multi-lingual labels and also variants generated from each, certain issues of payment arise:

1. Will there be a fee for providing and registering Variants
2. Will there be a fee for a registrant desirous of removing a variant granted to him (issues of cyber-squatting)
3. Will there be a concession for providing the registrant a label in multiple languages ?

4.5. Management Of Multi-Lingual gTLD's

The issues raised here are specific to gTLD's where TLD's are managed outside a country's law .

Certain issues need to be discussed in this area:

1. How are these to be allocated, especially when more than one country shares the same language?
2. Will there be a specific reservation for a country to register its societal and politically sensitive names such as political parties, name of a language etc ?

3. And corollary to the above which policy will apply for generation of variants ? Will the registrant be permitted to block out variants which are possible ? What would be the financial implications of the same ?

4. If a given corporate body is desirous of registering a gTLD in a variety of scripts, which policy will apply? It is suggested that the policy determined for each script/language be applied to resolve the issue.

5. REGISTRAR AND REGISTRY PERSPECTIVE

Within a registry, there is an important technical consideration when registering internationalized domain names. The domain name must be tagged with both a script indication and a language indication. In order to achieve this, a registry will have to establish certain policies that must be enforced when a request to register a domain name is received. The technical issues to be considered in the development of these policies are as follows.

In some cases, it may be sufficient to tag a domain name with either its script or its language. For example, the Gurumukhi script is only used for the Gurumukhi language. In this case the registry can infer the language when it receives a domain name with the Gurumukhi script tag.

Similarly, only the Tamil script supports the Tamil language. Thus when a domain name is tagged with the Tamil language the registry can infer the Tamil script tag.

However, either the Devanāgarī or Perso-Arabic script can support the Sindhi language. In this case when the registry receives a domain name to be registered it must be tagged with both its language and its script.

Also, the Devanāgarī script can be used to support many languages, e.g., Hindi or Nepali. In this case when the registry receives a domain name to be registered it must be tagged with both its language and its script.

The technical issue is that there is no standard way to do this in the standard EPP protocol used by gTLD registries and those ccTLD registries that choose to follow the ICANN recommendations. There is a defined extension for including each of these values but not both together. This issue is being currently pursued with the IETF.

This issue also affects registrars in two ways. To the extent there is no standard, a registrar will have to implement all EPP extensions that various registries may choose to specify to resolve this issue. For those cold's that do not use EPP registrars will have to implement whatever is required in order to support that ccTLD.

In addition, when registrars are present they are the interface to the registrant. Registrars that choose to support multiple scripts and languages will need to develop user interfaces that facilitate and simplify the identification of the script and language in use by a registrant.

Finally, with respect to the issue of a preferred variant, our discussions have noted that in general no variant is preferred over any other variant. However, RFC 3743 requires that at least one code point be specified in the preferred variant column of a language table. In the context of the Devanāgarī script it would be preferred if the preferred variant column could be left blank until a registrant chooses the desired code point. At that time, operationally, a

registry could then insert the chosen code point in to the preferred variant column before proceeding with the rest of the registration process.

5.1. DNS Technology and Operations Perspective

It is important to keep in mind that the DNS is technically a pure lookup protocol: a request is made for specific information (DNS record type) indexed by a domain name that is returned in a response. In the case of internationalized domain names, the domain name in the request is required to be an A-LABEL. Perhaps more importantly, the DNS is agnostic with respect to language and script as this information is neither stored in the DNS nor directly available in any part of the global DNS infrastructure. In that context, from a purely technical point of view, internationalized domain names do not present any unique challenges to the operation of the DNS.

However, a common point of discussion in the context of internationalized domain name TLDs is the desire to "alias" one TLD with another. The specifics of the desired "alias" behavior are varied but the intent, conceptually, is that a lookup of a domain name in one TLD return the same response as a corresponding lookup in the "aliased" TLD. For the two domain names to be corresponding the intent is usually that they be "variants" of each other, and therein lies the principal point of contention. There is no consensus as to the definition of "variant".

A full treatment of the possible definitions is beyond the scope of this comment. However, it is important to note that not all definitions can be fully implemented and enforced with today's DNS technology. This will have an effect on registry policies regarding "aliasing".

The critical gap is that policies regarding DNS behavior cannot be enforced beyond the level in the DNS hierarchy at which the policy is defined. Specifically, a registry may choose to establish a policy wherein all possible variants will behave the same (return the same response in the DNS) at the TLD level of the DNS hierarchy. Although this can work in many cases at the TLD level, the DNS cannot enforce this policy on the delegated second-level domain names in the TLD. This can have a dramatic affect on the user experience.

5.2. Security and Stability

A suggestion for evaluating variant policies and their implementation is to log, review, and analyze DNS query traffic. Specifically, the behavior of applications and services, and sometimes the users that use them, can be inferred from traffic patterns found in sequences of DNS queries and responses. For example, registries could review DNS traffic of the TLD for queries of non-existent domains (i.e., in DNS terms reviewing the NXDOMAIN responses). An analysis of these transactions may indicate that language tables are incomplete or that variant usage is not as expected.

Providing a consistent, uniform, and non-surprising (i.e., user expected) experience to the user is an essential component of stability. DNS transaction logs provide some insight into user

expectations and thus some ability to confirm that the needs of a user community are being met.

Some TLDs may wish to consider partnerships with second-level domain holders to continue the analysis at lower levels in the DNS hierarchy.

5.3. User Perspective

There are two issues to be considered from a user's perspective when introducing internationalized domain names: the submission and display of internationalized domain names. There are two underlying technical issues. First, can a user enter the desired Unicode code point in to the system? The answer depends in part on the hardware (does the keyboard in use make the code point available) and also on the software (will the software accept the code point value as a valid entry). Second, will the system in use display the Unicode code point in a way that is recognizable to the user? The answer depends in part on the availability of an appropriate font table indexed by the code point with a value representing a glyph that will be recognized by the user when displayed.

These issues are mostly straightforward to resolve in a local context but, when considered in a global context, they become challenging when you consider how a user is expected to maintain their environment such that it "works" in all cases. In this context, "works" means that the user experience remains uniform and consistent, i.e., the user is not surprised by any entry or presentation event. Specifically, consider the case of a web browser.

Web browsers today are commonly regionally packaged, which means it is possible to obtain a browser for whom its default behavior is optimized for the regional scripts or languages in use. However, this requires that appropriate hardware and software is available to support the browser (and the user). In addition, a user's usage of a browser frequently extends beyond the regional area, which means that a user may encounter web sites or information on web sites (documents) that cannot be displayed or used in the local environment without additional configuration (changes to the hardware or software or both).

The critical question is how the local environment (hardware and software) is maintained in the presence of changing entry and presentation needs or requirements?

5.4. System Administrator Perspective

The system administrator as a role is responsible for maintaining a local environment. In an enterprise situation there is a higher probability of greater skill being present and, thus, the maintenance of the local environment is more likely to be constrained by resources (e.g., staff or money). However, many users have mobile devices or other personal resources for which they serve the dual role of system administrator and end-user. These users are more likely to lack the skills necessary to properly maintain their local environment in order to achieve the best user experience possible.

5.5. End-User Perspective

Registration: It is important to keep in mind that the vast majority of users are monolingual and that in many cases the language and script are not Latin-based. The DNS requirement that queries of internationalized domain names be executed with the A-LABEL form of the name presents a burden for end-users. The A-LABEL form of the name is an encoding that transforms the name (using a reversible mapping) such that it is comprised only of US-ASCII characters. This transformation ensures that the use of internationalized domain names is backwards compatible with the existing DNS infrastructure. Working with the A-LABEL form is a burden for many end-users, in part because the encoding presents itself as a random sequence of US-ASCII characters but primarily because working with it is unnatural, even for those familiar with US-ASCII.

The use of appropriate software can mitigate this burden, the consequence of which is that users are constrained by their local hardware and software.

5.6. WHOIS Issues

The critical WHOIS issue facing the deployment of IDNs is the fact that the standard WHOIS protocol (as defined by RFC 3912) has not been internationalized, which means there is no standard way to indicate either a preferred language or script, or the actual language or script in use. The WHOIS protocol is a simple request and response transaction: a domain name is submitted to a server and output is returned. The predominant encoding in use on the Internet today is US-ASCII but a consequence of the lack of internationalization is that there is an increasing number of local, regional, and proprietary solutions that attempt to address the lack of internationalization. This variability has a dramatically adverse effect on the user experience.

For example, the labels used to tag the information in the WHOIS response are predominantly indicated in US-ASCII. It is straightforward to believe that the labels should be show in the same language or script as the data itself, but this is not possible with the standard WHOIS protocol.

Secondary to this issue, the question of what to display is a policy issue that will be guided, in part, by the variant registration policy. Consider the following questions.

1. If a variant domain name exists in the registry database but is not present in the DNS (i.e., the domain name is reserved), should a WHOIS request for the domain name return a referral indicating the name is a variant of a superordinate name or return the response for the superordinate name? Should the response indicate the name does not exist?
2. Should variant domain names be permitted to have different WHOIS information associated with them? The answer to this question should depend in part on whether different owners are permitted to register variant domain names.
3. If a variant domain name is a different language or script than its corresponding superordinate domain name, how is this to be presented to the user if the user does not

understand (or perhaps cannot display) the superordinate domain name's language or script?
4. If a WHOIS request is for a domain name with variants, should the variants be included in the response? What if the language or script of the variants cannot be understood or displayed by the user making the request?

5.7. Registration Process Issues

The critical technical issue facing the registration of IDNs and variants is the fact there is no standard way in the EPP protocol to indicate the language, script, or both in use by a domain name to be registered. As described in the Registry and Registrar perspective, this affects the user interface provided to a registrant as well as a registry's ability to know which domain name among a set of variants to register.

Secondary to this issue, a registry will need to have a policy specifying how it will deal with variants of prospective domain name registrations. Consider the following questions.

1. Are domain name variants to be considered equivalent, for an appropriate definition of equivalence?
2. If variants are equivalent, will all be registered (including DNS delegation) when the first one is presented? Will variants be reserved (does not include DNS delegation) and only registered upon request?
3. If variants are reserved for registration upon request, who is permitted to request registration? The owner of the first registered variant or anyone who requests it?

A critical technical issue to the question of equivalence is the implications to the DNS as described in the DNS Technology and Operations Perspective. The DNS behavior cannot be enforced beyond the level in the DNS hierarchy at which the policy is defined. This can have a dramatic effect on the user experience.

Finally, from a business perspective, a registry will need to have a policy specifying how it will charge (or not charge) for variants of registered domain names.

5.8. DNSSEC Issues

There are no IDN or variant specific issues that affect the deployment of DNSSEC.

From the point of view of DNSSEC, an IDN or variant TLD is simply another zone. Recall from the DNS Technology and Operations Perspective discussion that the DNS has no context with respect to the purpose or value judgment of the labels in a zone. The DNS is technically a pure lookup protocol.

A common point of discussion is to correlate the issue of TLD "aliasing" with the key management issues that must ordinarily be resolved when deploying DNSSEC. This coupling is an unnecessary complexity since the questions related to implementing key management

should be answered only in the context of DNS and DNSSEC, i.e., an IDN or a variant should be just a "label" to the DNS and DNSSEC.

6. SELECT BIBLIOGRAPHY

The bibliography given below and sorted thematically is a set of documents, books, articles and webographies consulted in the drafting of this report

WRITING SYSTEMS

Dillinger. D., The Alphabet. A Key to the History of Mankind. 3rd Edition in 2 Volumes. Hutchison. London. 1968.

DEVANĀGARĪ

Agrawala, V. S. (1966). The Devanāgarī script. In: Indian Systems of Writing. (Pp. 12-16) Delhi: Publications Division.

Agyeya, Sacchindanand Hiranand Vatsyayan. 1972. Bhavanti. Delhi: Rajpal and Sons.

Beames, John. 1872-79. A Comparative Grammar of the Modern Aryan Languages of India. 3 vols. London, Trubner and Co. [Reprinted by Munshiram Manoharlal, New Delhi, 1966.]

Bhatia, Tej K. 1987. A History of the Hindi Grammatical Tradition: Hindi-Hindustani Grammar, Grammarians, History and Problems. Leiden/New York: E. J. Brill.

Bright, W. (1996). The Devanāgarī script. In P. Daniels and W. Bright (eds), The World's Writing Systems. (Pp. 384-390). New York: Oxford University Press.

Cardona, George. 1987. Sanskrit. In The World's Major Languages. Bernard Comrie (ed.). London: Croom Helm. 448-469.

Dwivedi, Ram Awadh. 1966. A Critical Survey of Hindi Literature. Delhi:Motilal Banarsidass.

Faruqi, Shamsur Rahman. 2001. Early Urdu Literary Culture and History.Delhi: Oxford University Press.

Guru, Kamta Prasad. 1919. Hindi Vyakaran. Varanasi: Nagari Pracharini Sabha. (1962 edition).

Kachru, Yamuna. 1965. A Transformational Treatment of Hindi Verbal Syntax. London: University of London Ph.D. dissertation (Mimeographed).

Kachru, Yamuna. 1966. An Introduction to Hindi Syntax. Urbana: University of Illinois, Department of Linguistics.

Kalyan Kale and Anjali Soman, 1986. Learning Marathi. Shri Vishakha Prakashan, Pune :

McGregor, R. S. (1977). Outline of Hindi Grammar. 2nd edn. Delhi: Oxford University Press.

McGregor, R. S. 1972. Outline of Hindi Grammar with Exercises. Delhi: Oxford University Press.

McGregor, R. S. 1974. Hindi Literature of the Nineteenth and Early Twentieth Centuries. Wiesbaden: Harrassowitz.

McGregor, R. S. 1984. Hindi Literature from Its Beginnings to the Nineteenth Century. Wiesbaden: Harrassowitz.

Pandey, P. K. (2007). Phonology-orthography interface in Devanāgarī for Hindi. Written Language and Literacy, 10 (2): 139-156. 2007.

Rai, Amrit. 1984. A House Divided. The Origin and Development of Hindi/Hindavi. Delhi: Oxford University Press.

Sharad, Onkar. 1969. Lohiya ke Vicar. Allhabad: Lokbharati Prakashan.

Singh, A. K. (2007). Progress of modification of Brāhmī alphabet as revealed by the inscriptions of sixth-eighth centuries. In P.G. Patel, P. Pandey and D. Rajgor (eds), The Indic Scripts: Paleographic and Linguistic Perspectives. (Pp. 85-107). New Delhi: DK Printworld.

Sproat, R. (2000). A Computational Theory of Writing Systems. Cambridge University Press.

Tiwari, Pandit Udaynarayan. 1961. Hindi Bhasha ka Udgam aur Vikas [The Origin and Development of the Hindi Language]. Prayag: Leader Press.
Verma, M. K. 1971. The Structure of the Noun Phrase in English and Hindi.Delhi: Motilal Banarsidass.

MULTILINGUALISM

GENERIC

Multilingual Internet Names Consortium. MINC.
Dam, Mohan, Karp, Kane & Hotta, IDN Guidelines 1.0, ICANN, June 2003
Martin J. (December 20, 1996). "URLs and internationalization". World Wide Web Consortium. IDN TABLES: http://www.iana.org/domains/idn-tables/

LANGUAGE SPECIFIC

1. *INDIAN SCRIPTS AND LANGUAGES*

IS 10401: 8-bit code for information interchange. 1982
IS 10315: 7-bit coded character set for information interchange. 1985
IS 12326: 7-bit and 8-bit coded character sets-Code extension techniques. 1987
ISO 15919, Information and documentation - Transliteration of Devanāgarī and related Indic scripts into Latin characters. 2001
ISO 2375: Procedure for registration of escape sequences. 2003
ISO 8859: 8-bit single-byte coded graphic character sets - Parts 1-13. 1998-2001
IDN POLICY http://mit.gov.in/sites/upload_files/dit/files/India-IDN-Policy.pdf

Romanisation of Indian scripts

Library of Congress. Romanization Standards.. USA. 2002
Stone. Anthony., http://homepage.ntlworld.com/stone-catend/trind.htm

2. *CHINESE*

CHINESE: Chinese Domain Name Consortium". CDNC. 2000-05-19

3. *URDU*

URDU: http://urduworkshop.sdnpk.org

Romanisation of Indian scripts

RFC's

RFC 2181, Clarifications to the DNS Specification: section 11 explicitly allows any binary string
RFC 2870 Root Name Server Operational Requirements June 2000
RFC 3490 Internationalizing Domain Names in Applications (IDNA) March 2003
RFC 3492, Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA), A. Costello, The Internet Society (March 2003)
RFC 5890 "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework"
RFC 5891 "Internationalized Domain Names in Applications (IDNA): Protocol"
RFC 5892 "The Unicode Code Points and Internationalized Domain Names for Applications (IDNA)" August 2010

RFC 5893 "Right-to-Left Scripts for Internationalized Domain Names for Applications (IDNA)"

UNICODE

Unicode Consortium. Unicode ver.3.0.
---. Unicode ver.3.2.
---. Online version of Unicode ver.4.1 . (archived).
----. Online version of Unicode ver. 5.0 & 5.1. (www.unicode.org)
----. Online version of Unicode ver.6.0 (www.unicode.org)

LIST OF APPENDICES

APPENDIX 1:

Devanāgarī Team Members

Member	Role
Dr. Govind	Case Study Coordinator
Dr. Mahesh Kulkarni	Team Member
K. B. Narayanan	Team Member
Dr. James Galvin	Team Member
Amardeep Singh Chawla	Team Member
Tulika Pandey	Team Member
Jitender Kumar	Team Member
Rajiv Kumar	Team Member
Bhavin Turakhia	Team Member
Shashi Bharadwaj	Team Member
Prof Pramod Pandey	Team Member
Dr. Raiomond Doctor	Team Member
Dr. Kalyan Kale	Team Member
Prabhakar Kshotriya	Team Member
Manish Dalal	Team Member
Basanta Shrestha	Team Member
Bal Krishna Bal	Team Member
Satyendra Kumar Pandey	Team Member
Neha Gupta	Team Member
Akshat Joshi	Team Member

STAFF MEMBERS

Member (staff)	Role
Francisco Arias	Subject Matter Expert (Registry Ops)
Naela Sarras	Case Study Liaison
Nicholas Ostler	Subject Matter Expert (Linguistics)
Steve Sheng	Subject Matter Expert (Policy)
Andrew Sullivan	Subject Matter Expert (Protocol)
Kim Davies	Subject Matter Expert (Security)

APPENDIX II:
List of Official Languages of India[10]

India is a linguist's hunting ground with 4 major language families, over 6616 languages (Census of India 2001) and 20000+ dialects having been identified[11] (SIL report). To face this vast diversity, a considerable amount of accommodation has been made by the Constitution of India which has stipulated the usage of Hindi and English to be the two languages of official communication for the national government. In addition a set of 22 scheduled languages have been identified which are languages that can be

a. officially adopted by different states for administrative purposes,
b. as a medium of communication between the national and the state governments,
c. for examinations at the University as well as government levels.
d. for national databases such as voter lists, Unique Identity Number program (UIDAI) etc.

The 22 scheduled languages are represented table wise as under. Languages using Devanāgarī script are highlighted:

Language	ISO	Official Language	Family	Script
Assamese	asm	Assam	Indo-Aryan	Assamese
Bengali	ben	Tripura and West Bengal	Indo-Aryan	Bangla
Manipuri	mni	Meitei	Tibeto-Burman	Bangla Meitei-Meyek
Boro	brx	Assam	Tibeto-Burman	Devanāgarī (modified)
Dogri	dgr	Jammu and Kashmir	Indo-Aryan	Devanāgarī (modified)
Hindi	hin	Andaman and Nicobar Islands, Bihar, Chandigarh, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttaranchal	Indo-Aryan	Devanāgarī
Konkani	kok	Goa	Indo-Aryan	Devanāgarī Roman (Latin)
Maithili	mai	Bihar	Indo-Aryan	Devanāgarī
Marathi	mar	Maharashtra	Indo-	Devanāgarī

[10] This section has been contributed by GIST Group. CDAC
[11] http://www.ethnologue.com/show_country.asp?name=in

			Aryan	
Nepali	nep	Sikkim	Indo-Aryan	Devanāgarī
Sanskrit	san	Pan-Indian	Indo-Aryan	Devanāgarī
Gujarati	guj	Dadra and Nagar Haveli, Daman and Diu, and Gujarat	Indo-Aryan	Gujarati
Punjabi	pan	Punjab	Indo-Aryan	Gurmukhi
Kannada	kan	Karnataka	Dravidian	Kannada
Malayalam	mal	Kerala and Lakshadweep	Dravidian	Malayalam
Santali	sat	Jharkhand	Munda	Ol Ciki
Oriya	ori	Orissa	Indo-Aryan	Oriya
Kashmiri	kas		Indo-Aryan	Perso-Arabic Devanāgarī
Sindhi	snd	Pan-Indian	Indo-Aryan	Perso-Arabic Devanāgarī Gujarati Roman (Latin)
Urdu	urd	Jammu and Kashmir	Indo-Aryan	Perso-Arabic
Tamil	tam	Tamil Nadu and Pondicherry	Dravidian	Tamil
Telugu	tel	Andhra Pradesh	Dravidian	Telugu

Although these 22 languages belong to 4 distinct language families: Indo-Aryan, Dravidian, Munda and Tibeto-Burman, insofar as the writing system is concerned, two major families can be identified:
-Languages whose writing system has evolved from Brāhmī: e.g.. Hindi, Bangla, Punjabi and all the Dravidian languages
- Languages whose writing system is Perso-Arabic in nature. These are only three in number: Kashmiri, Sindhi, and Urdu. Of these Sindhi and Kashmiri can be written also using a Brāhmī based writing system viz. Devanāgarī .
Smaller sub-sets of writing systems can be seen in the case of languages such as Meitei and Ol Ciki which have indigenous script systems.

APPENDIX III:
List of Confusable Characters in Devanāgarī .

Character 1	Character 2
उ U+0909	ऊ U+090A
ङ U+0919	ड U+0921
ज U+091C	ञ U+091E
ब U+092C	व U+0935
ऋ U+090B	ॠ U+0960
थ U+0925	य U+092F
प U+092A	ष U+0937
भ U+092D	म U+092E
इ	ई

U+0907	U+0908
ए U+090F	ऐ U+0910
ओ U+0913	औ U+0914
क U+0915	फ U+092b
ट U+091F	ठ U+0920
त U+0924	ल U+0932
र U+0930	ऱ U+0931
ल U+0932	ळ U+0933

Appendix IV:
List of confusable Ligatures in Devanāgarī

1. LOOK-ALIKE PAIRS

UNICODE	H1	UNICODE	H2
U+0915 U+094D U+0915	क्क	U+0932 U+094D U+0932	ल्ल
U+0915 U+094D U+092F	क्य	U+092B U+094D U+092F	फ्य
U+0915 U+094D U+0932	क्ल	U+092B U+094D U+0932	फ्ल
U+0924 U+094D U+0924	त्त	U+0932 U+094D U+0932	ल्ल
U+0924 U+094D U+092A	त्प	U+0932 U+094D U+092A	ल्प
U+0924 U+094D U+0935	त्व	U+0932 U+094D U+092A	ल्व
U+0924 U+094D U+092F	थ्य	U+092F U+094D U+092F	य्य
U+092A U+094D U+092A	प्प	U+0937 U+094D U+092A	ष्प
U+092A U+094D U+092E	प्य	U+0937 U+094D U+092E	ष्य
U+092C U+094D U+092C	ब्ब	U+0935 U+094D U+0935	व्व
U+092C U+094D U+092F	ब्य	U+0935 U+094D U+092F	व्य
U+092D U+094D U+092F	भ्य	U+092E U+094D U+092F	म्य
U+0937 U+094D U+091F	ष्ट	U+0937 U+094D U+0920	ष्ठ
U+0936 U+094D U+0935	श्व	U+0936 U+094D U+0930 U+094D U+0935	श्च
U+0936 U+094D U+0928	श्न	U+0936 U+094D U+0930 U+094D U+0928	श्न
U+0936 U+094D U+0932	श्ल	U+0936 U+094D U+0930 U+094D U+0928	श्ल

2. LOOK-ALIKE TRPLETS

UNICODE	H1	UNICODE	H2	UNICODE	H3
U+0918	घ	U+0927	ध	U+0926 U+094D U+092f	द्य
U+0918 U+094D U+092F	घ्य	U+0927 U+094D U+092F	ध्य	U+0926 U+094D U+092f	द्य
U+091F U+094D U+091F	ट्ट	U+0920 U+094D U+0920	ठ्ठ	U+0922 U+094D U+0922	ढ्ढ
U+0924 U+094D U+092F	त्य	U+0932 U+094D U+092F	ल्य	U+0924 U+094D U+0924 U+094D U+092F	त्य

Appendix V

Topics extraneous to the Variant Issues Project, but deemed to be of interest.

Issues which are extraneous to the Variant Issues report but in which variants are involved, are presented here.

1. REGISTRY MANAGEMENT

Registry Management of ABNF[12], Restriction rules, Language Tables and Variant Tables

The issues arising from delegation of Devanāgarī labels were discussed above. These are closely allied to the issues arising from the manner in which the language and variant tables will be managed by the registry. This discussion is limited to the policy for भारत, although the issues raised, because of their generic nature, can have larger ramifications.

Some of the major issues that arise are as under:

1. In the case of Devanāgarī, a large number of languages use the code block U+900. Given that the registry for .भारत will have to provide language-wise solutions how will the registry maintain the language table ?
2. Corollary to the above, will the registry support a variant table for each language? The Hindi variant table has only two types of variants, whereas Marathi, Konkani and Nepali admit also the third type of variant table (cf. Section 2.2 supra)
3. In the case of TLD's other than.भारत, which rules will apply? It is suggested that in this case ICANN should deploy the rules and variant tables defined for each script/language

2. "Localization" of WHOIS

The term "Localization" has been used for WHOIS but the issues go far beyond. Two cases can be identified:

1. The label has no variant. In that case the major issue would be that of displaying the Information. Should the information be displayed in the language/script. Here language assumes priority. A Konkani speaker would not like information to be displayed in Hindi and vice-versa. Localization and language-wise information pertaining to WHOIS becomes a prime issue
2. Assuming that a given registrant is allocated variants (with/without payment of fees), this allocation raises the following issues:
    1. In a scenario where a user checks one variant should all the other variants linked to that variant be displayed. This becomes especially important in case ZWJ/ZWNJ are admitted, since on screen both variants will look alike

       e.g. In the case of a label such as गड्डा : pit

       गड्डा (without ZWNJ) गड्डा (with ZWNJ) give the same visual result

[12] Cf. footnote 12 supra. *ABNF is an acronym for Augmented Backus-Naur Formalism evolved to handle the Indic Akshar. Apart from rules governing Letters (L) it also handles Hyphen (H) and Digit (D)*

2. Corollary to the above should the WHOIS information be the same for a given label and its variant or should it be different ? The choice made will affect the registry functioning.
3. In a scenario where a variant is either deprecated or added at a later stage, how does the registry display such information. Will the registry have a systematic "re-indexing" and if so what will be the costs arising from it in terms of economics and logistics ?
4. The above case scenarios (1-3) are for variants which have been accepted. In the case of Type 2 variants where the variant is automatically blocked, should the registry display such variants also ?

Appendix VI:
Recommended List of Devanāgarī code points :

Continuous Range / Character	Validity Status as per RFC 5892	Unicode Name
0901..0903	PVALID	# DEVANAGARI SIGN CANDRABINDU..DEVANAGARI SIGN VISARGA
0905..0928	PVALID	DEVANAGARI LETTER A ..DEVANAGARI LETTER NA
092A..0933	PVALID	DEVANAGARI LETTER PA.. DEVANAGARI LETTER LLA
0935..0939	PVALID	DEVANAGARI LETTER VA.. DEVANAGARI LETTER HA
093A..093B	Reserved*	DEVANAGARI VOWEL SIGN OE..DEVANAGARI VOWEL SIGN OOE
093C..0945	PVALID	DEVANAGARI SIGN NUKTA..DEVANAGARI VOWEL SIGN CANDRA E
0947..094D	PVALID	DEVANAGARI VOWEL SIGN E..DEVANAGARI SIGNVIRAMA
094F	Reserved*	DEVANAGARI VOWEL SIGN AW
0950..0952	PVALID	DEVANAGARI OM..DEVANAGARI STRESS SIGN ANUDATTA
0956..0957	Reserved*	DEVANAGARI VOWEL SIGN UE..DEVANAGARI VOWEL SIGN UUE
0960..0963	PVALID	DEVANAGARI LETTER VOCALIC RR..DEVANAGARI VOWEL SIGN VOCALIC LL
0966..096F	PVALID	DEVANAGARI DIGIT ZERO..DEVANAGARI DIGIT NINE
0971..0972	PVALID	DEVANAGARI SIGN HIGH SPACING DOT..DEVANAGARI LETTER CANDRA A
0973..0975	Reserved*	DEVANAGARI LETTER OE..DEVANAGARI LETTER UUE
097B..097C	PVALID	DEVANAGARI LETTER GGA..DEVANAGARI LETTER JJA
097E..097F	PVALID	DEVANAGARI LETTER DDDA..DEVANAGARI LETTER BBA
02BC	PVALID	MODIFIER LETTER APOSTROPHE

* RFC 5892 since is based on Unicode 5.2 shows these Unicode 6.0 additions as "reserved" however they have been mentioned with assumption that when the IDNA protocol will migrate to successive version of Unicode, these characters be considered as characters that are a part of recommended Devanagari code points.